4-26-2013

New Non-Parametric Methods for Income Distributions

Shan Luo
Georgia State University

NEW NON-PARAMETRIC METHODS FOR INCOME DISTRIBUTIONS

by

SHAN LUO

Under the Direction of Dr. Gengsheng Qin

ABSTRACT

Low income proportion (LIP), Lorenz curve (LC) and generalized Lorenz curve (GLC) are important indexes in describing the inequality of income distribution. They have been widely used for measuring social stability by governments around the world. The accuracy of estimating those indexes is essential to quantify the economics of a country. Established statistical inferential methods for these indexes are based on an asymptotic normal distribution, which may have poor performance when the real income data is skewed or has outliers. Recent applications of nonparametric methods, though, allow researchers to utilize techniques

without giving data the parametric distribution assumption. For example, existing research proposes the plug-in empirical likelihood (EL)-based inferences for LIP, LC and GLC. However, this method becomes computationally intensive and mathematically complex because of the presence of nonlinear constraints in the underlying optimization problem. Meanwhile, the limiting distribution of the log empirical likelihood ratio is a scaled χ^2 distribution. The estimation of the scale constant will affect the overall performance of the plug-in EL method. To improve the efficiency of the existing inferential methods, this dissertation first proposes kernel estimators for LIP, LC and GLC, respectively. Then the cross-validation method is proposed to choose bandwidth for the kernel estimators. These kernel estimators are proved to have asymptotic normality. The smoothed jackknife empirical likelihood (SJEL) for LIP, LC and GLC are defined. Then the log-jackknife empirical likelihood ratio statistics are proved to follow the standard χ^2 distribution. Extensive simulation studies are conducted to evaluate the kernel estimators in terms of Mean Square Error and Asymptotic Relative Efficiency. Next, the SJEL-based confidence intervals and the smoothed bootstrap-based confidence intervals are proposed. The coverage probability and interval length for the proposed confidence intervals are calculated and compared with the normal approximation-based intervals. The proposed kernel estimators are found to be competitive estimators, and the proposed inferential methods are observed to have better finite-sample performance. All inferential methods are illustrated through real examples.

INDEX WORDS: Low income proportion, Lorenz curve, Generalized Lorenz curve, Kernel estimator, Bandwidth, Empirical likelihood, Bootstrap, Jackknife, Cross-validation

NEW NON-PARAMETRIC METHODS FOR INCOME DISTRIBUTIONS

by

SHAN LUO

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

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2013

NEW NON-PARAMETRIC METHODS FOR INCOME DISTRIBUTIONS

by

SHAN LUO

Committee Chair: Dr. Gengsheng Qin

Committee: Dr. Zhongshan Li

Dr. Xin Qi

Dr. Ye Shen

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2013

DEDICATION

This dissertation is dedicated to my dear parents, my advisor,
and all my dear friends.

ACKNOWLEDGEMENTS

I would like to give my deep thanks and sincere gratitude to everyone for helping me to complete this dissertation.

First and foremost I would like to thank Dr. Jeff Qin, my academic advisor for my Ph.D study, who is always walking with me in achieving my research goal. During my study, from the most fundamental mathematical proof, to the most basic language modification, Dr. Qin always instructs me with the most details in helping me get through different types of research difficulties. I can never be trained for the research ability as I have right now without Dr. Qin. I will never forget that Dr. Qin spent hours and hours in explaining the technical proofs to me on Saturday afternoons, and showed me the right direction.

I would like to express many thanks to my doctoral committee members Dr. Zhongshan Li, Dr. Xin Qi and Dr. Ye Shen who all agreed to come with me in thinking through my research topics. Without their support, I could not have done what I was able to do.

I would like to thank my parents who always want me to get a Ph.D degree, and provide me with the solid support.

My special thanks goes to my boyfriend, Xin Huang, who is always standing there with me. We take care of each other, and we grow up together. We have shown great strength in the past 2 years. While he is in Seattle and I'm in Atlanta, the distance between us is the diagonal line across the country, which is almost the longest distance in U.S.

I would also like to express many thanks to my director at work, Hicham Elhassani, who is a very encouraging boss and he always provides me with the strongest support in finishing my study, and my manger Tim Davis who leads me to the industry world step by step, and my colleagues Dr. Alex Liu and Dr. Scott Zrebiec who spend time in discussing the proof techniques with me and refresh me with new ideas, as well as my lovely team members Allen Li, Beining Zhang, Bijan Mizani, Cheng Lin, Emily Xu, Jih-Shiang Chern, Meimei Wu, Qianyi Zhao and Yvonne Phillips who make the strong team.

I would also express many thanks to my former manager Girmaye Gizaw at Georgia Department of Labor, who cares about my life and study at school, and Dr. Rosa Hayes who still keep me after I transfer my employment status, and my former co-workers as well as my good friends Helen Kim, Ann Hunter, Hans Friedrichsen, Roy Gains and Travis Williams. We care about each other as family members.

Last but not least, I would like to thank all professors at Mathematics & Statistics Department: Dr. Guantao Chen, Dr. Yu-Sheng Hsu, Dr. Jiawei, Liu, Dr. Yixin Fang, Dr. Jun Han, Dr. Xu Zhang, Dr. Yuanhui Xiao, and Dr. Ruiyan Luo for all that they have taught me, and the Ph.Ds Dr. Baoyi Yang, Dr. Binhuan Wang and Dr. Hanfang Yang, who greatly helped me with my research, as well as my dear friends, Amy Fomo, Dongmei Wang, Dong Yang, Fei Gong, Haochuan Zhou, Hongwei Wang, Huayu Liu, Meng Zhao, Shuang Liu, Shuman Guo, Tian Tang, Xiaoxi Wei, Xiaoxue Gao, Ye Cui, Yichao Yin, Yueheng An, Zhengbo Ma, Zhibo Wang, Zi Zhu, and many others.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- AMSE - Average Mean Square Error
- ARE - Asymptotic Relative Efficiency
- BCa - Bootstrap Bias Correction and Acceleration
- BT - Bootstrap
- CLT - Central Limit Theory
- CV - Cross-validation
- ECHP - European Community Household Panel
- EE - Estimating Equation
- EL - Empirical Likelihood
- EU - European Union
- GLC - Generalized Lorenz Curve
- JEL - Jackknife Empirical Likelihood
- KDE - Kernel Density Estimation
- LC - Lorenz Curve
- LIP - Low Income Proportion
- MAE - Mean Absolute Error
- MAR - Missing at random
- MCAR - Missing Completely at Random

- MLE - Maximum Likelihood Estimators
- MNAR - Missing not at Random
- MOM - Method of Moments
- MSE - Mean Square Error
- NA - Normal Approximation
- SES - Structure of Earnings Survey
- SJEL - Smoothed Jackknife Empirical Likelihood

PART 1

INTRODUCTION

1.1 Low Income Proportion

The low income proportion (LIP) is often used to evaluate social economic and poverty status. It is the proportion of the population income falling below a given fraction α ($0 < \alpha < 1$) of the β-th ($0 < \beta < 1$) quantile of the income distribution. Consider an income variable $X \in [0, \infty)$ with cumulative distribution function $F(x)$ and density function $f(x)$, the βth quantile $F^{-1}(\beta)$ is denoted as ξ_β. Then, the α fraction of the βth quantile $\alpha\xi_\beta$ of $F(x)$ is called low income line. The low income proportion is derived based on the low income line, which is actually the proportion falling below the low income line. Mathematically, the low income proportion is

$$\theta_{\alpha\beta} = P(X \leq \alpha\xi_\beta) = F(\alpha\xi_\beta). \tag{1.1}$$

Based in Luxembourg, functioned as the leading statistic department of European Union, Eurostat provides EU Member State with reliable statistics that allow comparisons across countries. It targets at offering a wide range of high quality data and statistics at country level to government, commercial business, education institute, non-profit organization and other public department. According to Eurostat 2012, low wage earners are defined as those employees who earn two thirds ($\alpha = \frac{2}{3}$) or less of the national median ($\beta = 0.5$) hourly earnings. Each Member State has different proportions of low-wage earners as shown in Table (A.1). There are 17 % of employees in EU are categorized as low-wage earners. The top five countries with the highest proportions of low-wage earners are Latvia (27.8 %), Lithuania (27.2 %), Romania (25.6 %), Poland (24.2 %), and Estonia (23.8 %), while the top five countries that own the lowest proportions of low-wage earners are Sweden (2.5

%), Finland (5.9 %), France (6.1 %), Belgium (6.4 %) and Denmark (7.7 %). European Community Household Panel (ECHP), which is a panel survey that runs from 1996 to 2001, interviews a sample of households and individuals to collect information such as living conditions, income, house, social, health and so on. According to the data estimated from 1996 ECHP survey, the proportion of low-wage earners is indicated to be 21% in UK (Yves and Chris, 2003).

A government should be on alert for a high value of low income proportion because it indicates a potential unstable social structure due to relative social wealth inequality. Further investigations and adjustments are suggested to maintain a healthy economic system. As a general social economic indicator, the estimation and inference of low income proportion may largely affect the decisions made by government regulators, business owners, and individual researchers.

Since the income distribution $F(x)$ is unknown, it is routine to estimate low income line $\alpha\xi_\beta$ first and then estimate the low income proportion $\theta_{\alpha\beta}$. The low income line has been widely applied in several poverty studies. For example, Preston (1995) discussed the reliability to estimate low income proportion based on simple random sample. Rongve (1997) proposed a statistical inference for the poverty index with fixed poverty lines. Zheng (2001) showed the poverty estimates are asymptotically normally distributed, and he developed an asymptotically distribution-free statistical inference for the poverty estimates. Yves and Chris (2003) showed how a linearization method of variance estimation can be applied to low income proportion based on Family Expenditure Survey data. However, most of the existing inferential methods are based on the limiting normal distribution or Gaussian processes. In practical situations, most income data is observed to be highly right-skewed. The skewness is due to quite small percentage of individuals with extremely high salaries, relative to the rest of others. Established statistical inferential methods may have poor finite-sample performance because of the skewness of the real income data. In regard to this challenge, recent efforts have yielded new statistical approaches to make inferences for low income proportion (Yang, Qin and Qin (2011)).

1.2 Lorenz Curve

The recent Occupy Wall Street Movement was induced by the uneven distributed social wealth and income. According to the report published by Congressional Budget Office in October 2011, in 2007, the top 1% of the richest population share 21% of market income, while the poorest 25% only obtain no more than 2% of market income. Those figures are actually inferring to Lorenz curve.

The Lorenz curve was first introduced by Lorenz (1905) for representing the inequality of the wealth distribution. It is the cumulative distribution function of the proportion of the distribution of wealth, and is often used to describe the degree of inequality in income.

The Lorenz curve has wide applications in the study of inequalities in disciplines of public health, medical research, social study, human service, industry, education, etc. For example, Lee (1999) applied the summary indices of the Lorenz curve in measuring the real-world medical diagnosis characteristics. Slottje (1989) used the Lorenz curve to analyze income inequality in economics study. Shorrocks (1983) extended the Lorenz dominance into social welfare analysis by partial orderings over income distributions for political study. Smith (1947) utilized the Lorenz curve directly in industry concentration analysis by census years and indicated that industry in the United States has been gradually decentralizing since 1899. There are many examples of development of the Lorenz curve in healthcare and clinical services research including graphical distribution of healthcare professionals (Kobayashi and Takaki (1992), Chang and Halfon (1997)), individual-level prescription data analysis (Hallas and Stovring (2006)). Other applications have been found in reliability (Gail and Gastwirth (1978)), ecology and bionomics study (Damgaard and Weiner (2000), Harvey, Gange, Hawes and Rink (2011)), agricultural analysis (Victor and Rodolfo (2004)), and human and poverty study (Foster, James and Shorrocks (1988)). Meanwhile, Lorenz curve is the foundation for many summary indices of income or wealth inequality such as the Gini index.

In economics, each point on the Lorenz curve can be used to represent the proportion of the total wealth owned by a certain proportion of households whose wealth rank from the

bottom.

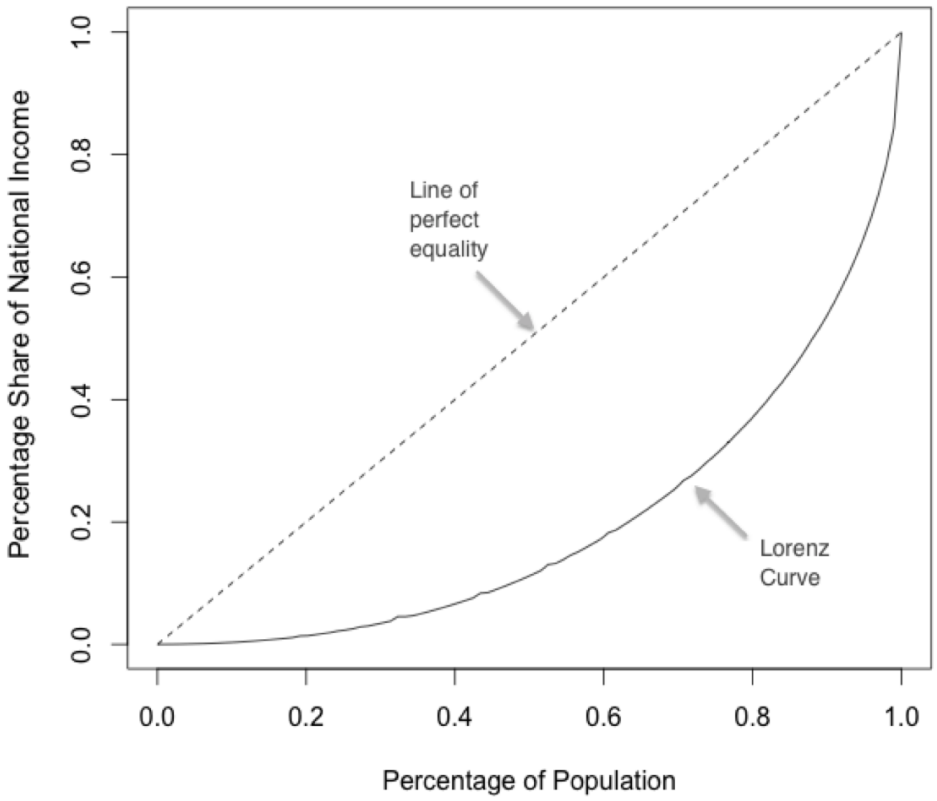

Figure 1.1 Lorenz curve based on real income data

Consider an income variable $X \in [0, \infty)$ with cumulative distribution function $F(x)$. The Lorenz ordinate is defined as follows:

$$\eta(t) = \frac{1}{\mu} \int_0^{\xi_t} x dF(x), \tag{1.2}$$

where $\mu = \int_0^{\infty} x dF(x)$ is the mean of $F(x)$, and $\xi_t = F^{-1}(t)$ is the t-th quantile of $F(x)$. For a fixed $t \in [0, 1]$, the Lorenz ordinate $\eta(t)$ is the percentage of total income owned by wealth-

holders of the lowest t-th percentage of incomes. Figure 1.1 refers to the Lorenz curve plot, the diagonal line is defined as the line of perfect equality, since the income distribution with perfect equality should be that each individual occupy the exactly the same social wealth, i.e., the bottom $t\%$ of society should own $t\%$ of wealth. However, in real world, the practical Lorenz curve would always fall below the line of perfect equality, due to the social wealth would never be equally distributed. So when the real Lorenz curve is too far below the line of perfect equality, it implies a high percentage of population own very low percentage of social wealth, which may cause social instability such as the Occupy Wall Street Movement.

To make inferences for the Lorenz ordinate, asymptotic theories have been developed and used to establish inference for Lorenz curve. Beach and Davidson (1983), and Beach and Richmond(1985) applied the statistical inference to the Lorenz curve and income shares by deriving the variance-covariance structure of (asymptotic) normal distribution, based only on the consistently estimated 1st and 2nd moments, and later a set of joint confidence bands are constructed by multiple comparisons. Zheng (2002) further derived the variance-covariance structure for Lorenz curves with non-simple random samples. Goldie (1977) proved the strong uniform consistency of the Lorenz curves and their inverses, and derived the limiting Gaussian processes for Lorenz process. Csorgo (1986) derived the empirical Lorenz process by the basic empirical diversity process. However, the existing methods made inferences based on the limiting normal distribution or Gaussian processes for Lorenz curves, which may have poor finite sample performance when the underlying distribution is skewed.

1.3 Generalized Lorenz Curve

The adoption of the Lorenz curve based on the assumption that the income distribution is independent of the population mean. The hypothesis of equal means, however, restricts the application of its use in the real world by requiring stronger conditions. Meanwhile, the assumption that Lorenz curves can't intersect with each other also limits its application on the real income data. With the motivation to overcome those limitations, Shorrocks (1983) and Kakwani (1984) extended the Lorenz curve to the generalized Lorenz curve by

taking the unequal means and the degree of income inequality into account at the same time. The height of the generalized Lorenz curve (GLC) is used to denote income level, and the convexity of GLC is used to denote the extent of income inequality.

Consider an income variable $X \in [0, \infty)$ with cumulative distribution function $F(x)$. The generalized Lorenz ordinate is defined as follows:

$$\theta(t) = \int_0^{\xi_t} x dF(x), \quad 0 \leq t \leq 1, \tag{1.3}$$

where $\xi_t = F^{-1}(t)$ is the t-th quantile of $F(x)$. For a fixed $t \in [0, 1]$, the generalized Lorenz ordinate $\theta(t)$ is the average wealth owned by the wealth-holders below the bottom t-th percent. (1.3) and (1.2) imply that the generalized Lorenz curve is derived by rescaling the Lorenz curve by the population mean μ of the income distribution.

Previous studies have been conducted on the properties of generalized Lorenz curve. Thistle (1989a) showed distribution functions are uniquely determined by their generalized Lorenz curve. Later, Thistle (1989b) derived the duality between generalized Lorenz curve and distribution functions, and showed that the generalized Lorenz dominance is equivalent to a second-order stochastic dominance. Kleiber and Kramer (2003) decomposed the generalized Lorenz order into two components: size and distribution. Beach and Davidson (1983) proved the asymptotic normality of the empirical estimator for GLC. Zheng (2002) further extended the asymptotical normality of the empirical estimator for GLC in non-simple random samples. Inferences can be made based on these asymptotic normal distributions for GLC. However, the normal approximation-based inference may have poor finite sample performance when income data is highly right skewed. Recently, generalized Lorenz curve has shown good results in empirical applications. Belinga-Hill(2007) applied interval estimation for the generalized Lorenz curve and concluded that empirical likelihood method performs better than normal approximation methods. Motivated by those previous researches, we develop several non-parametric methods to establish inference for the generalized Lorenz curve.

1.4 Bootstrap and Empirical Likelihood

Introduced by Efron (1981, 1982a), Bootstrap, a computationally intensive statistical technique, has been studied extensively in the literature. Four types of bootstrap confidence intervals are discussed by Diciccio and Efron (1996) including BC_a, Bootstrap-t, ABC and calibration. Efron (1987) constructed confidence interval for a single parameter in a multi-parameter family. Haukoos and Lewis (2004) demonstrated how to use bootstrap to get inferences for the median and Spearman rank correlation coefficient for data that does not follow normal distribution. Cambell and Torgerson (1999) used bootstrap methods in real data from clinical trials, and demonstrated that the confidence interval based on bootstrap is easy to built for cost-effectiveness ratios. Bootstrap has been proved to be a very powerful statistical method to construct confidence intervals. There are also some limitations of using bootstrap method. The assumption to use bootstrap is that the population distribution should be able to represent by distribution of the sample where the data is drawn from. If the sample is not larger enough, it will be difficulty to estimate efficient confidence intervals. And also bootstrap is heavily dependent on the skewness of sample. The sampling method used in generating bootstrap sample would also contribute to the sampling bias, according to Haukoos and Lewis (2005).

Empirical likelihood (EL), introduced by Owen (1988, 1990), allows researchers to utilize likelihood methods without any distribution assumptions. EL method has been shown to have diverse advantages over other methods. For example, we can use EL method to construct a confidence interval without choosing a parametric distribution. Next, the EL region is shaped by sample, especially in higher order asymptotic analysis, while the normal approximation method would assume a symmetrical shape. Based on these advantages, EL method is expected to work well for skewed data. Also, EL method is able to construct confidence interval without variance estimation. As mentioned by Wood et al (1996), under mild conditions, the log empirical likelihood ratio converges to a chi-square distribution, which is the Wilks' theorem. For more details, we refer to Owen's (2001) book for review.

EL method has been widely used in many applications. For example, many health care data analysis will not take the high skewness into account, which may cause significant deviations from actual average healthcare costs. Zhou, Qin, Lin and Li (2006) developed a new EL-based inference method in censored cost regression models and showed EL method outperforms the existing method in analyzing highly-skewed health care cost data. Chen and Qin (1993) applied EL estimation for finite population with efficient use of auxiliary information. Qin and Zhou (2006), Huang, Qin, Yuan and Zhou (2013) utilized EL inference based method for AUC. Liu, Zou and Zhang (2008) developed a EL method to make inferences for the difference of the means between two d-dimensional multivariate random samples with corrected Bartlett correction.

However, most of previous researches for EL are successfully applied in linear constraints. Significant computational intensity arise when EL is used in nonlinear statistics, due to the presence of nonlinear constraints in the underlying optimization problem. Suppose the sample size is n, to get the solution of the empirical likelihood ratio, we need to calculate $n+1$ number of nonlinear equations simultaneously. In this case, the selected optimization method become extremely important. Recently, Yang, Qin and Qin (2011) developed an EL method for low income proportion and showed that the EL inference achieved good performance on skewed income data. However, because their empirical likelihood method has been proven to follow a scaled Chi-square distribution, this method involves heavy computation burdens due to the estimation of the unknown scale constant. In order to alleviate the computational burden caused by the nonlinear statistics, Jing, Yuan and Zhou (2009) proposed a jackknife empirical likelihood (JEL) method for U-statistics. The general idea of JEL is to construct a jackknife sample which is shown to be asymptotically independent, then to implement EL on these jackknife pseudo-values (Quenouille 1956). Gong, Peng and Qi (2010) extended the JEL method with a smoothed ROC curve estimation, and observed that the JEL method results in shorter intervals than the naive boostrap intervals in most cases.

1.5 Brief Summary

In our research, kernel smoothed estimators for low income proportion, Lorenz curve, and generalized Lorenz curve have been proposed, and been compared to empirical estimators. The smoothed estimators have been proved to be asymptotically normal. Meanwhile, non-parametric inferences including bootstrap-based confidence intervals, and the smoothed jackknife empirical likelihood-based confidence intervals for low income proportion (LIP), Lorenz curve (LC), and generalized Lorenz curve (GLC) have been proposed and compared to the normal approximation-based confidence intervals and naive bootstrap-based confidence intervals. The limiting distributions of the log jackknife empirical likelihood ratio for LIP, LC and GLC have been proved to be a standard χ^2 distribution. After extensive simulation, the numeric study results show that the proposed smoothed estimators have smaller MSE and variance than the empirical estimators. Also, the smoothed jackknife empirical likelihood-based confidence intervals have the best finite sample performances in most cases, while the proposed bootstrap confidence intervals based on the smoothed estimators perform the second to the best.

This dissertation is organized as follows. In Part 2, we introduce the main methodology for low income proportion, including the discussion of how to choose a bandwidth based on a kernel estimator. A smoothed estimator is proposed for low income proportion(LIP), and this estimator is proved to be asymptotically normal. Then, we construct confidence intervals based on normal approximation, bootstrap and jackknife empirical likelihood methods. Next, extensive simulation studies are performed to evaluate the proposed point estimators and compare the performances of proposed intervals. Specifically, the empirical estimator and the proposed kernel estimator for LIP are evaluated in terms of Mean Square Error (MSE) and Asymptotic Relative Efficiency (ARE). Also, we evaluate normal approximation-based confidence intervals, bootstrap confidence intervals and the jackknife empirical likelihood confidence intervals based on empirical estimator and the proposed kernel estimator, respectively. As a real example, an individual income data set for full-time professors from Georgia

University System is used to illustrate the comparison. Proof of the theorems will be given at the end of this part. In Part 3, we first introduce the main methodology and theorem for the Lorenz curve, including the discussion of how to choose a bandwidth based on a kernel estimator. The smoothed estimator for Lorenz curve is proposed. Then, we establish the normal approximation-based inference, smoothed bootstrap-based inference and jackknife empirical likelihood-based inference. Those inferential methods are compared in the simulation section, as well as in a real example. Proof will be provided at the end of this part. In Part 4, a smoothed estimator is proposed for generalized Lorenz curve (GLC), and is proved to have asymptotic normality. Then it is compared with the empirical estimator for GLC in terms of Mean Square Error and Asymptotic Relative Efficiency. The smoothed jackknife empirical likelihood (SJEL) for generalized Lorenz curve is defined and the corresponding log-JEL ratio statistics is proved to follow the standard Chi-square distribution. Several interval estimations are evaluated through simulation and a real example. Finally, we discuss the conclusion and future work in Part 5.

PART 2

LOW INCOME PROPORTION

This part is organized as follows. In Section 2.1, we review the empirical estimator for low income proportion(LIP), then we define the kernel estimator for LIP. Next we propose methods to choose the bandwidth h for the kernel estimator. The proposed point estimator is evaluated together with the empirical estimator in terms of Mean Square Error (MSE) and Asymptotic Relative Efficiency (ARE). In Section 2.2, the jackknife pseudo-values for LIP is first defined, and jackknife empirical likelihood properties are then derived. In section 2.3, we construct the normal approximation confidence intervals and several bootstrap confidence intervals based on empirical estimator and the proposed kernel estimator, as well as the smoothed jackknife empirical likelihood-based confidence interval. In section 2.4, numerical studies are conducted to assess the performance of the proposed estimator, and of the proposed methods to build confidence interval. Confidence intervals based on the proposed methods are illustrated through a real example. Proof will be given at the end of this part.

2.1 Estimation of a Low Income Proportion

2.1.1 Empirical Estimator for Low Income Proportion

Low income proportion is defined as

$$\theta_{\alpha\beta} = P(X \leq \alpha\xi_\beta) = F(\alpha\xi_\beta), \tag{2.1}$$

Let $X_1, X_2, ..., X_n$ be a simple random sample drawn from the population X with cumulative distribution function F(x). Define the empirical estimator for $\theta_{\alpha\beta}$ to be $F_n(\alpha\xi_\beta)$. Since $F_n(\alpha\xi_\beta)$ depends on the αth fraction of the β-th quantile of the unknown population, we replace ξ_β with its consistent estimator $\hat{\xi}_\beta$, where $\hat{\xi}(\beta) = F_n^{-1}(\beta)$ is the β-th quantile of

$F_n(x)$. Then an empirical estimate for $\theta_{\alpha\beta}$ can be defined as

$$\hat{\theta}_{\alpha\beta} = F_n(\alpha\hat{\xi}_\beta) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq \alpha\hat{\xi}_\beta).$$

2.1.2 A Kernel Estimator for a Low Income Proportion

Since the empirical estimator $\hat{\theta}_{\alpha\beta}$ is a non-smoothing estimator for $\theta_{\alpha\beta}$, while in many applications, $\theta_{\alpha\beta}$ is a smoothing function. We will use kernel methods to develop a smoothed estimator for $\theta_{\alpha\beta}$. Extensive literature has shown the advantage of kernel estimation. Falk (1983, 1985) concluded that, for a distribution function $F(x)$, or its quantile function $F^{-1}(x)$, their corresponding kernel-based estimators asymptotically dominate their empirical estimators. Kernel estimations have been found in wide applications. Lloyd and Yong (1999) proved that the kernel estimator for the ROC curve performs better than the empirical estimator for its smaller mean-square error. The difference between empirical estimator and kernel estimator diminish as sample size increases. Hsieh and Turnbull (1966) showed that, for Youden index, the kernel-based estimator is superior to empirical estimators in that the MSE is asymptotically smaller by $O(h/n)$. In this section, we propose a kernel estimator for a low income proportion and develop confidence intervals based on bootstrap and jackknife empirical likelihood methods.

The kernel function is defined as $K(x) = \int_{-\infty}^{x} \omega(y)dy$, where ω is a density function. By substituting the indicator function $I(\alpha\xi_\beta - X_i \geq 0)$ by the kernel function $K(\frac{\alpha\hat{\xi}_\beta - X_i}{h})$, we construct a kernel estimator of the low income proportion $\theta_{\alpha\beta}$ as:

$$\hat{T}_n(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n} K(\frac{\alpha\hat{\xi}_\beta - X_i}{h}). \tag{2.2}$$

Kernel estimator $\hat{T}_n(\alpha,\beta)$ follows the asymptotic normal distribution as shown in Theorem 2.1:

Theorem 2.1. *Assume that the density function ω of the kernel function K has bounded support, its first derivative ω' exists and is bounded on its supporting set, and $\int_{-\infty}^{\infty} |\omega'(y)|dy <$*

∞. *If* $h = h(n) \to 0$, $\sqrt{n}h \to \infty$ *as* $n \to \infty$, *then*

$$\sqrt{n}\{\hat{T}_n(\alpha, \beta) - \theta_{\alpha\beta}\} \xrightarrow{d} N(0, \sigma^2_{\alpha\beta}),$$

where $\sigma^2_{\alpha\beta} = \frac{\alpha^2\beta(1-\beta)f^2(\alpha\xi_\beta))}{f^2(\xi_\beta)} + \theta_{\alpha\beta}(1 - \theta_{\alpha\beta})$, and $f(x)$ is the density function of the income distribution $F(x)$.

2.1.3 Bandwidth Selection for the Kernel Estimator by cross-validation Method

One of the difficulties in the calculation of the smoothed estimator $\hat{T}_n(\alpha, \beta)$ is to choose bandwidth h for the kernel estimator. The choice of the bandwidth h will strongly influence the performance of the kernel estimator. Extensive simulation analysis has been conducted to show that the choice of kernel function will not change the density estimation much. However, as in many kernel methods, the choice of the bandwidth h may influence the performance of the proposed kernel estimate.

Many methods have been proposed for selecting the bandwidth for kernel estimators.

In our study, we propose a cross-validation (CV) method for bandwidth selection. In order to ease the implementation, we utilize the 2-fold cross-validation method. The bandwidth h is suggested to be $h = cn^{-1/3}$, based on our simulation analysis. Then, the choice of h is controlled by the constant c. Here and thereafter, we denote $\hat{T}_{n,c}(\alpha, \beta) = \hat{T}_n(\alpha, \beta)$. For a given β, we select c by minimizing the Mean Squared Error(MSE).

$$MSE(c) = E[\hat{T}_{n,c}(\alpha, \beta) - \theta_{\alpha\beta}]^2.$$

For this purpose, we randomly split the sample into two parts, where the first part is treated as the training sample, and the other part is as the validation sample. The kernel estimator for low income proportion $\hat{T}^{(1)}_{n,c}(\alpha, \beta)$ is constructed based on the training sample, while the empirical estimator $\hat{\theta}^{(2)}_{\alpha\beta}$ is constructed from the validation sample. By repeating this random split many times, we will obtain the following cross-validation estimate of the

MSE

$$CV_c = \frac{1}{L}\sum_{l=1}^{L}[\hat{T}_{n,c}^{(1,l)}(\alpha,\beta) - \hat{\theta}_{\alpha,\beta}^{(2,l)}]^2,$$

where L is the number of random splits. Then, the value of c is chosen as the constant that minimize CV_c.

Figure 2.1 is a simulation example to illustrate the relationship between MSE and the constant c, which actually affects the bandwidth h. Clearly, the plot of MSE vs. bandwidth is a "smiling curve". The value of h corresponding to the lowest point of MSE will be the optimal bandwidth.

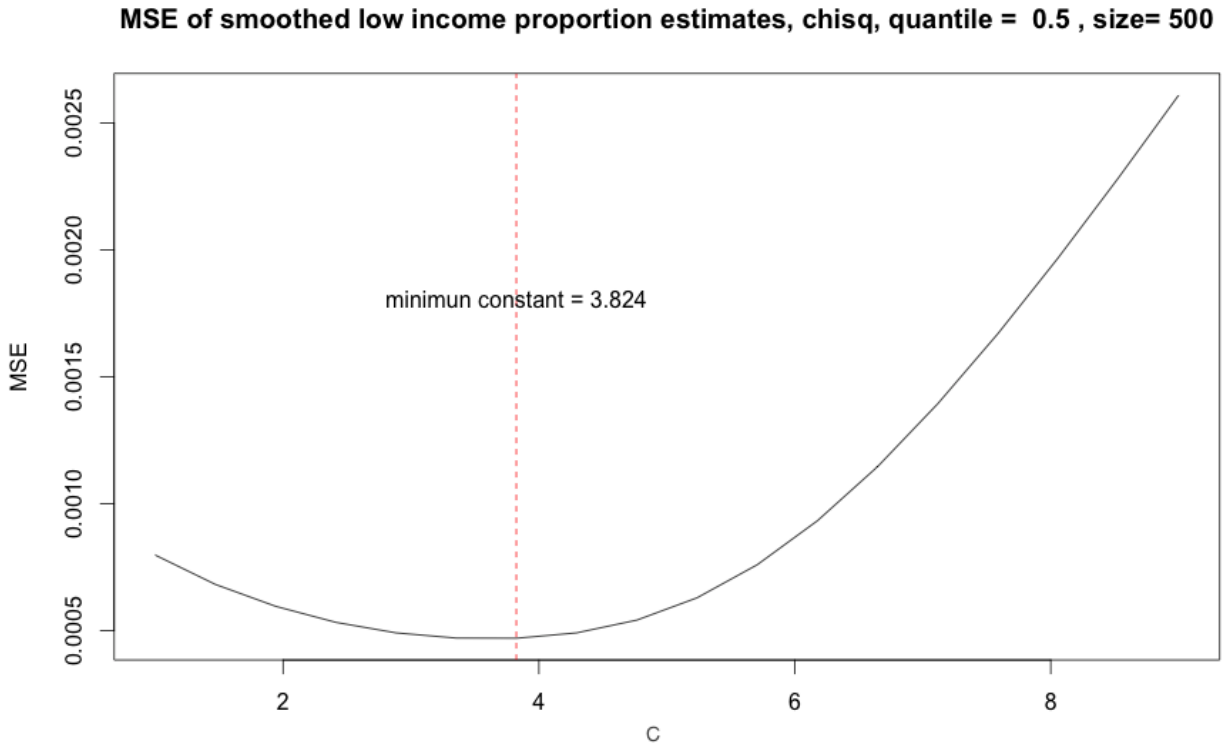

Figure 2.1 Bandwidth selection by MSE.

Alternatively, if we focus on the overall performance of the smoothed estimator for low income proportion across all β, we can use a similar cross-validation procedure for selecting

c by minimizing the Average Mean Squared Error (AMSE),

$$AMSE(c) = E\frac{1}{K}\sum_{k=1}^{K}[\hat{T}_{n,c}(\alpha, \beta_k) - \theta_{\alpha\beta_k}]^2.$$

where β_k is a fine grid of (0,1), and K is an integer.

And the cross-validation estimate of the AMSE is:

$$ACV_c = \frac{1}{L}\frac{1}{K}\sum_{l=1}^{L}\sum_{k=1}^{K}[\hat{T}_{n,c}^{(1,l)}(\alpha, \beta_k) - \hat{\theta}_{\alpha\beta_k}^{(2,l)}]^2.$$

Again, c is chosen as the one that minimize ACV_c.

Figure 2.2 illustrates the relationship between bandwidth and AMSE, and how we choose the constant c for bandwidth h.

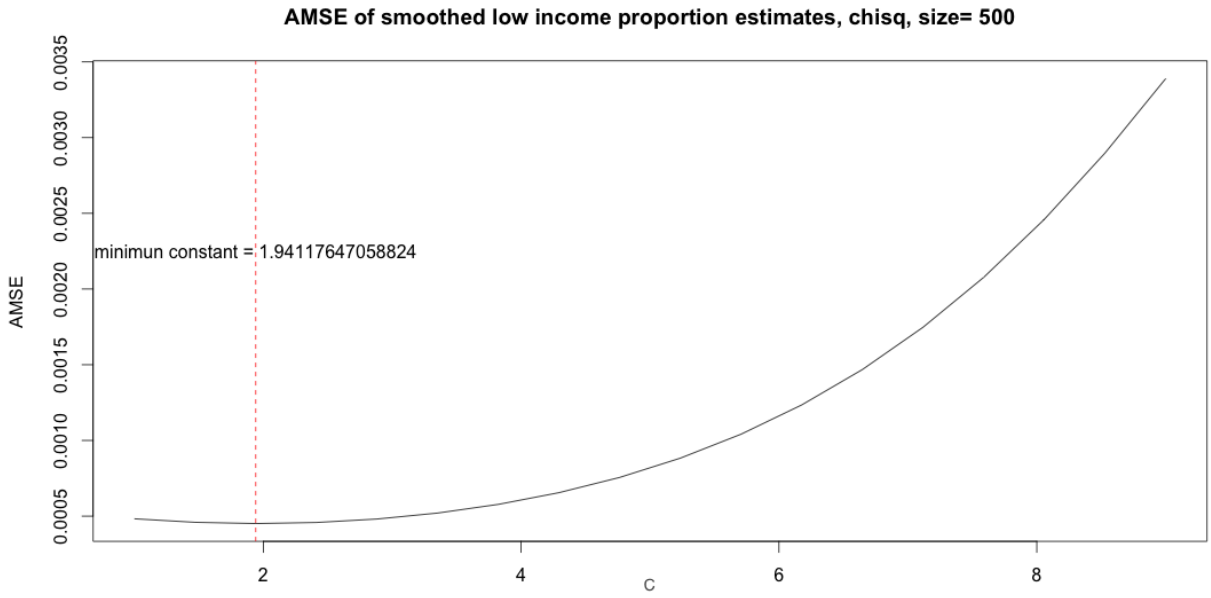

Figure 2.2 Bandwidth selection by AMSE.

Similarly, we choose the value of h corresponding to the lowest point of AMSE.

2.2 Smoothed Jackknife Empirical Likelihood for a Low Income Proportion

A smoothed version of jackknife empirical likelihood for the low income proportion will be defined in this section. Based on Tukey (1958) who used the jackknife method to estimate the variance, we define the jackknife pseudovalues for low income proportion as

$$\hat{V}_k(\alpha,\beta) = n\hat{T}_n(\alpha,\beta) - (n-1)\hat{T}_{n-1,k}(\alpha,\beta), k = 1,2,...,n. \tag{2.3}$$

where $\hat{T}_{n-1,k}(\alpha,\beta) = \frac{1}{n-1}\sum_{j\neq k}^{n} K(\frac{\alpha\hat{\xi}_{\beta,-j}-X_j}{h})$ is the given statistics T_{n-1} but computed on $n-1$ observations $X_1, X_2, ..., X_{k-1}, X_{k+1}, ..., X_n$, $F_{n,-j}(\alpha\hat{\xi}_\beta) = \frac{1}{n-1}\sum_{i\neq j}^{n} I(X_i \leq \alpha\hat{\xi}_\beta)$, and $\hat{\xi}_{\beta,-j} = F^{-1}_{n,-j}(\beta)$ is the βth quantile based on these $n-1$ observations.

Then, the jackknife empirical likelihood for $\theta_{\alpha\beta}$ can be defined as follows

$$L(\theta_{\alpha\beta}) = \sup\left\{\prod_{k=1}^{n} np_k : \sum_{k=1}^{n} p_k = 1, \sum_{k=1}^{n} p_k\hat{V}_k(\alpha,\beta) = \theta_{\alpha\beta}\right\}. \tag{2.4}$$

By using the Lagrange multiplier method, we obtain the maximization for (2.4) at

$$p_k = \frac{1}{n}\{1 + \lambda[\hat{V}_k(\alpha,\beta) - \theta_{\alpha\beta}]\}^{-1}, \tag{2.5}$$

where $\lambda = \lambda(\alpha,\beta,\theta_{\alpha\beta})$ is the solution to

$$\frac{1}{n}\sum_{k=1}^{n}\frac{\hat{V}_k(\alpha,\beta) - \theta_{\alpha\beta}}{1 + \lambda(\hat{V}_k(\alpha,\beta) - \theta_{\alpha\beta})} = 0. \tag{2.6}$$

Since $\prod_{k=1}^{n} p_k$ is subject to $\sum_{k=1}^{n} p_k = 1$, $p_k \geq 0$, k=1,2,..., n, $L(\theta_{\alpha\beta})$ will attain its maximum n^{-n} at $p_k = n^{-1}$. Thus, the jackknife empirical likelihood ratio for $\theta_{\alpha\beta}$ can be defined as

$$L_n(\theta_{\alpha\beta}) = \prod_{k=1}^{n}(np_k) = \prod_{k=1}^{n}\{1 + \lambda(\hat{V}_k(\alpha,\beta) - \theta_{\alpha\beta})\}^{-1}, \tag{2.7}$$

which gives the log jackknife empirical likelihood ratio as

$$l_n(\theta_{\alpha\beta}) = -2\log L_n(\theta_{\alpha\beta}) = 2\sum_{k=1}^{n}\log\{1+\lambda(\hat{V}_k(\alpha,\beta)-\theta_{\alpha\beta})\}. \tag{2.8}$$

Based on Turkey (1958), we conjecture that the pseudovalues $\hat{V}_i(\alpha,\beta)$, $i=1,...,n$ may be treated as though they were i.i.d, and $\hat{V}_i(\alpha,\beta)$ has approximately the same variance as $\sqrt{n}\hat{T}_n(\alpha,\beta)$. Therefore, the variance of $\sqrt{n}\hat{T}_n(\alpha,\beta)$, denoted as $var(\sqrt{n}\hat{T}_n(\alpha,\beta))$, can be estimated by the sample variance of $\hat{V}_1(\alpha,\beta)$, ..., $\hat{V}_n(\alpha,\beta)$. In order to prove that the above log jackknife empirical likelihood ratio converges to a χ^2 distribution, we construct the jackknife variance estimator for $\hat{V}_i(\alpha,\beta)$ as follows.

$$\begin{aligned} \upsilon_{JACK}(\alpha,\beta) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta)-\frac{1}{n}\sum_{j=1}^{n}\hat{V}_j(\alpha,\beta))^2 \\ &= \frac{n-1}{n}\sum_{i=1}^{n}(\hat{T}_{n-1,i}(\alpha,\beta)-\frac{1}{n}\sum_{j=1}^{n}\hat{T}_{n-1,j}(\alpha,\beta))^2. \end{aligned}$$

Then, the following theorems can be derived.

Theorem 2.2. *Under the conditions of Theorem 2.1, we have*

$$\upsilon_{JACK}(\alpha,\beta) \xrightarrow{p} \sigma^2_{\alpha\beta}, \tag{2.9}$$

where $\sigma^2_{\alpha\beta}$ is defined in Theorem 2.1.

Theorem 2.3. *Under the conditions of Theorem 2.1, if $\sqrt{n}h^2 \longrightarrow \infty$, we have*

$$l_n(\theta_{\alpha\beta}) \xrightarrow{d} \chi^2(1). \tag{2.10}$$

Detailed proofs for Theorem 2.2 and Theorem 2.3 will be given at the end of this part. In the next section, we will discuss methods for constructing confidence intervals of low

income proportion.

2.3 Confidence Interval for a Low Income Proportion

2.3.1 Normal Approximation-based Confidence Intervals

One of the most popular methods to construct a confidence interval for an unknown parameter is normal approximation. To construct a normal approximation based confidence interval for the Low Income Proportion $\theta_{\alpha\beta}$, we first need to obtain an appropriate estimator for $\theta_{\alpha\beta}$, and then derive its asymptotic normal distribution.

First of all, based on Preston (1995), the estimate $\hat{\theta}_{\alpha\beta}$ for the Low Income Proportion $\theta_{\alpha\beta}$ is asymptotically normal with variances σ_v^2,

$$\sqrt{n}(\hat{\theta}_{\alpha\beta} - \theta_{\alpha\beta}) \longrightarrow N(0, \sigma_v^2),$$

where

$$\sigma_v^2 = [\theta_{\alpha\beta}(1-\theta_{\alpha\beta})]^2 - 2\theta_{\alpha\beta}(1-\beta)\alpha\frac{f(\alpha\xi_\beta)}{f(\xi_\beta)} + \beta(1-\beta)\alpha^2\frac{f(\alpha\xi_\beta)}{f(\xi_\beta)}.$$

Therefore, the first $(1-\alpha)$ level normal approximation (NA1)-based confidence interval for $\theta_{\alpha\beta}$ can be constructed as follows

$$(l_1, u_1) = (\hat{\theta}_{\alpha\beta} - \frac{z_{1-\frac{\alpha}{2}}\hat{\sigma}_v}{\sqrt{n}}, \hat{\theta}_{\alpha\beta} + \frac{z_{1-\frac{\alpha}{2}}\hat{\sigma}_v}{\sqrt{n}}),$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})-th$ quantile of the standard normal distribution. Based on Preston (1995), a consistent estimate for σ_v^2 is

$$\hat{\sigma}_v^2 = [\hat{\theta}_{\alpha\beta}(1-\hat{\theta}_{\alpha\beta})]^2 - 2\hat{\theta}_{\alpha\beta}(1-\beta)\alpha\frac{\hat{f}(\alpha\xi_\beta)}{\hat{f}(\xi_\beta)} + \beta(1-\beta)\alpha^2\frac{\hat{f}(\alpha\xi_\beta)}{\hat{f}(\xi_\beta)},$$

where $\hat{f}(.)$ is the kernel density function estimate defined in Preston (1995).

As derived earlier, the smoothed estimator $\hat{T}_n(\alpha, \beta)$ for the low income proportion $\theta_{\alpha\beta}$

is asymptotically normal with variances $\sigma^2_{\alpha\beta}$,

$$\sqrt{n}(\hat{T}_n(\alpha,\beta) - \theta_{\alpha\beta}) \longrightarrow N(0, \sigma^2_{\alpha\beta}),$$

where $\sigma^2(t)$ is defined in Theorem 2.1.

Thus, the second $(1-\alpha)$ level normal approximation (NA2)-based confidence interval for $\theta_{\alpha\beta}$ can be constructed as

$$(l_2, u_2) = (\hat{T}_n(\alpha,\beta) - \frac{z_{1-\frac{\alpha}{2}}\hat{\sigma}_{\alpha\beta}}{\sqrt{n}}, \hat{T}_n(\alpha,\beta) + \frac{z_{1-\frac{\alpha}{2}}\hat{\sigma}_{\alpha\beta}}{\sqrt{n}}),$$

where $\hat{\sigma}_{\alpha\beta}$ is the standard deviation of the variance estimate $\hat{\sigma}^2_{\alpha\beta} = \frac{\alpha^2\beta(1-\beta)\hat{f}^2(\alpha\xi_\beta))}{\hat{f}^2(\xi_\beta)} + \hat{T}_n(\alpha,\beta)(1-\hat{T}_n(\alpha,\beta))$.

2.3.2 Bootstrap-based Confidence Intervals

The normal approximate-based confidence intervals may have poor performance since the income data is skewed or has outliers. Introduced in 1979, bootstrap is a powerful non-parametric approach for constructing confidence intervals when the asymptotic variance of an estimator is unknown and of a complex form. Besides, bootstrap-based confidence interval does not rely on the parametric assumption for the data.

Since our asymptotic variances of both estimators are very complicated, we apply two bootstrap methods to estimate the asymptotic variances. Both the empirical estimator and kernel estimator will be used to construct bootstrap confidence intervals for $\theta_{\alpha\beta}$. In total, 6 bootstrap-based confidence intervals will be built in this section. Throughout this section, $\theta_{\alpha\beta}$ will be denoted by θ, and $\hat{\theta}_{\alpha\beta}$ will be denoted by $\hat{\theta}$ for simplicity.

We draw a bootstrap resample $\{X_1^*, X_2^*, X_3^*, ..., X_n^*\}$ with replacement from the original data $\{X_1, X_2, X_3, ..., X_n\}$. The bootstrap versions of the low income proportion for the empirical estimator $\hat{\theta}$ is

$$\hat{\theta}^* = \frac{1}{n}\sum_{i=1}^{n} I(X_i^* \leq \alpha\hat{\xi}_\beta^*).$$

After repeating this bootstrap procedure B times, B bootstrap copies of $\hat{\theta}$ are obtained, denoted as $\{\hat{\theta}_b^*, b = 1, 2, ..., B\}$

Then, the bootstrap sample variance of $\hat{\theta}_b^*$'s,

$$V^* = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\theta}^*)^2,$$

where $\bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$, is used to estimate the asymptotic variance of $\hat{\theta}$.

Two bootstrap confidence intervals based on the empirical estimator are constructed as follows:

1. BT1 interval:

$$(l_3, u_3) = (\hat{\theta} - z_{1-\alpha/2}\sqrt{V^*}, \hat{\theta} + z_{1-\alpha/2}\sqrt{V^*}).$$

2. BT2 interval:

$$(l_4, u_4) = (\bar{\theta}^* - z_{1-\alpha/2}\sqrt{V^*}, \bar{\theta}^* + z_{1-\alpha/2}\sqrt{V^*}).$$

Another non-parametric method to construct a confidence interval for $\theta_{\alpha\beta}$ is the bootstrap bias correction and acceleration (BCa) method, which does not need a variance estimation.

3. BCa1 interval:

$$(l_5, u_5) = (\hat{\theta}^*_{([B\beta_1])}, \hat{\theta}^*_{([B\beta_2])}).$$

where

$$\beta_1 = \Phi(b + \frac{b + z_{\alpha/2}}{1 - a(b + z_{\alpha/2})}), \beta_2 = \Phi(b + \frac{b + z_{1-\alpha/2}}{1 - a(b + z_{1-\alpha/2})})$$

with correction constants a and b defined by

$$a = \frac{1}{6}\sum_{i=1}^{n}\varphi_i^3/(\sum_{i=1}^{n}\varphi_i^2)^{\frac{3}{2}}, b = \Phi^{-1}(\frac{1}{B}\sum_{b=1}^{B}I(\hat{\theta}_b^* \leq \hat{\theta}))$$

where $\varphi_i = \hat{\theta}_{(.)} - \hat{\theta}_{(-i)}$, and $\hat{\theta}_{(-i)}$ is the $\hat{\theta}$ computed by deleting the $i-th$ observation in original data, and $\hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(-i)}$.

Similar to build bootstrap confidence intervals based on the empirical estimator $\hat{\theta}$, we construct the bootstrap confidence intervals based on the kernel estimator $\hat{T}_n(\alpha,\beta)$ as follows.

First, the bootstrap version of $\hat{T}_n(\alpha,\beta)$ is

$$\hat{T}^*(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n} K(\frac{\alpha\hat{\xi}^*_\beta - X^*_i}{h}).$$

After repeating this bootstrap procedure B times, B bootstrap copies of $\hat{T}_n(\alpha,\beta)$ are obtained, denoted as $\{\hat{T}^*_b, b = 1,2,...,B\}$.

Thus, the bootstrap sample variance of $\hat{T}^*_b$'s

$$V^*_T = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{T}^*_b - \bar{T}^*)^2,$$

where $\bar{T}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{T}^*_b$, is used to estimate the asymptotic variance of $\hat{T}_n(\alpha,\beta)$.

Similar to the bootstrap confidence intervals based on the empirical estimator for LIP, three bootstrap confidence intervals based on the kernel estimator are constructed as follows:

4. BT3 interval:

$$(l_6, u_6) = (\hat{T}_n - z_{1-\alpha/2}\sqrt{V^*_T}, \hat{T}_n + z_{1-\alpha/2}\sqrt{V^*_T}).$$

5. BT4 interval:

$$(l_7, u_7) = (\bar{T}^* - z_{1-\alpha/2}\sqrt{V^*_T}, \bar{T}^* + z_{1-\alpha/2}\sqrt{V^*_T}).$$

6. BCa2 interval:

$$(l_8, u_8) = (\hat{T}^*_{([B\beta_1])}, \hat{T}^*_{([B\beta_2])}).$$

where

$$\beta_1 = \Phi(b + \frac{b + z_{\alpha/2}}{1 - a(b + z_{\alpha/2})}), \beta_2 = \Phi(b + \frac{b + z_{1-\alpha/2}}{1 - a(b + z_{1-\alpha/2})})$$

with correction constants a and b defined by

$$a = \frac{1}{6}\sum_{i=1}^{n} \varphi_i^3 / (\sum_{i=1}^{n} \varphi_i^2)^{\frac{3}{2}}, b = \Phi^{-1}(\frac{1}{B}\sum_{b=1}^{B} I(\hat{T}_b^* \leq \hat{T}_n))$$

where $\varphi_i = \hat{T}_{(.)} - \hat{T}_{(-i)}$, and $\hat{T}_{(-i)}$ is the $\hat{T}_n$ computed by deleting the $i-th$ observation in original data, and $\hat{T}_{(.)} = \frac{1}{n}\sum_{i=1}^{n} \hat{T}_{(-i)}$.

2.3.3 Smoothed Jackknife Empirical Likelihood-based Confidence Interval

The JEL theory discussed in Section 2.2 can be used to make inference for $\theta_{\alpha\beta}$. Based on theorem 2.3, the JEL-based confidence intervals for $\theta_{\alpha\beta}$ can be constructed as

$$(l_e, u_e) = \{\theta : l_n(\theta_{\alpha\beta}) \leq \chi^2_{1,1-\alpha}\}.$$

2.4 Numerical Studies and a Real Example

In this section, we first compare the proposed kernel estimator with the empirical estimator in terms of Mean Square Error (MSE), Bias and Asymptotic Relative Efficiency(ARE). Then, we present results for the coverage probability and the average length of NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL intervals for the low income proportion discussed in previous section. The methods are also evaluated by a real example.

2.4.1 Numerical Studies

Point Estimator Evaluation. It is interesting to evaluate the performance of the kernel estimator $\hat{T}_n(\alpha, \beta)$. In this section, we are going to compare the empirical estimator $\hat{\theta}_{\alpha\beta}$ and the kernel estimator $\hat{T}_n(\alpha, \beta)$.

The first evaluation method we discuss is Mean Square Error(MSE), which is frequently used in investigating finite-sample performance of a point estimator. The MSE of the em-

pirical estimator $\hat{\theta}_{\alpha\beta}$ is

$$MSE_{\hat{\theta}_{\alpha\beta}} = E[\hat{\theta}_{\alpha\beta} - \theta_{\alpha\beta}]^2,$$

and the MSE of the kernel estimator $\hat{T}_n(\alpha, \beta)$ is

$$MSE_{\hat{T}_n(\alpha,\beta)} = E[\hat{T}_n(\alpha, \beta) - \theta_{\alpha\beta}]^2.$$

Instead of an alternative evaluation, Mean Absolute Error $E_\theta|\hat{\theta} - \theta|$, MSE is even more tractable (Casella(2002)). MSE can be composed by two parts, the square of bias, which measure the accuracy, and the variance, which measure the precision of the estimator. Minimizing MSE can achieve the balance between the bias and the variance. Next, we will compare the bias of the two estimators. Moreover, Asymptotic Relative Efficiency (ARE) is used to evaluate these two estimators.

Preston (1995) proved the asymptotical normality of the empirical estimator $\hat{\theta}_{\alpha\beta}$:

$$\sqrt{n}\left(\hat{\theta}_{\alpha\beta} - \theta_{\alpha\beta}\right) \xrightarrow{d} N(0, \sigma_v^2),$$

where

$$\sigma_v^2 = [\theta_{\alpha\beta}(1 - \theta_{\alpha\beta})]^2 - 2\theta_{\alpha\beta}(1 - \beta)\alpha\frac{f(\alpha\xi_\beta)}{f(\xi_\beta)} + \beta(1 - \beta)\alpha^2\frac{f(\alpha\xi_\beta)}{f(\xi_\beta)}.$$

We also prove the asymptotical normality of the kernel estimator in Theorem 2.1 that

$$\sqrt{n}(\hat{T}_n(\alpha, \beta) - \theta_{\alpha\beta}) \xrightarrow{d} N(0, \sigma_{\alpha\beta}^2).$$

where $\sigma_{\alpha,\beta}^2$ is defined in Theorem 2.1.

Then the ARE of $\hat{T}_n(\alpha, \beta)$ with respect to $\hat{\theta}_{\alpha\beta}$ is

$$ARE(\hat{T}_n(t), \hat{\theta}(t)) = \frac{\sigma_v^2}{\sigma_{\alpha\beta}^2}.$$

$ARE = 1$ indicates that the empirical estimator and the kernel estimator are equal efficient asymptotically. $ARE > 1$ indicates that the kernel estimator is more asymptotically efficient than the empirical estimator.

Next, we are going to evaluate the kernel estimator and empirical estimator by simulation studies. The simulation setting is as follows: The fraction α of low income proportion will be fixed at 0.5. Monte Carlo simulations are employed to simulate samples from the Chi-square distribution with degree of freedom 1. The sample sizes n are chosen to be 500, 800, and 1000. One thousand random samples are generated from the above distribution. To see how the comparisons perform across different quantiles, β =0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8 are considered. MSE and Bias will be calculated independently for kernel estimator and empirical estimator. As for ARE, we will calculate the frequency of $\{\hat{\sigma}^2_{\alpha\beta} \leq \hat{\sigma}^2_v\}$ among the 1,000 samples.

Table (2.1) lists MSE, Bias and the frequency of $\{ARE > 1\}$ for kernel estimator and empirical estimator. $Bias_{\hat{\theta}_{\alpha\beta}}$ is the bias calculated for the empirical estimator, while $Bias_{\hat{T}_n(\alpha,\beta)}$ is the bias for the kernel estimator. We observe that the proposed kernel estimator has smaller MSE than the empirical estimator, although the bias of kernel estimator is slightly larger than the bias of empirical estimator. This observation is as expected because the incorporation of kernel function introduces some bias to the kernel estimator. Meanwhile, the percentages of $\{ARE > 1\}$ are all larger than 50%, which implies that, most of time, the asymptotic variance of the kernel estimator is less than that of the empirical estimator. This point estimation comparison results show that our proposed kernel estimator is a competitive estimator.

Interval Estimation Evaluation Next, we will evaluate confidence intervals based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL under the same simulation settings used for point estimation evaluation, except that Monte Carlo samples are generated from a Chi-square distribution with degree of freedom 3, and a standard Lognormal distribution $logN(0, 1)$. 95% and 90% confidence intervals for $\theta_{\alpha\beta}$ are constructed with

Table 2.1 MSE, bias and the percentage of ARE > 1 generated from Chi-square distribution(df=1) are compared for empirical estimator and the proposed kernel estimator for low income proportion with β range from 0.2 to 0.8

Sample Size	β	$Bias_{\hat{\theta}_{\alpha\beta}}$	$Bias_{\hat{T}_n(\alpha,\beta)}$	$MSE_{\hat{\theta}_{\alpha\beta}}$	$MSE_{\hat{T}_n(\alpha,\beta)}$	ARE> 1
500	0.2	0.0000151	0.0000864	0.0000809	0.0000186	62.7%
	0.3	0.0002177	0.0000886	0.0001270	0.0000383	65.6%
	0.4	0.0005388	0.0004251	0.0001642	0.0000631	65.9%
	0.5	0.0005045	0.0012869	0.0001996	0.0000836	65.7%
	0.6	0.0005159	0.0027048	0.0002387	0.0001101	65.7%
	0.7	0.0004883	0.0041538	0.0002508	0.0001321	62.1%
	0.8	0.0018987	0.0077198	0.0002834	0.0001662	55.2%
800	0.2	0.0002088	0.0000737	0.000049	0.0000137	64.9%
	0.3	0.0003675	0.0000579	0.0000733	0.0000266	64.3%
	0.4	0.0002526	0.0002618	0.0001000	0.0000417	67.6%
	0.5	0.0006438	0.0004305	0.0001200	0.0000551	65.6%
	0.6	0.0010366	0.0007160	0.0001552	0.0000733	67.6%
	0.7	0.0014917	0.0016281	0.0001641	0.0000850	65.6%
	0.8	0.0012780	0.0048371	0.0001855	0.0001018	54.3%
1000	0.2	0.0000599	0.0000187	0.0000401	0.0000107	60.6%
	0.3	0.0002257	0.0001994	0.0000626	0.0000226	66.2%
	0.4	0.0001292	0.0004355	0.0000792	0.0000369	65.7%
	0.5	0.0002355	0.0006652	0.0000959	0.0000488	68%
	0.6	0.0005141	0.0009720	0.0001155	0.0000602	69.8%
	0.7	0.0009827	0.0017942	0.0001309	0.0000695	65.1%
	0.8	0.0002987	0.0043738	0.0001412	0.0000873	55.2%

β=0.5, 0.6, 0.7, 0.8. Totally 5 kernel functions including Uniform, Triangular, Biweight, Triweight and Epanechnikov kernel functions are compared. Triweight kernel density function $\omega(t) = \frac{35}{32}(1-t^2)^2 I(|t| \leq 1)$ is finally selected for the kernel estimator based on the simulation results, and a constant c for the bandwidth $h = cn^{-1/3}$ is chosen via the proposed cross-validation method, where c is valued differently based on different β. For the bootstrap variance estimates, 500 bootstrap samples are drawn with replacement from the original sample.

Table (2.2) to Table (2.3) display the coverage probabilities and average lengths of various confidence intervals for the low income proportion with Chi-square distribution at 90% and 95% confidence level, and Table (2.4) to Table (2.5) display coverage probabilities and average lengths for the low income proportion with Lognormal distribution at 90% and 95% confidence levels, separately.

According to the simulation tables, we observe that all the confidence intervals perform better when sample size increases. As β increases, the average length of confidence intervals increases as well. BT1 and BT2 intervals have the same interval length, while BT3 and BT4 intervals have the same interval length. Although NA2 and BCa2 intervals are observed to have the shortest average length, their coverage probabilities are far below the nominal confidence level. Out of the 9 confidence intervals, the proposed SJEL-based confidence interval is observed to achieve the best performance for better coverage probability and shorter average interval length in most cases. The proposed smoothed bootstrap-based confidence intervals (BT3 and BT4) have good finite sample performances next to the SJEL interval.

Thus, we recommend the smoothed jackknife empirical likelihood (SJEL) interval, and the smoothed bootstrap-based (BT3 and BT4) confidence intervals for low income proportion. Meanwhile, the smoothed jackknife empirical likelihood method is shown to be less computationally intensive than the plug-in empirical likelihood (EL) method dicusssed by Yang, Qin and Qin (2011).

Figure 2.3 is used to present the consistency of the empirical estimator $\hat{\theta}_{\alpha\beta}$ and the

Table 2.2 Coverage probabilities and interval lengths at 90% confidence level for LIP with Chi-square distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and β from 0.5 to 0.8.

		$\beta = 50\%$		$\beta = 60\%$		$\beta = 70\%$		$\beta = 80\%$	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
500	NA1	0.872	0.0386	0.923	0.0468	0.936	0.0551	0.926	0.0492
	NA2	0.787	0.0386	0.888	0.0264	0.866	0.0242	0.931	0.0291
	BT1	0.922	0.0498	0.920	0.0535	0.928	0.0566	0.941	0.0591
	BT2	0.948	0.0498	0.944	0.0535	0.936	0.0566	0.953	0.0591
	BT3	0.911	0.0385	0.899	0.0421	0.885	0.0445	0.897	0.0450
	BT4	0.918	0.0385	0.906	0.0421	0.897	0.0445	0.900	0.0450
	BCa1	0.876	0.0491	0.879	0.0528	0.890	0.0560	0.907	0.0584
	BCa2	0.715	0.0376	0.728	0.0414	0.703	0.0438	0.706	0.0442
	SJEL	0.908	0.0395	0.896	0.0432	0.905	0.0456	0.908	0.0461
800	NA1	0.872	0.0344	0.867	0.0394	0.864	0.0401	0.889	0.0384
	NA2	0.870	0.0283	0.855	0.0189	0.820	0.0201	0.870	0.0230
	BT1	0.910	0.0388	0.910	0.0418	0.907	0.0443	0.920	0.0461
	BT2	0.936	0.0388	0.929	0.0418	0.932	0.0443	0.934	0.0461
	BT3	0.906	0.0312	0.898	0.0340	0.911	0.0363	0.902	0.0370
	BT4	0.922	0.0312	0.905	0.0340	0.912	0.0363	0.910	0.0370
	BCa1	0.869	0.0385	0.864	0.0415	0.861	0.0439	0.897	0.0456
	BCa2	0.734	0.0307	0.763	0.0338	0.752	0.0360	0.721	0.0367
	SJEL	0.898	0.0323	0.894	0.0354	0.899	0.0373	0.904	0.0379
1000	NA1	0.878	0.0316	0.875	0.0367	0.871	0.0379	0.871	0.0405
	NA2	0.870	0.0245	0.924	0.0166	0.868	0.0178	0.864	0.0197
	BT1	0.937	0.0345	0.917	0.0372	0.917	0.0395	0.921	0.0412
	BT2	0.948	0.0345	0.926	0.0372	0.927	0.0395	0.932	0.0412
	BT3	0.894	0.0282	0.892	0.0309	0.896	0.0329	0.902	0.0336
	BT4	0.892	0.0282	0.898	0.0309	0.898	0.0329	0.910	0.0336
	BCa1	0.899	0.0342	0.869	0.0369	0.873	0.0392	0.895	0.0409
	BCa2	0.738	0.0277	0.756	0.0306	0.748	0.0326	0.732	0.0330
	SJEL	0.902	0.0293	0.904	0.0320	0.901	0.0338	0.908	0.0345

Table 2.3 Coverage probabilities and interval lengths at 95% confidence level for LIP with Chi-square distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and β from 0.5 to 0.8.

		$\beta = 50\%$		$\beta = 60\%$		$\beta = 70\%$		$\beta = 80\%$	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
500	NA1	0.892	0.0460	0.911	0.0590	0.929	0.0599	0.967	0.0648
	NA2	0.902	0.0460	0.912	0.0315	0.979	0.0288	0.935	0.0347
	BT1	0.963	0.0593	0.965	0.0638	0.954	0.0674	0.971	0.0705
	BT2	0.976	0.0593	0.979	0.0638	0.964	0.0674	0.980	0.0705
	BT3	0.951	0.0459	0.936	0.0502	0.948	0.0532	0.943	0.0535
	BT4	0.955	0.0459	0.946	0.0502	0.956	0.0532	0.950	0.0535
	BCa1	0.930	0.0585	0.930	0.0632	0.939	0.0666	0.955	0.0695
	BCa2	0.791	0.0442	0.809	0.0490	0.789	0.0517	0.785	0.0521
	SJEL	0.950	0.0470	0.945	0.0513	0.944	0.0542	0.953	0.0547
800	NA1	0.934	0.0438	0.935	0.0432	0.930	0.0440	0.920	0.0479
	NA2	0.932	0.0337	0.912	0.0225	0.916	0.0239	0.919	0.0274
	BT1	0.958	0.0463	0.956	0.0498	0.961	0.0529	0.960	0.0550
	BT2	0.974	0.0463	0.973	0.0498	0.976	0.0529	0.968	0.0550
	BT3	0.965	0.0371	0.946	0.0407	0.955	0.0433	0.946	0.0441
	BT4	0.966	0.0371	0.949	0.0407	0.958	0.0433	0.949	0.0441
	BCa1	0.932	0.0458	0.924	0.0494	0.942	0.0525	0.941	0.0543
	BCa2	0.818	0.0361	0.835	0.0399	0.829	0.0423	0.801	0.0431
	SJEL	0.947	0.0383	0.950	0.0419	0.956	0.0443	0.951	0.0450
1000	NA1	0.924	0.0392	0.932	0.0399	0.929	0.0413	0.929	0.0445
	NA2	0.904	0.0292	0.940	0.0198	0.966	0.0212	0.972	0.0235
	BT1	0.957	0.0412	0.950	0.0444	0.960	0.0470	0.964	0.0491
	BT2	0.969	0.0412	0.964	0.0444	0.964	0.0470	0.975	0.0491
	BT3	0.948	0.0336	0.932	0.0368	0.950	0.0392	0.954	0.0398
	BT4	0.950	0.0336	0.936	0.0368	0.950	0.0392	0.962	0.0398
	BCa1	0.928	0.0409	0.927	0.0441	0.943	0.0468	0.945	0.0487
	BCa2	0.794	0.0325	0.828	0.0361	0.840	0.0385	0.820	0.0384
	SJEL	0.949	0.0348	0.945	0.0379	0.950	0.0401	0.954	0.0409

Table 2.4 Coverage probabilities and interval lengths at 90% confidence level for LIP with Log-normal distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and β from 0.5 to 0.8.

		$\beta = 50\%$		$\beta = 60\%$		$\beta = 70\%$		$\beta = 80\%$	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
500	NA1	0.842	0.0484	0.888	0.0567	0.857	0.0603	0.883	0.0611
	NA2	0.833	0.0483	0.882	0.0467	0.864	0.0401	0.870	0.0304
	BT1	0.901	0.0589	0.919	0.0662	0.902	0.0720	0.905	0.0768
	BT2	0.912	0.0589	0.931	0.0662	0.919	0.0720	0.929	0.0768
	BT3	0.897	0.0500	0.912	0.0560	0.904	0.0605	0.886	0.0613
	BT4	0.905	0.0500	0.919	0.0560	0.912	0.0605	0.895	0.0613
	BCa1	0.881	0.0584	0.895	0.0657	0.877	0.0712	0.880	0.0755
	BCa2	0.778	0.0495	0.764	0.0556	0.763	0.0599	0.723	0.0603
	SJEL	0.892	0.0513	0.901	0.0572	0.892	0.0612	0.892	0.0621
800	NA1	0.869	0.0378	0.856	0.0465	0.865	0.0417	0.864	0.0536
	NA2	0.935	0.0449	0.851	0.0433	0.857	0.0378	0.842	0.0278
	BT1	0.899	0.0462	0.898	0.0517	0.915	0.0567	0.911	0.0604
	BT2	0.917	0.0462	0.920	0.0517	0.932	0.0567	0.921	0.0604
	BT3	0.952	0.0461	0.952	0.0516	0.948	0.0578	0.939	0.0596
	BT4	0.954	0.0461	0.953	0.0516	0.954	0.0578	0.946	0.0596
	BCa1	0.884	0.0460	0.888	0.0513	0.884	0.0561	0.893	0.0596
	BCa2	0.882	0.0476	0.860	0.0530	0.876	0.0571	0.813	0.0581
	SJEL	0.901	0.0414	0.900	0.0461	0.910	0.0496	0.901	0.0508
1000	NA1	0.878	0.0326	0.865	0.0307	0.871	0.0319	0.881	0.0305
	NA2	0.882	0.0336	0.828	0.0322	0.866	0.0281	0.846	0.0211
	BT1	0.937	0.0345	0.917	0.0372	0.917	0.0395	0.921	0.0412
	BT2	0.948	0.0345	0.926	0.0372	0.927	0.0395	0.932	0.0412
	BT3	0.914	0.0333	0.894	0.0405	0.890	0.0389	0.904	0.0410
	BT4	0.918	0.0333	0.902	0.0405	0.900	0.0389	0.912	0.0410
	BCa1	0.899	0.0342	0.869	0.0369	0.873	0.0392	0.895	0.0409
	BCa2	0.800	0.0362	0.808	0.0404	0.792	0.0435	0.752	0.0444
	SJEL	0.902	0.0293	0.904	0.0320	0.901	0.0338	0.908	0.0345

Table 2.5 Coverage probabilities and interval lengths at 95% confidence level for LIP with Log-normal distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and β from 0.5 to 0.8.

		$\beta = 50\%$		$\beta = 60\%$		$\beta = 70\%$		$\beta = 80\%$	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
500	NA1	0.903	0.0676	0.896	0.0757	0.930	0.0780	0.930	0.0871
	NA2	0.914	0.0575	0.901	0.0556	0.914	0.0478	0.936	0.0362
	BT1	0.945	0.0703	0.958	0.0788	0.956	0.0859	0.958	0.0914
	BT2	0.963	0.0703	0.966	0.0788	0.962	0.0859	0.966	0.0914
	BT3	0.940	0.0598	0.957	0.0669	0.950	0.0721	0.940	0.0729
	BT4	0.943	0.0598	0.963	0.0669	0.956	0.0721	0.949	0.0729
	BCa1	0.931	0.0697	0.945	0.0779	0.940	0.0850	0.943	0.0900
	BCa2	0.849	0.0590	0.842	0.0657	0.832	0.0705	0.805	0.0704
	SJEL	0.947	0.0609	0.948	0.0680	0.941	0.0728	0.944	0.0738
800	NA1	0.897	0.0451	0.890	0.0535	0.901	0.0478	0.881	0.0681
	NA2	0.905	0.0449	0.891	0.0433	0.897	0.0378	0.902	0.0278
	BT1	0.949	0.0550	0.955	0.0616	0.958	0.0673	0.954	0.0720
	BT2	0.957	0.0550	0.962	0.0616	0.967	0.0673	0.967	0.0720
	BT3	0.952	0.0481	0.952	0.0536	0.948	0.0578	0.939	0.0596
	BT4	0.954	0.0481	0.953	0.0536	0.954	0.0578	0.946	0.0596
	BCa1	0.937	0.0548	0.942	0.0612	0.940	0.0667	0.943	0.0712
	BCa2	0.882	0.0476	0.860	0.0530	0.876	0.0571	0.813	0.0581
	SJEL	0.949	0.0492	0.949	0.0547	0.951	0.0590	0.946	0.0604
1000	NA1	0.883	0.0401	0.932	0.0399	0.929	0.0413	0.929	0.0445
	NA2	0.938	0.0400	0.882	0.0384	0.882	0.0335	0.918	0.0251
	BT1	0.954	0.0492	0.950	0.0444	0.960	0.0470	0.964	0.0491
	BT2	0.959	0.0492	0.964	0.0444	0.964	0.0470	0.975	0.0491
	BT3	0.954	0.0434	0.952	0.0424	0.942	0.0423	0.956	0.0437
	BT4	0.958	0.0434	0.958	0.0424	0.950	0.0423	0.956	0.0437
	BCa1	0.940	0.0490	0.927	0.0441	0.943	0.0468	0.945	0.0487
	BCa2	0.888	0.0429	0.882	0.0479	0.852	0.0516	0.850	0.0518
	SJEL	0.949	0.0445	0.945	0.0379	0.950	0.0401	0.954	0.0409

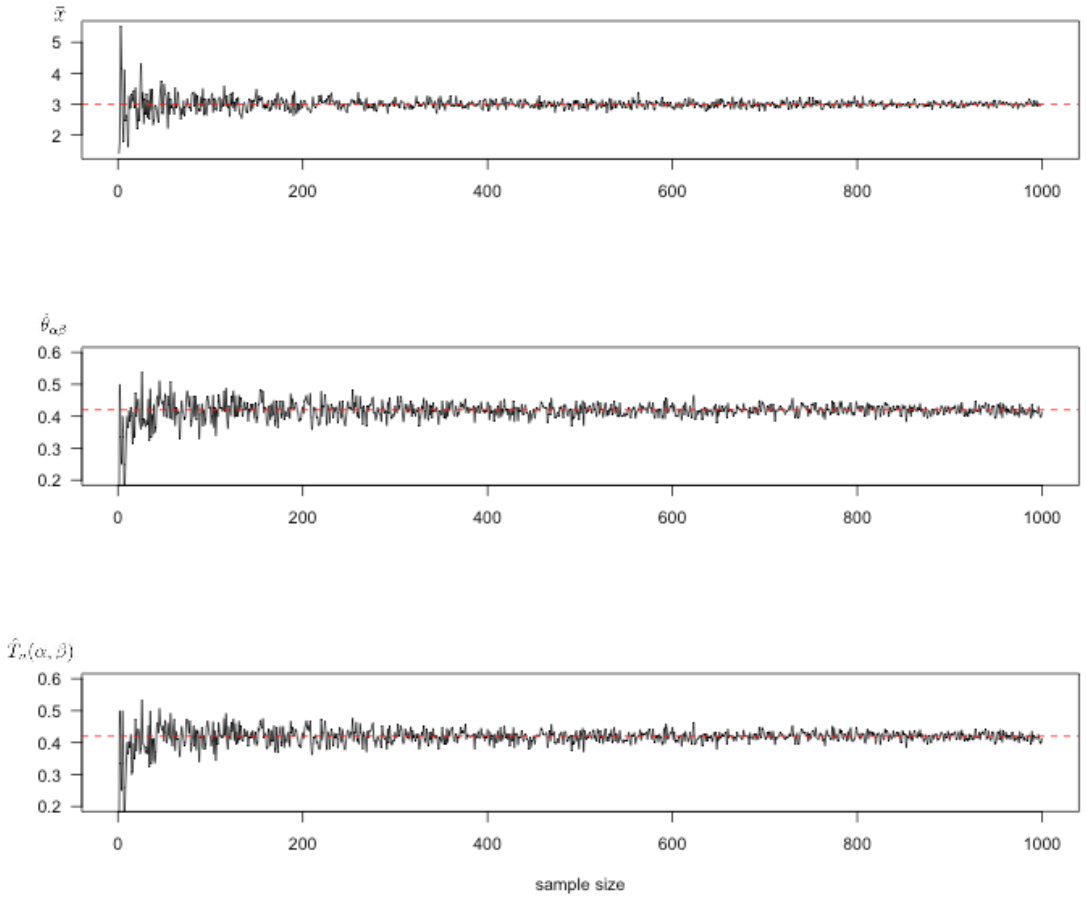

Figure 2.3 Consistency check for the empirical estimator $\hat{\theta}_{\alpha\beta}$ and the kernel estimator $\hat{T}_n(\alpha, \beta)$ for LIP

kernel estimator $\hat{T}_n(\alpha, \beta)$. We choose Chi-square distribution with degree of freedom 3. The fraction $\alpha = \frac{2}{3}$ and β =0.6. To illustrate the consistency, the first plot is the consistency check for the sample mean $\bar{x}$ vs. sample size. The population mean equals to 3 when X follows the Chi-sq distribution with degree of freedom 3. It is already known that $\bar{x}$ is a consistent estimator of the true population mean. The second plot is the empirical estimator $\hat{\theta}_{\alpha\beta}$ vs. sample size. As sample size increases, the empirical estimator will get closer to $\theta_{\alpha\beta}$. The third plot is the kernel estimator $\hat{T}_n(\alpha, \beta)$ vs. sample size. The kernel estimator is getting closer to $\theta_{\alpha\beta}$ as sample size increases.

2.4.2 Georgia Public University Employee Income Data Example

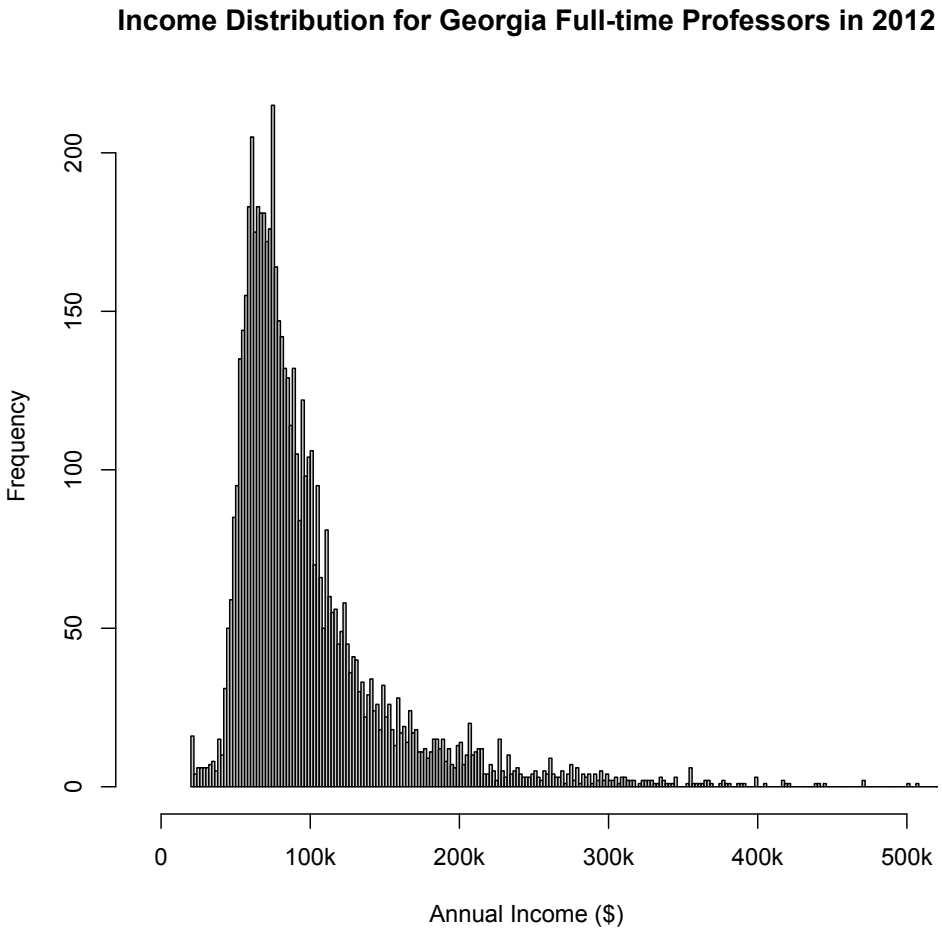

Figure 2.4 Income Distribution for Georgia full-time Professors in 2012

Georgia Department of Audits and Accounts compiled annually-updated salary information for all employees from each department, office, institution, board, commission, authority and agency of State government; every university or college in the University System of Georgia; any regional educational service agency; the General Assembly including all legislative offices and agencies; offices of the Judicial Branch; and local boards of education, etc. Each record will have ending periods in June 30, 2008, June 30, 2009, June 30, 2010, June 30, 2011 and June 30, 2012. These income data include a list of employee's name, title or functional

area, salary and travel reimbursement. The purpose of these income files is to strengthen the transparency and monitoring in government. Analysis will be based on these annual individual income data for low income proportion.

Since there is a certain percent of part-time employee or temporary employee, it will cause downward bias of income distribution and does not meet our annual salary definition. To minimize the downward bias introduced by the part-time employee or temporary employee, a homogenous income group is thus created for all faculty positions of universities and colleges in Georgia, which has a relatively evenly-distributed income. We limit the income data to all titles with Professor, Associate Professor and Assistant Professor from Units of University System and Georgia Military College from 2012 fiscal year. There are 10,332 individuals obtained initially. However, we observe some records having abnormally low salary, and we infer that those types of professors are not working full-time during 2012. It may be caused by several reasons. First of all, there are some newly-hired professors in 2012, who will not work for the whole 2012 fiscal year. After dropping those professors who don't have salary record in 2011, 9,229 observations are kept. Second of all, some professors may not be in full-time service in 2012, who may possibly either take leave or transfer to another organization. We filter this type of records out by dropping those whose income in 2012 is far less than that in 2011. Therefore, 6,195 observations are kept. Then, we drop the part-time professors whose salary is less than $20,000. In total, 5,921 observations are retained at last. By taking above steps, we create a relatively homogenous income group by retaining only full-time professors who provide full-time service during 2012 fiscal year. All the real example analysis will be based on these 5,921 observations.

We plot the histogram of 2012 annual salary for these full-time professors in Figure 2.4. It is observed that the income data is highly right skewed. Next, we present some basic statistics for annual income by job title. There are 2,266 full-time Assistant Professors recorded in 2012 with median salary $68,618 and mean salary $81,055. The number of recorded full-time Associate Professors in Georgia is 1,891. The median salary is $80,675, and mean salary is $91,760. While for the 1,764 full-time Professors, the median salary is

$109,044 and mean salary is $129,640. The maximum recorded salary for Assistant Professors, Associate Professors and Professors is $507,500, $633,260, and $949,419, respectively. The statistics for annual income by school in Georgia will be present in Appendix B.

To evaluate annual income by school, based on the 5,921 observations, University of Georgia (UGA), Georgia State University (GSU), Georgia Institute of Technology (GIT) and Georgia Health Sciences University (GHSU) are the top 4 universities that have the largest number of recorded full-time professors. According to the low wage definition (i.e., $\alpha = \frac{2}{3}$, β=0.5) by Eurostats 2012, we calculated the low wage, empirical estimator and kernel estimator of low income proportion for UGA, GSU, GIT and GHSU in Table (2.6). It is observed that the proportion of low wage earner is 8.13% in UGA, 7.34% in GSU, 13.51% in GIT, and 22.58% in GHSU. Obviously, GSU has smaller proportion of low-wage earners compared with the other three Georgia public universities.

Table 2.6 Empirical Estimator $\hat{\theta}_{\alpha\beta}$, Kernel Estimator $\hat{T}_n(\alpha, \beta)$ for Low Income Proportion and Low Wage $\alpha\xi_\beta$ defined by Eurostat 2012

Organization	Empirical Estimator	Kernel Estimator	Low Wage (Eurostat 2012)
UGA	8.161351%	8.137909%	$62,987.28
GSU	7.253086%	7.341518%	$62,188.08
GIT	13.52201%	13.51576%	$88,029.55
GHSU	22.72727%	22.58798%	$91,731.08

For the confidence interval evaluation, we fix the fraction α at $\frac{2}{3}$, and choose β equal to 0.5, 0.6, 0.7 and 0.8. The lower bound and upper bound for the proposed NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL intervals are calculated at 90% and 95% confidence levels. The final results are summarized in Table (2.7). Even the distribution of the real income data is highly right-skewed, by Central Limit Theory (CLT), when the sample size is large enough, the NA1-based confidence interval can have relatively shorter average length than most of other intervals. As β increases, the lower bound and upper bound of the confidence intervals will also increase, and the average length will increase as

well in most cases. For the same β, the average length will increase when confidence level increases. According to the low wage defined by Eurostat 2012, at 95% confidence level, in 2012, there are 8.92% to 10.8% of professors can be categorized as low-wage earners based on the SJEL interval, and 9.0% to 10.7% of professors based on BT3 interval in 2012. This findings will provide meaningful information for the government.

Table 2.7 Georgia Individual Income Example: Confidence interval and interval length at 90% and 95% confidence levels for LIP for professor's real income data are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with β from 0.5 to 0.8.

		90% Confidence level		95% Confidence level	
β	Method	Confidence Interval	Length	Confidence Interval	Length
0.5	NA1	(0.09005536, 0.1055197)	0.01546436	(0.08857408, 0.1070010)	0.01842692
	NA2	(0.08971204, 0.1080916)	0.01837951	(0.08795153, 0.1098521)	0.02190053
	BT1	(0.08969452, 0.1058806)	0.01618604	(0.08830381, 0.1072713)	0.01896746
	BT2	(0.09020321, 0.1063893)	0.01618604	(0.08861153, 0.1075790)	0.01896746
	BT3	(0.09118705, 0.1066165)	0.01542950	(0.09005919, 0.1077444)	0.01768522
	BT4	(0.09124759, 0.1066771)	0.01542950	(0.08998855, 0.1076738)	0.01768522
	BCa1	(0.08951191, 0.1064009)	0.01688904	(0.08883635, 0.1082587)	0.01942239
	BCa2	(0.08930053, 0.1045310)	0.01523051	(0.08786997, 0.1050712)	0.01720122
	SJEL	(0.09076837, 0.1067684)	0.016	(0.08924836, 0.1082484)	0.019
0.6	NA1	(0.1815626, 0.1987785)	0.01721590	(0.1799136, 0.2004276)	0.02051401
	NA2	(0.1775827, 0.2050476)	0.02746493	(0.1749519, 0.2076784)	0.03272648
	BT1	(0.1753438, 0.2049974)	0.02965359	(0.1720306, 0.2083105)	0.03627986
	BT2	(0.1761355, 0.2057891)	0.02965359	(0.1724424, 0.2087223)	0.03627986
	BT3	(0.1797684, 0.2028619)	0.02309353	(0.1772200, 0.2054103)	0.02819024
	BT4	(0.1797652, 0.2028587)	0.02309353	(0.1771416, 0.2053318)	0.02819024
	BCa1	(0.1759838, 0.2055396)	0.02955582	(0.1734504, 0.2089174)	0.03546698
	BCa2	(0.1777831, 0.2003407)	0.02255765	(0.1741475, 0.2031689)	0.02902142
	SJEL	(0.1795816, 0.2035816)	0.024	(0.1770349, 0.2060349)	0.029
0.7	NA1	(0.3040480, 0.3205086)	0.01646056	(0.3024713, 0.3220853)	0.01961397
	NA2	(0.2963817, 0.3302967)	0.03391499	(0.2931331, 0.3335453)	0.04041221
	BT1	(0.2990196, 0.3255371)	0.02651744	(0.2970666, 0.3274901)	0.03042344
	BT2	(0.2982295, 0.3247470)	0.02651744	(0.2966451, 0.3270685)	0.03042344
	BT3	(0.3004532, 0.3262252)	0.02577202	(0.2984255, 0.3282529)	0.02982737
	BT4	(0.3006397, 0.3264118)	0.02577202	(0.2987662, 0.3285936)	0.02982737
	BCa1	(0.3002871, 0.3278162)	0.02752913	(0.2975849, 0.3276474)	0.03006249
	BCa2	(0.2989087, 0.3231805)	0.02427183	(0.2958069, 0.3253234)	0.02951655
	SJEL	(0.3008217, 0.3268217)	0.026	(0.2983017, 0.3293017)	0.031
0.8	NA1	(0.4919766, 0.5034633)	0.01148670	(0.4908764, 0.5045636)	0.01368725
	NA2	(0.4781557, 0.5173900)	0.03923427	(0.4743976, 0.5211481)	0.04675052
	BT1	(0.4843998, 0.5110402)	0.02664041	(0.4813239, 0.5141161)	0.03279225
	BT2	(0.4846457, 0.5112861)	0.02664041	(0.4816849, 0.5144772)	0.03279225
	BT3	(0.4842442, 0.5113015)	0.02705726	(0.4815802, 0.5139654)	0.0323852
	BT4	(0.4842554, 0.5113126)	0.02705726	(0.4815346, 0.5139198)	0.0323852
	BCa1	(0.4818443, 0.5085290)	0.02668468	(0.4801554, 0.5117379)	0.0315825
	BCa2	(0.4829940, 0.5091359)	0.02614181	(0.4801499, 0.5123030)	0.03215309
	SJEL	(0.4837620, 0.5117620)	0.028	(0.4808233, 0.5148233)	0.034

2.5 Discussion

Development of accurate and robust measurements to estimate low income proportion and making inference for the Low Income Proportion are increasingly important. In this part, we have proposed a kernel estimator for LIP, and illustrated the selection of bandwidth for the kernel estimator by 2-fold cross-validation method. The asymptotic normality of the proposed kernel estimator is also proved in Theorem 2.1. Later, the jackknife empirical likelihood for LIP is defined, and the log-jackknife empirical likelihood ratio statistics is proved to follow the standard Chi-square distribution. To evaluate the proposed kernel estimator, we compare its MSE, Bias, and ARE with the empirical estimator. Later, the coverage probability and interval length are compared for the two normal approximation-based confidence intervals (NA1 and NA2), four bootstrap based confidence intervals (BT1, BT2, BT3, and BT4), two bootstrap bias correction and acceleration (BCa1 and BCa2) intervals and the smoothed jackknife empirical likelihood (SJEL) confidence interval. Chi-squre distribution with df=3 and the standard Lognormal distribution $log\text{N}(0,\ 1)$ are used to generated the simulation data.

Simulation studies indicate that the proposed smoothed jackknife empirical likelihood based interval which combines the power of both jackknife and empirical likelihood methods performs better than any other intervals, while the proposed confidence intervals BT3 and BT4 perform the second to the best. Lastly, thousands of real income data of Georgia Public Sector employee is used to illustrate our methods. Based on this study, we recommend the use of the proposed SJEL confidence interval for the Low Income Proportion.

2.6 Proof

Proof of Theorem 2.1, Theorem 2.2 and Theorem 2.3

Theorem 2.1. Under the conditions in Theorem 2.1, we have

$$\sqrt{n}\{\hat{T}_n(\alpha,\beta)-\theta_{\alpha\beta}\} \xrightarrow{d} N(0,\sigma^2_{\alpha\beta}).$$

Proof: We have the following decomposition

$$\begin{aligned}
&\sqrt{n}\{\hat{T}_n(\alpha,\beta)-\theta_{\alpha\beta}\}\\
&=\sqrt{n}[\frac{1}{n}\sum_{i=1}^{n}K(\frac{\alpha\hat{\xi}_\beta-X_i}{h})-F(\alpha\xi_\beta)]\\
&=\sqrt{n}[\frac{1}{n}\sum_{i=1}^{n}K(\frac{\alpha\hat{\xi}_\beta-X_i}{h})-\frac{1}{n}\sum_{i=1}^{n}K(\frac{\alpha\xi_\beta-X_i}{h})]\\
&+\sqrt{n}[\frac{1}{n}\sum_{i=1}^{n}K(\frac{\alpha\xi_\beta-X_i}{h})-F(\alpha\xi_\beta)]\\
&\equiv I_1+I_2. \qquad (2.11)
\end{aligned}$$

I_1 from (2.11) can be written as

$$\begin{aligned}
&I_1=\sqrt{n}[\int_{-\infty}^{\infty}[K(\frac{\alpha\hat{\xi}_\beta-x}{h})dF_n(x)-K(\frac{\alpha\xi_\beta-x}{h})\ dF_n(x)]\\
&=\int_{-\infty}^{\infty}[K(\frac{\alpha\hat{\xi}_\beta-x}{h})-K(\frac{\alpha\xi_\beta-x}{h})]d[\sqrt{n}(F_n(x)-F(x))]\\
&+\sqrt{n}\int_{-\infty}^{\infty}[K(\frac{\alpha\hat{\xi}_\beta-x}{h})-K(\frac{\alpha\xi_\beta-x}{h})]dF(x)\\
&\equiv I_{11}+I_{12}. \qquad (2.12)
\end{aligned}$$

Then using Taylor Series and the Bahadur's representation for sample quantile (See

Bahadur (1996))

$$\hat{\xi}_\beta - \xi_\beta = \frac{\beta - \frac{1}{n}\sum_{i=1}^{n} I(X_i \le \xi_\beta)}{f(\xi_\beta)} + o_p(n^{-\frac{1}{2}}).$$

I_{12} from (2.12) can be written as

$$\begin{aligned}
I_{12} &= \sqrt{n}\int_{-\infty}^{\infty}[K(\frac{\alpha\hat{\xi}_\beta - x}{h}) - K(\frac{\alpha\xi_\beta - x}{h})]dF(x) \\
&= \sqrt{n}\int_{-\infty}^{\infty}\omega(\frac{\alpha\xi_\beta - x}{h})\frac{\alpha\hat{\xi}_\beta - \alpha\xi_\beta}{h}dF(x) + o_p(1) \\
&= -\sqrt{n}\int_{-\infty}^{\infty}\omega(\frac{\alpha\xi_\beta - x}{h})\frac{\alpha}{h}\frac{\frac{1}{n}\sum_{i=1}^{n} I(X_i \le \xi_\beta) - \beta}{f(\xi_\beta)}dF(x) + o_p(1) \\
&= -\int_{-\infty}^{\infty}\frac{\alpha}{h}\omega(z)\frac{U_n(\beta)}{f(\xi_\beta)}dF(\alpha\xi_\beta - zh) + o_p(1) \\
&= -\frac{\alpha}{h}\int_{-\infty}^{\infty}\omega(z)\frac{U_n(\beta)}{f(\xi_\beta)}f(\alpha\xi_\beta - zh)d(\alpha\xi_\beta - zh) + o_p(1) \\
&= \frac{\alpha U_n(\beta)}{f(\xi_\beta)}\int_{-\infty}^{\infty}\omega(z)f(\alpha\xi_\beta - zh)dz + o_p(1) \\
&= \frac{\alpha f(\alpha\xi_\beta)U_n(\beta)}{f(\xi_\beta)} + o_p(1), \qquad (2.13)
\end{aligned}$$

where

$$\begin{aligned}
U_n(\beta) &= \sqrt{n}[\frac{1}{n}\sum_{i=1}^{n} I(X_i \le \xi_\beta) - \beta] \\
&= \sqrt{n}[\frac{1}{n}\sum_{i=1}^{n} I(F(X_i) \le \beta) - \beta]. \qquad (2.14)
\end{aligned}$$

Since $\sqrt{n}[F_n(x) - F(x)] \to B(x)$, which is a Gaussian Process, $\sqrt{n}(\hat{\xi}_\beta - \xi_\beta) = O_p(1)$, and $\sqrt{n}h \to \infty$, we get $I_{11} = o_p(1)$. Therefore, $I_1 = \frac{\alpha f(\alpha\xi_\beta)U_n(\beta)}{f(\xi_\beta)} + o_p(1)$.

Next, let's consider I_2 from (2.11). We are going to prove

$$EK(\frac{\alpha\xi_\beta - X}{h}) = \int_{-\infty}^{\infty} K(\frac{\alpha\xi_\beta - x}{h})]dF(x) \longrightarrow F(\alpha\xi_\beta),\ as\ h \to 0.$$

and

$$EK^2(\frac{\alpha\xi_\beta - X}{h}) = \int_{-\infty}^{\infty} K^2(\frac{\alpha\xi_\beta - x}{h})]dF(x) \longrightarrow F(\alpha\xi_\beta),\ as\ h \to 0.$$

Notice that

$$\begin{aligned}
&\lim_{h\to 0}\int_{-\infty}^{\infty} K(\frac{\alpha\xi_\beta - x}{h})f(x)dx \\
&= \lim_{h\to 0}\int_{-\infty}^{\infty}\int_{-\infty}^{\frac{\alpha\xi_\beta - x}{h}} \omega(y)dyf(x)dx \\
&= \int_{-\infty}^{\infty}[\lim_{h\to 0}\int_{-\infty}^{\frac{\alpha\xi_\beta - x}{h}} \omega(y)dy]f(x)dx \\
&= \int_{-\infty}^{\infty}\{0 * I[\alpha\xi_\beta < x] + \int_{-\infty}^{0}\omega(y)dy * I[\alpha\xi_\beta = x] + 1 * I[\alpha\xi_\beta > x]\}f(x)dx \\
&= \int_{-\infty}^{\infty} I[\alpha\xi_\beta > x]f(x)dx \\
&= \int_{-\infty}^{\infty} I[F(x) < F(\alpha\xi_\beta)]dF(x) \\
&= F(\alpha\xi_\beta) \\
&= \theta_{\alpha\beta}. \qquad (2.15)
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\lim_{h\to 0}\int_{-\infty}^{\infty} K^2(\frac{\alpha\xi_\beta - x}{h})f(x)dx \\
&= \lim_{h\to 0}\int_{-\infty}^{\infty}(\int_{-\infty}^{\frac{\alpha\xi_\beta - x}{h}} \omega(y)dy)^2 f(x)dx \\
&= \int_{-\infty}^{\infty}[\lim_{h\to 0}\int_{-\infty}^{\frac{\alpha\xi_\beta - x}{h}} \omega(y)dy]^2 f(x)dx \\
&= \int_{-\infty}^{\infty}\{0 * I[\alpha\xi_\beta < x] + \int_{-\infty}^{0}\omega(y)dy * I[\alpha\xi_\beta = x] + 1 * I[\alpha\xi_\beta > x]\}^2 f(x)dx \\
&= \int_{-\infty}^{\infty}\{I[\alpha\xi_\beta > x]\}^2 f(x)dx \\
&= \int_{-\infty}^{\infty} I[\alpha\xi_\beta > x]f(x)dx \\
&= \int_{-\infty}^{\infty} I[F(x) < F(\alpha\xi_\beta)]dF(x) \\
&= F(\alpha\xi_\beta) \\
&= \theta_{\alpha\beta}. \qquad (2.16)
\end{aligned}$$

Let $\omega_i = K(\frac{\alpha\xi_\beta - X_i}{h})$. So I_2 from (2.11) can be rewritten as

$$\begin{aligned}
I_2 \quad &= \sqrt{n}[\frac{1}{n}\sum_{i=1}^{n} K(\frac{\alpha\xi_\beta - X_i}{h}) - EK(\frac{\alpha\xi_\beta - X}{h})] + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\omega_i - E\omega_i) + o_p(1). \qquad (2.17)
\end{aligned}$$

Since $F(X_i) \overset{i.i.d.}{\sim} U(0,1)$, $U_n(\beta)$ is an ancillary statistic. By Basu's Lemma in Shao(2003), $U_n(\beta)$ is independent of $\frac{1}{n}\sum_{i=1}^{n}(\omega_i - E\omega_i)$.

Also,

$$
\begin{aligned}
& Var(\frac{\alpha f(\alpha\xi_\beta)U_n(\beta)}{f(\xi_\beta)}) + Var(\omega_i) \\
&= \frac{\alpha^2\beta(1-\beta)f^2(\alpha\xi_\beta))}{f^2(\xi_\beta)} + Var[K(\frac{\alpha\xi_\beta - x}{h})] \\
&= \frac{\alpha^2\beta(1-\beta)f^2(\alpha\xi_\beta))}{f^2(\xi_\beta)} + EK^2(\frac{\alpha\xi_\beta - x}{h}) - [EK(\frac{\alpha\xi_\beta - x}{h})]^2 \\
&\longrightarrow \frac{\alpha^2\beta(1-\beta)f^2(\alpha\xi_\beta))}{f^2(\xi_\beta)} + \theta_{\alpha\beta} - \theta_{\alpha\beta}^2 \\
&= \frac{\alpha^2\beta(1-\beta)f^2(\alpha\xi_\beta))}{f^2(\xi_\beta)} + \theta_{\alpha\beta}(1-\theta_{\alpha\beta}) \\
&\equiv \sigma_{\alpha\beta}^2. \qquad (2.18)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
I_1 + I_2 \quad &= \frac{\alpha f(\alpha\xi_\beta)U_n(\beta)}{f(\xi_\beta)} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\omega_i - E\omega_i) + o_p(1) \\
&\xrightarrow{d} N(0, \sigma_{\alpha\beta}^2). \qquad (2.19)
\end{aligned}
$$

We need Lemma 1 and Lemma 2 to prove Theorem 2.2

Lemma 1. Under the conditions in Theorem 2.1, we have

$$\sqrt{n}\{\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k(\alpha,\beta)-\theta_{\alpha\beta}\}\xrightarrow{d}N(0,\sigma^2_{\alpha\beta}). \tag{2.20}$$

where $\sigma^2_{\alpha\beta}$ is defined in Theorem 2.1.

Proof:

Define $\hat{T}_{n-1,k}(\alpha,\beta)=\frac{1}{n-1}\sum_{j\neq k}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})$.

Note that $\frac{1}{n}\sum_{k=1}^{n}V_k(\alpha,\beta)$ can be decomposed into

$$\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k(\alpha,\beta)=\frac{n-1}{n}\sum_{k=1}^{n}[\hat{T}_n(\alpha,\beta)-\hat{T}_{n-1,k}(\alpha,\beta)]+\hat{T}_n(\alpha,\beta), \tag{2.21}$$

while

$$\begin{aligned}
&\hat{T}_n(\alpha,\beta)-\hat{T}_{n-1,k}(\alpha,\beta)\\
&=\frac{1}{n}\sum_{i=1}^{n}K(\frac{\alpha\hat{\xi}_\beta-X_i}{h})-\frac{1}{n-1}\sum_{j\neq k}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})\\
&=\frac{1}{n}\sum_{i=1}^{n}K(\frac{\alpha\hat{\xi}_\beta-X_i}{h})-\frac{1}{n}\sum_{j\neq k}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})\\
&+\frac{1}{n}\sum_{j\neq k}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})-\frac{1}{n-1}\sum_{j\neq k}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h}).
\end{aligned} \tag{2.22}$$

So

$$
\begin{aligned}
&\sum_{k=1}^{n}(\hat{T}_n(\alpha,\beta)-\hat{T}_{n-1,k}(\alpha,\beta))\\
&=\frac{1}{n}[\sum_{k=1}^{n}\sum_{i=1}^{n}K(\frac{\alpha\hat{\xi}_\beta-X_i}{h})-\sum_{k=1}^{n}\sum_{j\neq k}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})]\\
&+\frac{1}{n}\sum_{k=1}^{n}\sum_{j\neq k}^{n}[K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})]-\frac{1}{n-1}\sum_{k=1}^{n}\sum_{j\neq k}^{n}[K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})]\\
&=\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}[K(\frac{\alpha\hat{\xi}_\beta-X_i}{h})-K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_i}{h})]+\frac{1}{n}\sum_{k=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_k}{h})\\
&+\frac{1}{n}\sum_{k=1}^{n}\sum_{j\neq k}^{n}[K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})]-\frac{1}{n-1}\sum_{k=1}^{n}\sum_{j\neq k}^{n}[K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})]\\
&=\{\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}[K(\frac{\alpha\hat{\xi}_\beta-X_i}{h})-K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_i}{h})]\}\\
&+\{\frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{n}[K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})]-\frac{1}{n-1}\sum_{k=1}^{n}\sum_{j=1}^{n}[K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})]\\
&+\frac{1}{n-1}\sum_{k=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_k}{h})\}\\
&\equiv I_1+I_2. \qquad (2.23)
\end{aligned}
$$

Using Talyor series, I_1 from (2.23) can be written as

$$
\begin{aligned}
I_1&=\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}[K(\frac{\alpha\hat{\xi}_\beta-X_i}{h})-K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_i}{h})]\\
&=\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}[-\omega(\frac{\alpha\hat{\xi}_\beta-X_i}{h})\frac{\alpha\hat{\xi}_{\beta,-k}-\alpha\hat{\xi}_\beta}{h}-\frac{1}{2}\omega'(\frac{\alpha\hat{\xi}_\beta-X_i}{h})(\frac{\alpha\hat{\xi}_{\beta,-k}-\alpha\hat{\xi}_\beta}{h})^2]\\
&=\frac{1}{n}\sum_{i=1}^{n}\{-\omega(\frac{\alpha\hat{\xi}_\beta-X_i}{h})\sum_{k=1}^{n}\frac{\alpha\hat{\xi}_{\beta,-k}-\alpha\hat{\xi}_\beta}{h}-\frac{1}{2}\sum_{k=1}^{n}\omega'(\frac{\alpha\hat{\xi}_\beta-X_i}{h})(\frac{\alpha\hat{\xi}_{\beta,-k}-\alpha\hat{\xi}_\beta}{h})^2\}\\
&=-\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n}\omega'(\frac{\alpha\hat{\xi}_\beta-X_i}{h})(\frac{\alpha\hat{\xi}_{\beta,-k}-\alpha\hat{\xi}_\beta}{h})^2+o_p(1). \qquad (2.24)
\end{aligned}
$$

That is because by Bahadur (1966), we have

$$
\begin{aligned}
&\hat{\xi}_{\beta,-k} - \hat{\xi}_{\beta} \\
&= (\hat{\xi}_{\beta,-k} - \xi_{\beta}) - (\hat{\xi}_{\beta} - \xi_{\beta}) \\
&= [\frac{\beta - \frac{1}{n-1}\sum_{j\neq k}^{n} I(X_j \leq \xi_{\beta})}{f(\xi_{\beta})}] - [\frac{\beta - \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq \xi_{\beta})}{f(\xi_{\beta})}] + o_p(n^{-1}) \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq \xi_{\beta}) - \frac{1}{n-1}\sum_{j\neq k}^{n} I(X_j \leq \xi_{\beta})}{f(\xi_{\beta})} + o_p(n^{-1}) \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq \xi_{\beta}) - \frac{1}{n-1}\sum_{i=1}^{n} I(X_i \leq \xi_{\beta})}{f(\xi_{\beta})} \\
&+ \frac{\frac{1}{n-1}\sum_{j=1}^{n} I(X_j \leq \xi_{\beta}) - \frac{1}{n-1}\sum_{j\neq k}^{n} I(X_j \leq \xi_{\beta})}{f(\xi_{\beta})} + o_p(n^{-1}) \\
&= -\frac{\frac{1}{n(n-1)}\sum_{i=1}^{n} I(X_i \leq \xi_{\beta}) - \frac{1}{n-1} I(X_k \leq \xi_{\beta})}{f(\xi_{\beta})} + o_p(n^{-1}) \\
&= \frac{\frac{1}{n-1}[I(X_k \leq \xi_{\beta}) - \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq \xi_{\beta})]}{f(\xi_{\beta})} + o_p(n^{-1}) \\
&= \frac{\frac{1}{n-1}[I(X_k \leq \xi_{\beta}) - F_n(\xi_{\beta})]}{f(\xi_{\beta})} + o_p(n^{-1}) \\
&= O_p(\frac{1}{n-1}) + o_p(n^{-1}) \\
&= O_p(\frac{1}{n}).
\end{aligned} \tag{2.25}
$$

And we also infer from (2.25) that $(\frac{\alpha\hat{\xi}_{\beta,-k} - \alpha\hat{\xi}_{\beta}}{h})^2 = O_p(\frac{1}{n^2h^2})$.

Under conditions of Theorem 2.1, I_1 from (2.23) will be derived to

$$
\begin{aligned}
I_1 \quad &= -\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n}\omega^{'}(\frac{\alpha\hat{\xi}_\beta - X_i}{h})(\frac{\alpha\hat{\xi}_{\beta,-k} - \alpha\hat{\xi}_\beta}{h})^2 + o_p(1) \\
&= -\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\omega^{'}(\frac{\alpha\hat{\xi}_\beta - X_i}{h})\sum_{k=1}^{n}(\frac{\alpha\hat{\xi}_{\beta,-k} - \alpha\hat{\xi}_\beta}{h})^2 \\
&= -\int_{-\infty}^{\infty}\omega^{'}(\frac{\alpha\hat{\xi}_\beta - x}{h})dF_n(x)nO_p(\frac{1}{(n-1)^2h^2}) \\
&= O_p(\frac{1}{nh^2})\left\{\frac{1}{\sqrt{n}}\int_{-\infty}^{\infty}\omega^{'}(\frac{\alpha\hat{\xi}_\beta - x}{h})d\sqrt{n}[F_n(x) - F(x)] + \int_{-\infty}^{\infty}\omega^{'}(\frac{\alpha\hat{\xi}_\beta - x}{h})dF(x)\right\} \\
&= O_p(\frac{1}{nh^2})\int_{-\infty}^{\infty}\omega^{'}(\frac{\alpha\hat{\xi}_\beta - x}{h})dF(x) \\
&= O_p(\frac{1}{nh^2})\int_{-\infty}^{\infty}\omega^{'}(y)dF(\alpha\hat{\xi}_\beta - yh) \\
&= O_p(\frac{1}{nh})\int_{-\infty}^{\infty}\omega^{'}(y)f(\alpha\hat{\xi}_\beta - yh)dy \\
&= O_p(\frac{1}{nh}). \qquad (2.26)
\end{aligned}
$$

That is because $\sqrt{n}[F_n(x) - F(x)] \to B(x)$, which is a Gaussian Process.

Meanwhile, I_2 from (2.23) can be written to

$$
\begin{aligned}
I_2 &= \sum_{k=1}^{n}[(\frac{1}{n}-\frac{1}{n-1})\sum_{j=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})+\frac{1}{n-1}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_k}{h})] \\
&= \sum_{k=1}^{n}\{\frac{-1}{n(n-1)}\sum_{j=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})+\frac{1}{n-1}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_k}{h})\} \\
&= \frac{-1}{n-1}\sum_{k=1}^{n}[\frac{1}{n}\sum_{j=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})-K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_k}{h})] \\
&= \frac{-1}{n-1}\sum_{k=1}^{n}\{[\frac{1}{n}\sum_{j=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})-\frac{1}{n}\sum_{j=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta}-X_j}{h})+\frac{1}{n}\sum_{j=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta}-X_j}{h})] \\
&-[K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_k}{h})-K(\frac{\alpha\hat{\xi}_{\beta}-X_k}{h})+K(\frac{\alpha\hat{\xi}_{\beta}-X_k}{h})]\} \\
&= \frac{-1}{n-1}\sum_{k=1}^{n}[\frac{1}{n}\sum_{j=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})-\frac{1}{n}\sum_{j=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta}-X_j}{h})] \\
&+\frac{-1}{n-1}\sum_{k=1}^{n}\frac{1}{n}\sum_{j=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta}-X_j}{h})-\frac{-1}{n-1}\sum_{k=1}^{n}[K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_k}{h})-K(\frac{\alpha\hat{\xi}_{\beta}-X_k}{h})] \\
&-\frac{-1}{n-1}\sum_{k=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta}-X_k}{h}) \\
&= \frac{-1}{n(n-1)}\sum_{k=1}^{n}\sum_{j=1}^{n}[K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})-K(\frac{\alpha\hat{\xi}_{\beta}-X_j}{h})]+\frac{-1}{n(n-1)}\sum_{k=1}^{n}\sum_{j=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta}-X_j}{h}) \\
&-\frac{-1}{n-1}\sum_{k=1}^{n}[K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_k}{h})-K(\frac{\alpha\hat{\xi}_{\beta}-X_k}{h})]-\frac{-1}{n-1}\sum_{k=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta}-X_k}{h}) \\
&= O_p(\frac{1}{n(n-1)h})-\frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{j=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta}-X_j}{h}) \\
&-O_p(\frac{1}{(n-1)^2h})+\frac{1}{n-1}\sum_{k=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta}-X_k}{h}) \\
&= -\frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{j=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta}-X_j}{h})+\frac{1}{n-1}\sum_{k=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta}-X_k}{h})+O_p(\frac{1}{n^2h}) \\
&= -\frac{1}{n-1}\sum_{j=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta}-X_j}{h})+\frac{1}{n-1}\sum_{k=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta}-X_k}{h})+O_p(\frac{1}{n^2h}) \\
&= O_p(\frac{1}{n^2h}). \qquad (2.27)
\end{aligned}
$$

From (2.26) and (2.27), we get $I_1 + I_2 = O_p(\frac{1}{nh})$, which imply that

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k(\alpha,\beta) \\
&= \frac{n-1}{n}\sum_{k=1}^{n}[\hat{T}_n(\alpha,\beta) - \hat{T}_{n-1,k}(\alpha,\beta)] + \hat{T}_n(\alpha,\beta) \\
&= \hat{T}_n(\alpha,\beta) + \frac{n-1}{n}O_p(\frac{1}{nh}) \\
&= \hat{T}_n(\alpha,\beta) + O_p(\frac{1}{nh}). \qquad (2.28)
\end{aligned}$$

Therefore

$$\begin{aligned}
&\sqrt{n}\{\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k(\alpha,\beta) - \theta_{\alpha\beta}\} \\
&= \sqrt{n}[\hat{T}_n(\alpha,\beta) - \theta_{\alpha\beta}] + O_p(\frac{1}{\sqrt{n}h}) \\
&\xrightarrow{d} N(0,\sigma^2_{\alpha\beta}). \qquad (2.29)
\end{aligned}$$

Thus, Lemma 1 holds.

Lemma 2. Under the conditions in Theorem 2.1, we have

$$\frac{1}{n}\sum_{k=1}^{n}\{\hat{V}_k(\alpha,\beta)-\theta_{\alpha\beta}\}^2 \xrightarrow{p} \sigma^2_{\alpha\beta}. \tag{2.30}$$

Proof:

Based on conclusion from Lemma 1, we can write that

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\{\hat{V}_k(\alpha,\beta)-\theta_{\alpha\beta}\}^2\\
&=\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k^2(\alpha,\beta)-2\theta_{\alpha\beta}\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k(\alpha,\beta)+\frac{1}{n}\sum_{k=1}^{n}\theta^2_{\alpha\beta}\\
&\xrightarrow{p}\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k^2(\alpha,\beta)-2\theta_{\alpha\beta}\theta_{\alpha\beta}+\frac{1}{n}n\theta^2_{\alpha\beta}\\
&=\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k^2(\alpha,\beta)-\theta^2_{\alpha\beta}.
\end{aligned} \tag{2.31}$$

As we define earlier that the jackknife pseudovalue is

$$\hat{V}_k(\alpha,\beta)=n\hat{T}_n(\alpha,\beta)-(n-1)\hat{T}_{n-1,k}(\alpha,\beta). \tag{2.32}$$

By substituting this into $\hat{V}_k(\alpha,\beta)$, we get

$$\begin{aligned}
&\hat{V}_k(\alpha,\beta)\\
&=\sum_{i=1}^{n}K(\frac{\alpha\hat{\xi}_\beta-X_i}{h})-\sum_{j\neq k}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_j}{h})\\
&=\sum_{i=1}^{n}[K(\frac{\alpha\hat{\xi}_\beta-X_i}{h})-K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_i}{h})]+K(\frac{\alpha\hat{\xi}_{\beta,-k}-X_k}{h}),
\end{aligned} \tag{2.33}$$

which imply

$$
\begin{aligned}
&\hat{V}_k^2(\alpha,\beta)\\
&=\{\sum_{i=1}^{n}[K(\frac{\alpha\hat{\xi}_\beta - X_i}{h}) - K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_i}{h})]\}^2\\
&+K^2(\frac{\alpha\hat{\xi}_{\beta,-k} - X_k}{h})\\
&+2K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_k}{h})\sum_{i=1}^{n}[K(\frac{\alpha\hat{\xi}_\beta - X_i}{h}) - K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_i}{h})]. \qquad (2.34)
\end{aligned}
$$

Then,

$$
\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k^2(\alpha,\beta)\\
&=\frac{1}{n}\sum_{k=1}^{n}\{\sum_{i=1}^{n}[K(\frac{\alpha\hat{\xi}_\beta - X_i}{h}) - K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_i}{h})]\}^2\\
&+\frac{1}{n}\sum_{k=1}^{n}K^2(\frac{\alpha\hat{\xi}_{\beta,-k} - X_k}{h})\\
&+\frac{2}{n}\sum_{k=1}^{n}K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_k}{h})\sum_{i=1}^{n}[K(\frac{\alpha\hat{\xi}_\beta - X_i}{h}) - K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_i}{h})]\\
&\equiv J_1 + J_2 + J_3. \qquad (2.35)
\end{aligned}
$$

Based on Lemma 1, we have $\hat{\xi}_{\beta,-k} - \hat{\xi}_\beta = \frac{\frac{1}{n-1}[I(x_k\le\xi_\beta) - \frac{1}{n}\sum_{i=1}^{n}I(X_i\le\xi_\beta)]}{f(\xi_\beta)} + o_p(n^{-\frac{1}{2}})$ and $\hat{\xi}_{\beta,-k} - \hat{\xi}_\beta = O_p(\frac{1}{n})$. By applying this conclusion and Taylor series, J_1 from (2.35) can be written as

$$
\begin{aligned}
J_1 &= \frac{1}{n}\sum_{k=1}^{n}\{\sum_{i=1}^{n}[K(\frac{\alpha\hat{\xi}_\beta - X_i}{h}) - K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_i}{h})]\}^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\{\sum_{i=1}^{n}[K(\frac{\alpha\hat{\xi}_\beta - X_i}{h}) - K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_i}{h})]\}^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\{\sum_{i=1}^{n}\omega(\frac{\alpha\hat{\xi}_\beta - X_i}{h})\frac{\alpha\hat{\xi}_{\beta,-k} - \alpha\hat{\xi}_\beta}{h}\}^2 + o_p(1) \\
&= \frac{1}{n}\sum_{k=1}^{n}(\frac{\alpha\hat{\xi}_{\beta,-k} - \alpha\hat{\xi}_\beta}{h})^2\{\sum_{i=1}^{n}\omega(\frac{\alpha\hat{\xi}_\beta - X_i}{h})\}^2 + o_p(1) \\
&= \frac{1}{n}\frac{\alpha^2}{h^2}\sum_{k=1}^{n}(\hat{\xi}_{\beta,-k} - \hat{\xi}_\beta)^2\{n\int_{-\infty}^{\infty}\omega(\frac{\alpha\hat{\xi}_\beta - x}{h})dF_n(x)\}^2 + o_p(1) \\
&= \frac{\alpha^2}{nh^2}\sum_{k=1}^{n}(\hat{\xi}_{\beta,-k} - \hat{\xi}_\beta)^2\{\frac{n}{\sqrt{n}}\int_{-\infty}^{\infty}\omega(\frac{\alpha\hat{\xi}_\beta - x}{h})d\sqrt{n}[F_n(x) - F(x)] \\
&+ n\int_{-\infty}^{\infty}\omega(\frac{\alpha\hat{\xi}_\beta - x}{h})dF(x)\}^2 + o_p(1) \\
&= \frac{\alpha^2}{nh^2}\sum_{k=1}^{n}(\hat{\xi}_{\beta,-k} - \hat{\xi}_\beta)^2\{n\int_{-\infty}^{\infty}\omega(\frac{\alpha\hat{\xi}_\beta - x}{h})dF(x)\}^2 + o_p(1) \\
&= \frac{n\alpha^2}{h^2}\sum_{k=1}^{n}(\hat{\xi}_{\beta,-k} - \hat{\xi}_\beta)^2[\int_{-\infty}^{\infty}\omega(z)dF(\alpha\xi_\beta - zh)]^2 + o_p(1) \\
&= \frac{n\alpha^2}{f^2(\xi_\beta)h^2}\sum_{k=1}^{n}\frac{1}{(n-1)^2}[I(x_k \le \xi_\beta)I(x_k \le \xi_\beta) - 2I(x_k \le \xi_\beta)\frac{1}{n}\sum_{i=1}^{n}I(X_i \le \xi_\beta) \\
&+ \frac{1}{n^2}\sum_{i=1}^{n}I(X_i \le \xi_\beta)\sum_{i=1}^{n}I(X_i \le \xi_\beta)][\int_{-\infty}^{\infty}\omega(z)dF(\alpha\xi_\beta - zh)]^2 + o_p(1) \\
&= \frac{n^2\alpha^2(-h)^2}{f^2(\xi_\beta)h^2(n-1)^2}[F_n(\xi_\beta) - F_n^2(\xi_\beta)][\int_{-\infty}^{\infty}\omega(z)f(\alpha\xi_\beta - zh)dz]^2 + o_p(1) \\
&= \frac{n^2\alpha^2 f^2(\alpha\xi_\beta)}{f^2(\xi_\beta)(n-1)^2}[\beta - \beta^2]1^2 + o_p(1) \\
&= \frac{n^2\alpha^2\beta(1-\beta)f^2(\alpha\xi_\beta)}{f^2(\xi_\beta)(n-1)^2} \\
&\xrightarrow{p} \frac{\alpha^2\beta(1-\beta)f^2(\alpha\xi_\beta)}{f^2(\xi_\beta)}. \qquad (2.36)
\end{aligned}
$$

By Taylor's series, J_2 from (2.35) can be written as

$$
\begin{aligned}
&J_2 \\
&= \frac{1}{n}\sum_{k=1}^{n} K^2(\frac{\alpha\hat{\xi}_{\beta,-k} - X_k}{h}) \\
&= \frac{1}{n}\sum_{k=1}^{n}[K^2(\frac{\alpha\hat{\xi}_{\beta,-k} - X_k}{h}) - K^2(\frac{\alpha\xi_\beta - X_k}{h})] + \frac{1}{n}\sum_{k=1}^{n} K^2(\frac{\alpha\xi_\beta - X_k}{h}) \\
&= \frac{1}{n}\sum_{k=1}^{n}\{2K(\frac{\alpha\xi_\beta - X_k}{h})\omega(\frac{\alpha\xi_\beta - X_k}{h})\frac{\alpha\hat{\xi}_{\beta,-k} - \alpha\xi_\beta}{h}\} - \frac{1}{2}[2\omega^2(\frac{\alpha\xi_\beta - X_k}{h})(\frac{\alpha\hat{\xi}_{\beta,-k} - \alpha\xi_\beta}{h})^2] \\
&+\frac{1}{n}\sum_{k=1}^{n} K^2(\frac{\alpha\xi_\beta - X_k}{h}) + o_p(1) \\
&= \frac{1}{n}\sum_{k=1}^{n} K^2(\frac{\alpha\xi_\beta - X_k}{h}) + o_p(1) \\
&= EK^2(\frac{\alpha\xi_\beta - x}{h}) + o_p(1) \\
&= \theta_{\alpha\beta} + o_p(1).
\end{aligned} \tag{2.37}
$$

Based on the proof of Lemma 1, we have $\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}[K(\frac{\alpha\hat{\xi}_\beta - X_i}{h}) - K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_i}{h})] = O_p(\frac{1}{nh})$.

Since $K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_k}{h})$ is a cumulative distribution function, it ranges from 0 to 1. Under

the conditions of Theorem 2.1, J_3 from (2.35) can be written as

$$
\begin{aligned}
&J_3 \\
&= \frac{2}{n}\sum_{k=1}^{n} K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_k}{h}) \sum_{i=1}^{n}[K(\frac{\alpha\hat{\xi}_{\beta} - X_i}{h}) - K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_i}{h})] \\
&\leq \frac{2}{n}\sum_{k=1}^{n} |K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_k}{h})| \sum_{i=1}^{n}[|K(\frac{\alpha\hat{\xi}_{\beta} - X_i}{h}) - K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_i}{h})|] \\
&\leq \frac{2}{n}\sum_{k=1}^{n} 1 \sum_{i=1}^{n}[|K(\frac{\alpha\hat{\xi}_{\beta} - X_i}{h}) - K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_i}{h})|] \\
&= 2\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}[|K(\frac{\alpha\hat{\xi}_{\beta} - X_i}{h}) - K(\frac{\alpha\hat{\xi}_{\beta,-k} - X_i}{h})|] \\
&= \frac{2}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}[|-\omega(\frac{\alpha\hat{\xi}_{\beta} - X_i}{h})\frac{\alpha\hat{\xi}_{\beta,-k} - \alpha\hat{\xi}_{\beta}}{h} - \frac{1}{2}\omega'(\frac{\alpha\hat{\xi}_{\beta} - X_i}{h})(\frac{\alpha\hat{\xi}_{\beta,-k} - \alpha\hat{\xi}_{\beta}}{h})^2|] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} |\omega'(\frac{\alpha\hat{\xi}_{\beta} - X_i}{h})(\frac{\alpha\hat{\xi}_{\beta,-k} - \alpha\hat{\xi}_{\beta}}{h})^2| + o_p(1) \\
&= \int_{-\infty}^{\infty} |\omega'(\frac{\alpha\hat{\xi}_{\beta} - x}{h})| dF_n(x) n O_p(\frac{1}{(n-1)^2h^2}) \\
&= O_p(\frac{1}{nh^2}) \left\{ \frac{1}{\sqrt{n}} \int_{-\infty}^{\infty} |\omega'(\frac{\alpha\hat{\xi}_{\beta} - x}{h})| d\sqrt{n}[F_n(x) - F(x)] + \int_{-\infty}^{\infty} |\omega'(\frac{\alpha\hat{\xi}_{\beta} - x}{h})| dF(x) \right\} \\
&= O_p(\frac{1}{nh^2}) \int_{-\infty}^{\infty} |\omega'(\frac{\alpha\hat{\xi}_{\beta} - x}{h})| dF(x) \\
&= O_p(\frac{1}{nh^2}) \int_{-\infty}^{\infty} |\omega'(y)| dF(\alpha\hat{\xi}_{\beta} - yh) \\
&= O_p(\frac{1}{nh}) \int_{-\infty}^{\infty} |\omega'(y)| f(\alpha\hat{\xi}_{\beta} - yh) dy \\
&= O_p(\frac{1}{nh}).
\end{aligned} \tag{2.38}
$$

Based on (2.36), (2.37) and (2.38), we have

$$
\frac{1}{n}\sum_{k=1}^{n} \hat{V}_k^2(\alpha,\beta) \xrightarrow{p} \frac{\alpha^2\beta(1-\beta)f^2(\alpha,\beta)}{f^2(\xi_\beta)} + \theta_{\alpha\beta}. \tag{2.39}
$$

In sum

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\{\hat{V}_k(\alpha,\beta)-\theta_{\alpha\beta}\}^2 \\
&\xrightarrow{p} \frac{\alpha^2\beta(1-\beta)f^2(\alpha,\beta)}{f^2(\xi_\beta)}+\theta_{\alpha\beta}-\theta_{\alpha\beta}^2 \\
&=\sigma_{\alpha\beta}^2. \qquad (2.40)
\end{aligned}$$

Proof of Theorem 2.2. It follows immediately from Lemma 1 and Lemma 2.

Proof of Theorem 2.3. According to Gong, Peng and Li (2010), define $g(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta}}{1+\lambda(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})}$. It is easy to check that

$$\begin{aligned}
0=|g(\lambda)| &= \frac{1}{n}|\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})-\lambda\sum_{i=1}^{n}\frac{(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})^2}{1+\lambda(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})}| \\
&\geq |\frac{\lambda}{n}\sum_{i=1}^{n}\frac{(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})^2}{1+\lambda(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})}|-|\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta}| \\
&\geq \frac{|\lambda|S_n}{1+|\lambda|Z_n}-|\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})|, \qquad (2.41)
\end{aligned}$$

where $S_n=\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})^2$ and $Z_n = max_{1\leq i\leq n}|\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta}|$.

From Lemma 1 and Lemma 2, we have $|\lambda|=O_p\{n^{-\frac{1}{2}}\}$.

Put $\gamma_i=\lambda(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})$, then we have $max_{1\leq i\leq n}|\gamma_i|=O_p(1)$.

$$\begin{aligned}
0=g(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})(1-\gamma_i+\frac{\gamma_i^2}{1+\gamma_i}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})-S_n\lambda+\frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})\gamma_i^2}{1+\gamma_i} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})-S_n\lambda+O_p(\frac{1}{n}), \qquad (2.42)
\end{aligned}$$

which implies that $\lambda=S_n^{-1}\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})+\beta_n$, where $\beta_n=O_p(\frac{1}{n})$.

So

$$
\begin{aligned}
& l_n(\theta_{\alpha\beta}) \\
&= -2log\ L_n(\theta_{\alpha\beta}) \\
&= 2\sum_{i=1}^{n} log\{1+\lambda(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})\} \\
&= 2\sum_{i=1}^{n}\gamma_i - \sum_{i=1}^{n}\gamma_i^2 + 2\sum_{i=1}^{n}\eta_i \\
&= 2n\lambda\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta}) - nS_n\lambda^2 + 2\sum_{i=1}^{n}\eta_i \\
&= \frac{n\{\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})\}^2}{S_n} - nS_n\beta_n^2 + 2\sum_{i=1}^{n}\eta_i \\
&= \frac{n\{\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta)-\theta_{\alpha\beta})\}^2}{S_n} + O_p(1) \\
&\xrightarrow{d} \chi^2(1). \qquad (2.43)
\end{aligned}
$$

Theorem 2.3 holds.

PART 3

LORENZ CURVE

This part is organized as follows. In Section 3.1, we first define a kernel estimator for Lorenz curve, then propose methods to choose bandwidth for the kernel estimator. Later, the empirical estimator and kernel estimator are evaluated through MSE and ARE. In Section 3.2, the jackknife pseudo-values and the jackknife empirical likelihood properties for Lorenz curve is derived. In section 3.3, normal approximation-based confidence intervals and several bootstrap-based confidence intervals for Lorenz curve are presented, together with the proposed smoothed jackknife empirical likelihood-based confidence interval. In section 3.4, numerical studies are conducted to compare the performance of the proposed estimators, as well as the proposed methods to build confidence intervals. The confidence intervals based on proposed methods are illustrated through a real example. Technique details for proof will be given at the end of this part.

3.1 Estimation of a Lorenz Curve

3.1.1 Empirical Estimator for Lorenz Curve

Lorenz Curve is defined as follows:

$$\eta(t) = \frac{1}{\mu}\int_0^{\xi_t} x dF(x), \tag{3.1}$$

where $\mu = \int_0^{\infty} x dF(x)$ is the mean of $F(x)$, and $\xi_t = F^{-1}(t)$ is the $t-th$ quantile of $F(x)$. For a fixed $t \in [0,1]$, the Lorenz ordinate $\eta(t)$ is the percentage of total income owned by wealth-holders of the lowest $t-th$ percentage of incomes.

Let $X_1, X_2, ..., X_n$ be a simple random sample drawn from the population X with c.d.f.

$F(x)$. The Lorenz ordinate η(t) satisfies

$$E[X(I(X \leq \xi_t) - \eta(t))] = 0.$$

An empirical estimate for $\eta(t)$ can be found from the following estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} X_i I(F_n(X_i) \leq t) - \eta(t)\bar{X} = 0,$$

where

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

Therefore, the empirical estimator for Lorenz Curve $\eta(t)$ is

$$\hat{\eta}(t) = \frac{1}{\bar{X}}\frac{1}{n}\sum_{i=1}^{n} X_i I(X_i \leq F_n^{-1}(t)).$$

3.1.2 A Kernel Estimator for a Lorenz Curve

$\eta(t)$ is a smoothing function in many applications. To find a smoothing estimator for $\eta(t)$, we apply the kernel method. A smoothed estimator of $\eta(t)$ may have nice properties. Lloyd and Yong (1999) proved that the kernel estimator for the ROC curve performs better than the empirical estimator for its smaller mean-square error, especially when sample size increases. In regard of the sample quantile, Yang and Shie-Shien (1985) illustrated that the smooth estimator to the conventional sample quantize function is essentially a kernel estimator, which has the same asymptotic distribution as the conventional sample quantile without smoothness. In this part, we develop a kernel method to estimate the Lorenz ordinate.

We define a kernel function $K(x) = \int_{-\infty}^{x} \omega(y)dy$, where ω is a probability density func-

tion. The smoothed estimator of the Lorenz ordinate can be constructed as follows

$$\hat{T}_n(t) = \frac{1}{n\bar{X}} \sum_{i=1}^{n} X_i K(\frac{t - F_n(X_i)}{h}). \tag{3.2}$$

We derive the asymptotic normality of the smoothed estimator $\hat{T}_n(t)$ in the following theorem.

Theorem 3.1. *Assume ω is a probability density function with bounded support and its first derivative exists on its supporting set. If $h = h(n) \to 0$, $\frac{1}{\sqrt{n}h^2} \to 0$ as $n \to \infty$, then*

$$\sqrt{n}\{\hat{T}_n(t) - \eta(t)\} \xrightarrow{d} N(0, \sigma^2(t)),$$

where $\sigma^2(t) = \frac{1}{\mu^2}[\xi_t^2 t(1-t) + (1 - 2\eta(t)) \int_0^{\xi_t} x^2 dF(x) + \eta^2(t) \int_{-\infty}^{\infty} x^2 dF(x)]$.

3.1.3 Bandwidth Selection for the Kernel Estimator by Cross-validation Method

The choice of bandwidth h is one of the most difficulties in studying kernel function related topics. Multiple methods have been used to select bandwidth. In our research, we propose a cross-validation (CV) method to choose bandwidth. In order to ease the implement, we utilize a 2-fold cross-validation method. From the previous section, the smoothed Lorenz ordinate estimate for $\eta(t)$ is as in (3.2), where h is chosen to be $h = cn^{-1/3}$, based on our simulation experience. This formula implies that the choice of h is controlled by the constant c. In the following descriptions, we denote $\hat{T}_{n,c}(t) = \hat{T}_n(t)$. For a given t, we propose a cross-validation (CV) procedure for selecting c by minimizing the Mean Squared Error(MSE),

$$MSE(c) = E[\hat{T}_{n,c}(t) - \eta(t)]^2.$$

By simple random sampling, we split the sample into two parts, where the first part is treated as the training sample, based on which we construct the smoothed estimator for Lorenz ordinate $\hat{T}^{(1)}_{n,c}(t)$. And the second part is treated as the validation sample, based on

which we construct the empirical estimator $\hat{\eta}^{(2)}(t)$. The following cross-validation estimate of the MSE is obtained by repeating the random split enough times.

$$CV_c = \frac{1}{L}\sum_{l=1}^{L}[\hat{T}_{n,c}^{(1,l)}(t) - \hat{\eta}^{(2,l)}(t)]^2,$$

where L is the number of random splits. A constant c is chosen such that CV_c is minimized.

A small value of bandwidth h will cause small bias and large variance, while a large value of h will lead to small variance in sacrificing bias. The plot of MSE vs. bandwidth can be a"smiling curve" which can be refer to Part 2. So the value of MSE is a trade-off between bias and variance; minimizing MSE can compromise these two terms.

Meanwhile if we want to study the overall performance of the kernel estimator for the Lorenz curve across all t, we can select the constant c by minimizing the Average Mean Squared Error (AMSE).

$$AMSE(c) = E\frac{1}{K}\sum_{k=1}^{K}[\hat{T}_{n,c}(t_k) - \eta(t_k)]^2, k = 1, 2, ..., K,$$

where t_k is in a fine grid of (0,1), K is an integer.

Similar to MSE, the cross-validation estimate of the AMSE is

$$ACV_c = \frac{1}{L}\frac{1}{K}\sum_{l=1}^{L}\sum_{k=1}^{K}[\hat{T}_{n,c}^{(1,l)}(t_k) - \hat{\eta}^{(2,l)}(t_k)]^2.$$

In our study, constant c is chosen by using the cross-validation method to minimize AMSE.

3.1.4 Point Estimator Evaluation

In this section, we are going to compare the empirical estimator $\hat{\eta}(t)$ and the smoothed estimator $\hat{T}_n(t)$. There are several point estimator evaluation methods including MSE, best unbiased estimator, sufficiency and unbiasedness, etc. To ease the implementation and explanation, Mean Square Error(MSE) is chosen as our first evaluation method.

The MSE of the empirical estimator $\hat{\eta}(t)$ is

$$MSE_{\hat{\eta}} = E[\hat{\eta} - \eta(t)]^2,$$

and the MSE of the kernel estimator $\hat{\eta}(t)$ is

$$MSE_{\hat{T}_n(t)} = E[\hat{T}_n(t) - \eta(t)]^2.$$

The second method we use to evaluate these two estimators is the Asymptotic Relative Efficiency (ARE). Based on Zheng (2002), the empirical estimator satisfies

$$\sqrt{n}\,(\hat{\eta}(t) - \eta(t)) \xrightarrow{d} N(0, \sigma_v^2),$$

where $\sigma_v^2 = \int_0^\infty [(x - \xi_{t_0})I(x \leq \xi_{t_0}) - x\eta]^2 dF(x) - (t_0\xi_{t_0})^2$.

And we proved in Theorem 3.1 that the kernel estimator satisfies

$$\sqrt{n}(\hat{T}_n(t) - \eta(t)) \xrightarrow{d} N(0, \sigma^2(t)).$$

Then the ARE of $\hat{T}_n(t)$ with respect to $\hat{\eta}(t)$ is

$$ARE(\hat{T}_n(t), \hat{\eta}(t)) = \frac{\sigma_v^2}{\sigma^2(t)}.$$

$ARE > 1$ indicates that the kernel estimator $\hat{T}_n(t)$ is more efficient than the empirical estimator $\hat{\eta}(t)$.

3.2 Smoothed Jackknife Empirical Likelihood for a Lorenz Curve

Based on Tukey (1958) who used the jackknife method to estimate the variance, the jackknife pseudo-values for Lorenz curve can be defined as

$$\hat{V}_i(t) = n\hat{T}_n(t) - (n-1)\hat{T}_{n-1,i}(t), \tag{3.3}$$

where $\hat{T}_{n-1,i}(t) = \frac{1}{(n-1)\bar{X}_{n-1,i}} \sum_{j\neq i} X_j K(\frac{t-F_{n,i}(t)}{h})$ is the given statistics T_{n-1} but computed on $n-1$ observations $X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_n$, $i = 1, ..., n$, $F_{n,i}(t) = \frac{1}{n-1} \sum_{j\neq i} I(X_j \leq t)$ and $\bar{X}_{n-1,i} = \frac{1}{n-1} \sum_{j\neq i}^{n} X_j$ is the sample mean based on these $n-1$ observations.

Next, we define the jackknife empirical likelihood for $\eta = \eta(t)$ as

$$L(t, \eta) = \sup \left\{ \prod_{i=1}^{n} p_i : \sum_{i=1}^{n} np_i = 1, \sum_{i=1}^{n} p_i \hat{V}_i(t) = \eta \right\}. \tag{3.4}$$

By using the Lagrange multiplier method, the maximization for (3.4) is obtained at

$$p_i = \frac{1}{n}\{1 + \lambda[\hat{V}_i(t) - \eta]\}^{-1}, \tag{3.5}$$

where $\lambda = \lambda(t, \eta)$ is the solution to

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\hat{V}_i(t) - \eta}{1 + \lambda(\hat{V}_i(t) - \eta)} = 0. \tag{3.6}$$

Clearly $\prod_{i=1}^{n} p_i$ is subject to $\sum_{i=1}^{n} p_i = 1$, $p_i \geq 0$, $i = 1, 2, ..., n$. Because of $L(t, \eta)$ attains its maximization n^{-n} at $p_i = n^{-1}$, the jackknife empirical likelihood ratio for η can be defined as

$$L_n(\eta(t)) = \prod_{i=1}^{n} (np_i) = \prod_{i=1}^{n} \{1 + \lambda(\hat{V}_i(t) - \eta)\}^{-1}, \tag{3.7}$$

which gives the log jackknife empirical likelihood ratio as

$$l_n(\eta(t)) = -2 \log L_n(\eta(t)) = 2 \sum_{i=1}^{n} \log\{1 + \lambda(\hat{V}_i(t) - \eta)\}. \tag{3.8}$$

The jackknife variance estimator for $\hat{T}_n(t)$ is defined as follows

$$\upsilon_{JACK}(t) = \frac{1}{n(n-1)} \sum_{i=1}^{n} [\hat{V}_i(t) - \frac{1}{n} \sum_{j=1}^{n} \hat{V}_j(t)]^2 = \frac{n-1}{n} \sum_{i=1}^{n} [\hat{T}_{n-1,i}(t) - \frac{1}{n} \sum_{j=1}^{n} \hat{T}_{n-1,j}(t)]^2. \tag{3.9}$$

We derive the following theorems to show that the jackknife variance estimator is a consistent estimator for $\sigma^2(t)$.

Theorem 3.2. *Under conditions of Theorem 3.1, we have*

$$\upsilon_{JACK}(t) \xrightarrow{p} \sigma^2(t), \tag{3.10}$$

where $\sigma^2(t)$ is defined in Theorem 3.1

Theorem 3.3. *Under the conditions of Theorem 3.1, we have*

$$l_n(\eta(t)) \xrightarrow{d} \chi^2(1). \tag{3.11}$$

3.3 Confidence Intervals for a Lorenz Curve

In this section, we compare confidence intervals for $\eta(t)$ including the most commonly used normal approximation-based confidence intervals, the proposed bootstrap-based confidence intervals, and the proposed jackknife empirical likelihood-based confidence interval.

3.3.1 Normal Approximation-based Confidence Intervals

The first confidence interval we are going to discuss here is a normal approximation-based confidence interval, which is a very popular interval estimation method for an unknown parameter.

Zheng (2002) showed that the empirical estimator $\hat{\eta}(t)$ for Lorenz curve $\eta(t)$ is asymptotically normal with variances σ_v^2 , i.e.,

$$\sqrt{n}\,(\hat{\eta}(t) - \eta(t)) \longrightarrow N(0, \sigma_v^2),$$

where $\sigma_v^2 = \int_0^\infty [(x - \xi_{t_0})I(x \le \xi_{t_0}) - x\eta]^2 dF(x) - (t_0\xi_{t_0})^2$.

Therefore, based on the empirical estimator, the first $(1-\alpha)$ level normal approximate

(NA1)-based confidence interval for $\eta(t)$ can be constructed as

$$(l_1, u_1) = (\hat{\eta}(t) - \frac{z_{1-\frac{\alpha}{2}}\hat{\sigma}_v}{\sqrt{n}}, \hat{\eta}(t) + \frac{z_{1-\frac{\alpha}{2}}\hat{\sigma}_v}{\sqrt{n}}),$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})-th$ quantile of the standard normal distribution, and

$$\hat{\sigma}_v^2 = \int_0^\infty [(x - \hat{\xi}_{t_0})I(x \leq \hat{\xi}_{t_0}) - x\hat{\eta}]^2 dF_n(x) - (t_0\hat{\xi}_{t_0})^2$$

is a consistent estimate of σ_v^2.

Based on Theorem 3.1, the kernel estimator $\hat{T}_n(t)$ for the Lorenz curve $\eta(t)$ is asymptotically normal with variances $\sigma^2(t)$, i.e.,

$$\sqrt{n}(\hat{T}_n(t) - \eta(t)) \longrightarrow N(0, \sigma^2(t)),$$

where $\sigma^2(t)$ is defined in Theorem 3.1.

Therefore, the second $(1-\alpha)$ level normal approximate (NA2) based confidence interval for $\eta(t)$ built on the smoothed estimator can be constructed as

$$(l_2, u_2) = (\hat{T}_n(t) - \frac{z_{1-\frac{\alpha}{2}}\hat{\sigma}(t)}{\sqrt{n}}, \hat{T}_n(t) + \frac{z_{1-\frac{\alpha}{2}}\hat{\sigma}(t)}{\sqrt{n}}),$$

where

$$\hat{\sigma}^2(t) = \frac{1}{\bar{X}^2}\{\hat{\xi}_t^2 t(1-t) + (1 - 2\hat{T}_n(t)) \int_0^{\hat{\xi}_t} x^2 dF_n(x) + \hat{T}_n^2(t) \int_{-\infty}^{\infty} x^2 dF_n(x)\}.$$

3.3.2 Bootstrap-based Confidence Intervals

For most of interval estimations, confidence intervals are mainly built based on the parametric distribution assumption of the data. However, the parametric assumptions such as Gaussian distribution may not work well because income data is usually skewed or has outliers. Bootstrapping has gradually become a computationally intensive statistical tech-

nique for constructing confidence intervals without assumption of parametric distribution for the sample data. As in our case, we have complicated variance estimates based on both the empirical estimator and the kernel estimator, so we will develop multiple bootstrap methods to estimate the asymptotic variance and thus construct confidence intervals. Both the empirical estimator and smoothed estimator will be used to construct the bootstrap-based confidence intervals for the Lorenz curve $\eta(t)$.

First of all, the bootstrap resample $\{X_1^*, X_2^*, X_3^*, ..., X_n^*\}$ is drawn from the original data $\{X_1, X_2, X_3, ..., X_n\}$ with replacement. The bootstrap versions of the empirical estimator for Lorenz curve is

$$\hat{\eta}^* = \frac{\sum_{i=1}^n X_i^* I(X_i^* \leq \xi_{t_0}^*)}{\sum_{i=1}^n X_i^*}.$$

By repeating this bootstrap procedure for B times, B bootstrap copies of $\hat{\eta}$ are obtained. We denoted them as $\{\hat{\eta}_b^*, b = 1, 2, ..., B\}$. Thus, the bootstrap sample variance of $\hat{\eta}_b^*$'s

$$V_L^* = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\eta}_b^* - \bar{\eta}^*)^2,$$

where $\bar{\eta}^* = \frac{1}{B} \sum_{b=1}^B \hat{\eta}_b^*$, is used to estimate the asymptotic variance of $\hat{\eta}$.

Two bootstrap confidence intervals for $\eta(t)$ based on the empirical estimator are constructed as follows:

1. BT1 interval:

$$(l_3, u_3) = (\hat{\eta} - z_{1-\alpha/2}\sqrt{V_L^*}, \hat{\eta} + z_{1-\alpha/2}\sqrt{V_L^*}),$$

2. BT2 interval:

$$(l_4, u_4) = (\bar{\eta}^* - z_{1-\alpha/2}\sqrt{V_L^*}, \bar{\eta}^* + z_{1-\alpha/2}\sqrt{V_L^*}).$$

The next method to construct a confidence interval for $\eta(t)$ is the bootstrap bias correction and acceleration (BCa1) method, which does not need a variance estimation.

3. BCa1 interval:

$$(l_5, u_5) = (\hat{\eta}^*_{([B\beta_1])}, \hat{\eta}^*_{([B\beta_2])}).$$

where

$$\beta_1 = \Phi(b + \frac{b + z_{\alpha/2}}{1 - a(b + z_{\alpha/2})}), \beta_2 = \Phi(b + \frac{b + z_{1-\alpha/2}}{1 - a(b + z_{1-\alpha/2})})$$

with correction constants a and b defined by

$$a = \frac{1}{6}\sum_{i=1}^{n}\varphi_i^3/(\sum_{i=1}^{n}\varphi_i^2)^{\frac{3}{2}}, b = \Phi^{-1}(\frac{1}{B}\sum_{b=1}^{B} I(\hat{\eta}^*_b \leq \hat{\eta}))$$

where $\varphi_i = \hat{\eta}_{(.)} - \hat{\eta}_{(-i)}$, and $\hat{\eta}_{(-i)}$ is the $\hat{\eta}$ computed by deleting the $i - th$ observation in original data, and $\hat{\eta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_{(-i)}$.

Next, the bootstrap version of the kernel estimator for the Lorenz curve $\eta(t)$ is

$$\hat{T}^*(t) = \frac{1}{n\bar{X}^*}\sum_{i=1}^{n} X_i^* K(\frac{t - F_n(X_i^*)}{h}).$$

After repeating this bootstrap procedure for B times, B bootstrap copies of $\hat{T}_n$ are obtained, denoted as $\{\hat{T}^*_b, b = 1, 2, ..., B\}$.

Thus, the bootstrap sample variance of $\hat{T}^*_b$'s is used to estimate the asymptotic variance of $\hat{T}_n(t)$,

$$V^*_{LT} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{T}^*_b - \bar{T}^*)^2,$$

where $\bar{T}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{T}^*_b$.

Similarly, three bootstrap confidence intervals based on the kernel estimator for Lorenz curve $\eta(t)$ are constructed as follows:

4. BT3 interval:

$$(l_6, u_6) = (\hat{T}_n - z_{1-\alpha/2}\sqrt{V^*_{LT}}, \hat{T}_n + z_{1-\alpha/2}\sqrt{V^*_{LT}}),$$

5. BT4 interval:

$$(l_7, u_7) = (\bar{T}^* - z_{1-\alpha/2}\sqrt{V^*_{LT}}, \bar{T}^* + z_{1-\alpha/2}\sqrt{V^*_{LT}}).$$

6. BCa2 interval:

$$(l_8, u_8) = (\hat{T}^*_{([B\beta_1])}, \hat{T}^*_{([B\beta_2])}).$$

where

$$\beta_1 = \Phi(b + \frac{b + z_{\alpha/2}}{1 - a(b + z_{\alpha/2})}), \beta_2 = \Phi(b + \frac{b + z_{1-\alpha/2}}{1 - a(b + z_{1-\alpha/2})})$$

with correction constants a and b defined by

$$a = \frac{1}{6}\sum_{i=1}^{n}\varphi_i^3/(\sum_{i=1}^{n}\varphi_i^2)^{\frac{3}{2}}, b = \Phi^{-1}(\frac{1}{B}\sum_{b=1}^{B} I(\hat{T}^*_b \leq \hat{T}_n))$$

where $\varphi_i = \hat{T}_{(.)} - \hat{T}_{(-i)}$, and $\hat{T}_{(-i)}$ is the $\hat{T}_n$ computed by deleting the $i-th$ observation in original data, and $\hat{T}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{T}_{(-i)}$.

3.3.3 Smoothed Jackknife Empirical Likelihood-based Confidence Interval

Based on the jackknife empirical likelihood theory discussed in Section 3.2, we can make inference for Lorenz curve (LC). According to Theorem 3.3, the smoothed jackknife empirical likelihood (SJEL)-based confidence interval for $\eta(t)$ can be constructed as

$$(l_e, u_e) = \{\eta : l_n(\eta(t)) \leq \chi^2_{1,1-\alpha}\}.$$

3.4 Numerical Studies and a Real Example

The proposed smoothed estimator will be compared with the empirical estimator in this section. Then, the coverage probabilities and interval lengths are evaluated for NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL intervals for Lorenz curve. Later, a real example is used to illustrate the proposed methods.

3.4.1 Numerical Studies

In the numerical studies, the empirical estimator $\hat{\eta}(t)$ and the kernel estimator $\hat{T}_n(t)$ are compared in terms of MSE and ARE. Next, the simulation results for the coverage probabilities and the average interval lengths of the proposed confidence intervals for the Lorenz ordinate are presented and discussed.

One thousand random samples are generated from the Chi-square distribution with degree of freedom 3 under the following settings. The sample size is selected to be 100, 200, and 500, respectively, and t is chosen to be 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8. Then, Table (3.1) contains the comparison of the two estimators. $Bias_{\hat{\eta}}$ is the bias calculated for the empirical estimator, and $Bias_{\hat{T}_n(t)}$ is the bias for the kernel estimator. $\{ARE > 1\}$ indicates the frequency of $\hat{\sigma}^2(t) \leq \hat{\sigma}_v^2$ based on the 1,000 samples.

Table (3.1) shows that, even with a slightly larger bias, the kernel estimator $\hat{T}_n(t)$ has a smaller MSE than the empirical estimator $\hat{\eta}(t)$. At the same time, the percentage of $\{ARE > 1\}$ is all larger than 50 % across all different sample sizes, which means, in most cases, the kernel estimator has a smaller variance than the empirical estimator. Clearly, the proposed kernel estimator is able to compete the empirical estimator.

For the interval estimation evaluation, the same simulation settings will be used, except that samples are generated from the Chi-square distribution with degree of freedom 3, and the Wellbull distribution with shape=1 and scale=2. We construct the 95% and 90% confidence intervals for Lorenz curve $\eta(t)$ with t=0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Five kernel density functions including Uniform, Triangular, Biweight, Triweight and Epanechnikov are compared and finally the Quartic/Triweight kernel density function $\omega(x) = \frac{35}{32}(1-t^2)^2 I(|t| \leq 1)$ is chosen for the kernel estimator, and the constant c for bandwidth $h = cn^{-1/3}$ is selected via the proposed cross-validation method, where c are valued differently based on different quantiles. For the bootstrap variance estimate, 500 bootstrap re-samples are drawn from the original sample based on $F(x)$.

First of all, we calculate the coverage probabilities and average lengths for the Lorenz curve with Chi-square distribution, for t from 0.1 to 0.9, shown in Table (3.2) and Table

Table 3.1 MSE, bias and the percentage of $ARE > 1$ generated from Chi-square distribution(df=3) are compared for empirical estimator and the proposed smoothed estimator for LC with t ranges from 0.2 to 0.8

Sample Size	t	$Bias_{\hat{\eta}}$	$Bias_{\hat{T}_n(t)}$	$MSE_{\hat{\eta}}$	$MSE_{\hat{T}_n(t)}$	ARE> 1
100	0.2	0.0007988	0.0000303	0.0000457	0.0000420	63.3%
	0.3	0.0012336	0.0002300	0.0001167	0.0001095	57.8%
	0.4	0.0015107	0.0023934	0.0002197	0.0002090	67.0%
	0.5	0.0016503	0.0023045	0.0003441	0.0003266	71.7%
	0.6	0.0018173	0.0024518	0.0004751	0.0004490	72.4%
	0.7	0.0021618	0.0031213	0.0005804	0.0005455	73.2%
	0.8	0.0023247	0.0052457	0.0006236	0.0005741	62.4%
200	0.2	0.0008439	0.0005584	0.0000207	0.0000194	70.2%
	0.3	0.0011933	0.0005786	0.0000510	0.0000486	60.1%
	0.4	0.0014810	0.0024493	0.0000945	0.0000941	50.3%
	0.5	0.0017496	0.0026769	0.0001430	0.0001418	50.5%
	0.6	0.0020858	0.0030494	0.0001919	0.0001900	52.8%
	0.7	0.0021838	0.0035613	0.0002383	0.0002350	51.5%
	0.8	0.0020453	0.0047925	0.0002670	0.0002655	55.1%
500	0.2	0.0002354	0.0002046	0.0000079	0.0000078	73.6%
	0.3	0.0004232	0.0002602	0.0000200	0.0000197	59.3%
	0.4	0.0006112	0.0013564	0.0000385	0.0000392	51.4%
	0.5	0.0007637	0.0015491	0.0000627	0.0000634	53.3%
	0.6	0.0009519	0.0018398	0.0000911	0.0000916	53.4%
	0.7	0.0010729	0.0022663	0.0001160	0.0001170	53.1%
	0.8	0.0011004	0.0030915	0.0001269	0.0001300	51.2%

Table 3.2 Coverage probabilities and interval lengths at 90% confidence level for LC with Chi-square distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.1 to 0.4

		t=10%		t=20%		t=30%		t=40%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.880	0.0096	0.884	0.0213	0.891	0.0338	0.896	0.0462
	NA2	0.852	0.0092	0.875	0.0204	0.850	0.0329	0.932	0.0447
	BT1	0.950	0.0116	0.932	0.0239	0.923	0.0370	0.926	0.0508
	BT2	0.901	0.0116	0.892	0.0239	0.896	0.0370	0.895	0.0508
	BT3	0.927	0.0086	0.920	0.0200	0.919	0.0315	0.899	0.0436
	BT4	0.895	0.0086	0.905	0.0200	0.895	0.0315	0.895	0.0436
	BCa1	0.915	0.0097	0.914	0.0221	0.919	0.0352	0.928	0.0488
	BCa2	0.857	0.0087	0.890	0.0201	0.880	0.0317	0.892	0.0438
	SJEL	0.907	0.0104	0.908	0.0216	0.897	0.0340	0.900	0.0465
200	NA1	0.882	0.0066	0.906	0.0149	0.904	0.0237	0.905	0.0327
	NA2	0.862	0.0066	0.902	0.0147	0.935	0.0235	0.910	0.0321
	BT1	0.920	0.0073	0.920	0.0159	0.912	0.0248	0.917	0.0343
	BT2	0.890	0.0073	0.907	0.0159	0.905	0.0248	0.907	0.0343
	BT3	0.916	0.0063	0.910	0.0144	0.905	0.0228	0.902	0.0315
	BT4	0.895	0.0063	0.902	0.0144	0.912	0.0228	0.897	0.0315
	BCa1	0.901	0.0067	0.919	0.0153	0.908	0.0243	0.907	0.0338
	BCa2	0.875	0.0063	0.892	0.0145	0.895	0.0230	0.890	0.0316
	SJEL	0.915	0.0075	0.919	0.0156	0.911	0.0244	0.912	0.0334
500	NA1	0.883	0.0041	0.904	0.0093	0.901	0.0149	0.891	0.0206
	NA2	0.912	0.0041	0.917	0.0093	0.930	0.0149	0.922	0.0205
	BT1	0.899	0.0043	0.914	0.0096	0.904	0.0152	0.901	0.0211
	BT2	0.888	0.0043	0.903	0.0096	0.908	0.0152	0.906	0.0211
	BT3	0.901	0.0040	0.907	0.0092	0.904	0.0147	0.903	0.0203
	BT4	0.910	0.0040	0.905	0.0092	0.903	0.0147	0.905	0.0203
	BCa1	0.883	0.0041	0.907	0.0094	0.903	0.0151	0.900	0.0210
	BCa2	0.900	0.0040	0.902	0.0092	0.905	0.0146	0.917	0.0203
	SJEL	0.904	0.0049	0.910	0.0103	0.910	0.0159	0.907	0.0216

Table 3.3 Coverage probabilities and interval lengths at 90% confidence level for LC with Chi-square distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.5 to 0.9.

		t=50%		t=60%		t=70%		t=80%		t=90%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.884	0.0584	0.892	0.0688	0.883	0.0765	0.888	0.0791	0.855	0.0695
	NA2	0.877	0.0551	0.889	0.0641	0.867	0.0710	0.855	0.0697	0.872	0.0497
	BT1	0.918	0.0639	0.914	0.0756	0.919	0.0858	0.932	0.0919	0.954	0.0916
	BT2	0.888	0.0639	0.890	0.0756	0.890	0.0858	0.880	0.0919	0.875	0.0916
	BT3	0.882	0.0566	0.885	0.0671	0.888	0.0727	0.882	0.0744	0.880	0.0645
	BT4	0.892	0.0566	0.895	0.0671	0.897	0.0727	0.882	0.0744	0.879	0.0645
	BCa1	0.918	0.0621	0.902	0.0738	0.920	0.0847	0.917	0.0902	0.924	0.0871
	BCa2	0.890	0.0566	0.877	0.0670	0.875	0.0721	0.872	0.0731	0.815	0.0616
	SJEL	0.893	0.0589	0.894	0.0700	0.895	0.0777	0.903	0.0807	0.886	0.0715
200	NA1	0.906	0.0412	0.901	0.0487	0.884	0.0543	0.876	0.0563	0.888	0.0507
	NA2	0.867	0.0400	0.862	0.0470	0.867	0.0525	0.867	0.0534	0.857	0.0455
	BT1	0.921	0.0430	0.917	0.0511	0.913	0.0577	0.911	0.0608	0.942	0.0586
	BT2	0.899	0.0430	0.890	0.0511	0.888	0.0577	0.876	0.0608	0.891	0.0586
	BT3	0.895	0.0403	0.882	0.0477	0.887	0.0525	0.887	0.0543	0.885	0.0473
	BT4	0.857	0.0403	0.892	0.0477	0.907	0.0525	0.894	0.0543	0.897	0.0473
	BCa1	0.907	0.0426	0.912	0.0511	0.910	0.0578	0.907	0.0612	0.915	0.0590
	BCa2	0.860	0.0403	0.850	0.0477	0.840	0.0523	0.850	0.0538	0.862	0.0461
	SJEL	0.903	0.0421	0.901	0.0498	0.892	0.0556	0.895	0.0580	0.909	0.0532
500	NA1	0.879	0.0260	0.898	0.0308	0.897	0.0344	0.894	0.0358	0.900	0.0325
	NA2	0.897	0.0257	0.900	0.0303	0.932	0.0339	0.912	0.0351	0.922	0.0319
	BT1	0.889	0.0265	0.902	0.0313	0.900	0.0352	0.904	0.0369	0.914	0.0345
	BT2	0.878	0.0265	0.890	0.0313	0.896	0.0352	0.892	0.0369	0.893	0.0345
	BT3	0.892	0.0257	0.897	0.0305	0.901	0.0338	0.903	0.0351	0.907	0.0315
	BT4	0.892	0.0257	0.890	0.0305	0.907	0.0338	0.909	0.0351	0.911	0.0315
	BCa1	0.895	0.0264	0.905	0.0314	0.914	0.0353	0.910	0.0373	0.920	0.0352
	BCa2	0.892	0.0257	0.900	0.0305	0.887	0.0338	0.895	0.0350	0.895	0.0312
	SJEL	0.893	0.0265	0.908	0.0310	0.901	0.0350	0.901	0.0365	0.903	0.0344

Table 3.4 Coverage probabilities and interval lengths at 95% confidence level for LC with Chi-square distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.1 to 0.4

		t=10%		t=20%		t=30%		t=40%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.929	0.0114	0.938	0.0254	0.945	0.0403	0.950	0.0551
	NA2	0.915	0.0109	0.923	0.0243	0.930	0.0392	0.932	0.0533
	BT1	0.979	0.0138	0.977	0.0286	0.973	0.0442	0.974	0.0605
	BT2	0.947	0.0138	0.945	0.0286	0.948	0.0442	0.944	0.0605
	BT3	0.934	0.0103	0.958	0.0238	0.938	0.0375	0.945	0.0518
	BT4	0.948	0.0103	0.946	0.0238	0.948	0.0375	0.952	0.0518
	BCa1	0.951	0.0114	0.967	0.0264	0.963	0.0420	0.973	0.0582
	BCa2	0.935	0.0103	0.937	0.0239	0.935	0.0377	0.940	0.0517
	SJEL	0.951	0.0118	0.955	0.0261	0.952	0.0403	0.955	0.0552
200	NA1	0.937	0.0079	0.952	0.0177	0.951	0.0283	0.950	0.0390
	NA2	0.932	0.0078	0.962	0.0175	0.960	0.0280	0.952	0.0382
	BT1	0.963	0.0087	0.963	0.0189	0.962	0.0296	0.961	0.0408
	BT2	0.934	0.0087	0.954	0.0189	0.952	0.0296	0.953	0.0408
	BT3	0.945	0.0075	0.952	0.0172	0.950	0.0272	0.950	0.0377
	BT4	0.952	0.0075	0.955	0.0172	0.957	0.0272	0.955	0.0377
	BCa1	0.953	0.0079	0.957	0.0182	0.955	0.0290	0.958	0.0402
	BCa2	0.940	0.0075	0.960	0.0173	0.945	0.0271	0.945	0.0376
	SJEL	0.962	0.0090	0.956	0.0185	0.950	0.0289	0.949	0.0396
500	NA1	0.937	0.0049	0.951	0.0111	0.945	0.0178	0.955	0.0246
	NA2	0.957	0.0049	0.960	0.0111	0.972	0.0178	0.965	0.0244
	BT1	0.953	0.0051	0.955	0.0114	0.951	0.0181	0.958	0.0251
	BT2	0.945	0.0051	0.950	0.0114	0.949	0.0181	0.948	0.0251
	BT3	0.950	0.0048	0.953	0.0110	0.953	0.0175	0.961	0.0241
	BT4	0.953	0.0048	0.957	0.0110	0.955	0.0175	0.951	0.0241
	BCa1	0.946	0.0049	0.954	0.0113	0.949	0.0180	0.952	0.0250
	BCa2	0.952	0.0048	0.952	0.0110	0.962	0.0175	0.960	0.0241
	SJEL	0.962	0.0057	0.957	0.0120	0.954	0.0187	0.958	0.0255

Table 3.5 Coverage probabilities and interval lengths at 95% confidence level for LC with Chi-square distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.5 to 0.9.

		t=50%		t=60%		t=70%		t=80%		t=90%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.940	0.0695	0.935	0.0821	0.932	0.0912	0.935	0.0943	0.907	0.0828
	NA2	0.925	0.0657	0.937	0.0763	0.937	0.0846	0.929	0.0830	0.782	0.0592
	BT1	0.957	0.0761	0.958	0.0902	0.963	0.1024	0.966	0.1095	0.986	0.1091
	BT2	0.936	0.0761	0.933	0.0902	0.934	0.1024	0.938	0.1095	0.929	0.1091
	BT3	0.945	0.0673	0.942	0.0798	0.939	0.0863	0.938	0.0888	0.927	0.0771
	BT4	0.949	0.0673	0.938	0.0798	0.937	0.0863	0.940	0.0888	0.940	0.0771
	BCa1	0.960	0.0737	0.957	0.0882	0.963	0.1011	0.959	0.1073	0.974	0.1026
	BCa2	0.950	0.0674	0.925	0.0794	0.932	0.0858	0.917	0.0872	0.892	0.0733
	SJEL	0.940	0.0701	0.946	0.0831	0.942	0.0921	0.945	0.0949	0.939	0.0824
200	NA1	0.955	0.0491	0.952	0.0580	0.952	0.0647	0.933	0.0671	0.943	0.0605
	NA2	0.838	0.0477	0.947	0.0560	0.926	0.0626	0.939	0.0637	0.937	0.0542
	BT1	0.960	0.0514	0.962	0.0609	0.960	0.0686	0.962	0.0727	0.970	0.0699
	BT2	0.957	0.0514	0.950	0.0609	0.947	0.0686	0.934	0.0727	0.938	0.0699
	BT3	0.959	0.0480	0.940	0.0571	0.943	0.0626	0.959	0.0648	0.932	0.0564
	BT4	0.940	0.0480	0.953	0.0571	0.959	0.0626	0.960	0.0648	0.932	0.0564
	BCa1	0.955	0.0508	0.956	0.0606	0.956	0.0686	0.963	0.0728	0.959	0.0713
	BCa2	0.922	0.0480	0.925	0.0569	0.912	0.0624	0.962	0.0641	0.915	0.0551
	SJEL	0.956	0.0500	0.951	0.0593	0.956	0.0661	0.947	0.0689	0.949	0.0621
500	NA1	0.942	0.0310	0.938	0.0367	0.947	0.0410	0.944	0.0426	0.944	0.0387
	NA2	0.945	0.0306	0.922	0.0361	0.952	0.0404	0.960	0.0419	0.962	0.0381
	BT1	0.945	0.0315	0.942	0.0373	0.947	0.0420	0.949	0.0441	0.955	0.0411
	BT2	0.939	0.0315	0.932	0.0373	0.938	0.0420	0.948	0.0441	0.937	0.0411
	BT3	0.945	0.0307	0.942	0.0363	0.947	0.0404	0.960	0.0417	0.954	0.0375
	BT4	0.947	0.0307	0.942	0.0363	0.952	0.0404	0.955	0.0417	0.952	0.0375
	BCa1	0.945	0.0314	0.946	0.0374	0.955	0.0421	0.952	0.0445	0.961	0.0421
	BCa2	0.942	0.0307	0.950	0.0363	0.940	0.0403	0.957	0.0416	0.947	0.0371
	SJEL	0.953	0.0314	0.952	0.0373	0.952	0.0421	0.948	0.0441	0.950	0.0410

(3.3) at 90% confidence levels, and Table (3.4) and Table (3.5) at 95% confidence levels. Secondly, we calculate coverage probabilities and average lengths for the Lorenz ordinate with Weibull distribution, for t ranges from 0.1 to 0.9, shown in Table (3.6) and Table (3.7) at 90% confidence level, and Table (3.8) and Table (3.9) at 95% confidence level.

Based on the simulation tables, all the confidence intervals are observed to perform better as sample size increases. As t increases, the average length of confidence intervals increases as well. BT1 and BT2 intervals have comparable performance, while BT3 and BT4 intervals have comparable performance. When t is chosen to be in the middle percent ($t = 40\%$ to $t = 80\%$), SJEL interval is observed to have good performance. However, when t increases or decreases to the two sides, SJEL interval becomes gradually underperformed. In most cases, the smoothed jackknife empirical likelihood-based (SJEL) confidence interval and the proposed smoothed bootstrap-based (BT3 and BT4) confidence intervals are observed to perform better than all other methods. Thus, SJEL interval outperforms the others for better coverage probabilities and shorter interval lengths, while BT3 and BT4 intervals perform the second to the best.

Based on the simulation results, we recommend the smoothed jackknife empirical likelihood-based confidence interval (SJEL) and the smoothed Boostrap-based confidence intervals (BT3, BT4) for the Lorenz ordinate since income data is skewed. Meanwhile, the smoothed jackknife empirical likelihood method is proved to be less computationally intensive than the plug-in empirical likelihood (EL) method proposed by Yang, Qin and Qin (2011).

3.4.2 A Real Example

Georgia Department of Audits and Accounts provides an open resource for Georgia public institute employee's annually-updated salary information. Our study will focus on professor in public school that cause more interests. To create a relative homogenous income group, we limit our analysis to the income of full-time assistant professors, associated professors, and full professors from Units of University System and Georgia Military Col-

Table 3.6 Coverage probabilities and interval lengths at 90% confidence level for LC with Weibull distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.1 to 0.4

		t=10%		t=20%		t=30%		t=40%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.876	0.0063	0.887	0.0171	0.891	0.0303	0.898	0.0448
	NA2	0.937	0.0064	0.867	0.0169	0.910	0.0298	0.915	0.0437
	BT1	0.941	0.0076	0.930	0.0190	0.927	0.0329	0.923	0.0488
	BT2	0.892	0.0076	0.897	0.0190	0.894	0.0329	0.896	0.0488
	BT3	0.897	0.0057	0.885	0.0157	0.897	0.0284	0.897	0.0421
	BT4	0.907	0.0057	0.897	0.0157	0.900	0.0284	0.897	0.0421
	BCa1	0.889	0.0060	0.909	0.0170	0.916	0.0307	0.912	0.0464
	BCa2	0.847	0.0058	0.890	0.0158	0.892	0.0285	0.895	0.0424
	SJEL	0.889	0.0071	0.896	0.0174	0.897	0.0304	0.894	0.0448
200	NA1	0.873	0.0043	0.899	0.0117	0.914	0.0211	0.915	0.0315
	NA2	0.827	0.0043	0.882	0.0117	0.882	0.0210	0.875	0.0314
	BT1	0.919	0.0047	0.916	0.0124	0.930	0.0220	0.930	0.0330
	BT2	0.875	0.0047	0.904	0.0124	0.913	0.0220	0.920	0.0330
	BT3	0.872	0.0040	0.857	0.0112	0.870	0.0205	0.865	0.0308
	BT4	0.882	0.0040	0.880	0.0112	0.879	0.0205	0.876	0.0308
	BCa1	0.900	0.0042	0.911	0.0118	0.919	0.0213	0.915	0.0323
	BCa2	0.845	0.0040	0.852	0.0113	0.865	0.0206	0.860	0.0309
	SJEL	0.911	0.0051	0.901	0.0122	0.918	0.0218	0.913	0.0321
500	NA1	0.916	0.0027	0.899	0.0074	0.888	0.0132	0.887	0.0198
	NA2	0.915	0.0026	0.902	0.0074	0.920	0.0133	0.907	0.0198
	BT1	0.925	0.0028	0.907	0.0076	0.891	0.0135	0.893	0.0202
	BT2	0.912	0.0028	0.897	0.0076	0.885	0.0135	0.881	0.0202
	BT3	0.900	0.0026	0.910	0.0072	0.911	0.0131	0.910	0.0196
	BT4	0.900	0.0026	0.905	0.0072	0.909	0.0131	0.902	0.0196
	BCa1	0.914	0.0026	0.909	0.0074	0.886	0.0133	0.893	0.0201
	BCa2	0.887	0.0026	0.900	0.0072	0.905	0.0131	0.900	0.0195
	SJEL	0.905	0.0028	0.900	0.0077	0.906	0.0136	0.906	0.0207

Table 3.7 Coverage probabilities and interval lengths at 90% confidence level for LC with Weibull distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.5 to 0.9

		t=50%		t=60%		t=70%		t=80%		t=90%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.912	0.0606	0.909	0.0751	0.899	0.0881	0.868	0.0957	0.870	0.0899
	NA2	0.857	0.0593	0.868	0.0727	0.885	0.0843	0.869	0.0886	0.885	0.0750
	BT1	0.928	0.0654	0.940	0.0816	0.927	0.0968	0.909	0.1085	0.935	0.1124
	BT2	0.909	0.0654	0.906	0.0816	0.892	0.0968	0.873	0.1085	0.879	0.1124
	BT3	0.887	0.0580	0.887	0.0719	0.880	0.0839	0.888	0.0903	0.865	0.0818
	BT4	0.889	0.0580	0.892	0.0719	0.889	0.0839	0.889	0.0903	0.880	0.0818
	BCa1	0.916	0.0629	0.912	0.0794	0.919	0.0947	0.904	0.1070	0.914	0.1083
	BCa2	0.892	0.0581	0.887	0.0719	0.840	0.0836	0.842	0.0894	0.817	0.0789
	SJEL	0.909	0.0605	0.906	0.0755	0.905	0.0889	0.890	0.0970	0.883	0.0932
200	NA1	0.898	0.0425	0.899	0.0531	0.896	0.0624	0.895	0.0682	0.881	0.0648
	NA2	0.879	0.0420	0.897	0.0524	0.897	0.0613	0.897	0.0662	0.865	0.0617
	BT1	0.902	0.0441	0.906	0.0554	0.912	0.0655	0.919	0.0725	0.923	0.0727
	BT2	0.898	0.0441	0.891	0.0554	0.889	0.0655	0.886	0.0725	0.878	0.0727
	BT3	0.896	0.0413	0.892	0.0515	0.887	0.0608	0.875	0.0659	0.877	0.0609
	BT4	0.898	0.0413	0.891	0.0515	0.888	0.0608	0.879	0.0659	0.879	0.0609
	BCa1	0.900	0.0434	0.905	0.0549	0.911	0.0653	0.916	0.0731	0.918	0.0734
	BCa2	0.890	0.0415	0.860	0.0517	0.860	0.0607	0.865	0.0656	0.840	0.0600
	SJEL	0.899	0.0431	0.900	0.0540	0.904	0.0636	0.906	0.0697	0.889	0.0679
500	NA1	0.898	0.0269	0.893	0.0337	0.895	0.0397	0.903	0.0435	0.895	0.0411
	NA2	0.865	0.0268	0.862	0.0336	0.867	0.0394	0.869	0.0431	0.932	0.0413
	BT1	0.901	0.0273	0.894	0.0343	0.901	0.0405	0.909	0.0445	0.908	0.0431
	BT2	0.897	0.0273	0.897	0.0343	0.898	0.0405	0.891	0.0445	0.892	0.0431
	BT3	0.895	0.0266	0.897	0.0333	0.890	0.0391	0.895	0.0427	0.905	0.0402
	BT4	0.898	0.0266	0.898	0.0333	0.895	0.0391	0.892	0.0427	0.907	0.0402
	BCa1	0.894	0.0272	0.897	0.0342	0.900	0.0405	0.911	0.0449	0.912	0.0440
	BCa2	0.892	0.0267	0.877	0.0334	0.882	0.0391	0.877	0.0427	0.902	0.0400
	SJEL	0.906	0.0271	0.902	0.0341	0.906	0.0405	0.909	0.0447	0.906	0.0432

Table 3.8 Coverage probabilities and interval lengths at 95% confidence level for LC with Weibull distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.1 to 0.4

		t=10%		t=20%		t=30%		t=40%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.932	0.0076	0.957	0.0203	0.961	0.0360	0.958	0.0538
	NA2	0.925	0.0077	0.937	0.0201	0.972	0.0356	0.960	0.0520
	BT1	0.982	0.0090	0.979	0.0226	0.974	0.0393	0.975	0.0585
	BT2	0.961	0.0090	0.968	0.0226	0.962	0.0393	0.949	0.0585
	BT3	0.937	0.0068	0.937	0.0187	0.947	0.0337	0.937	0.0503
	BT4	0.942	0.0068	0.945	0.0187	0.952	0.0337	0.940	0.0503
	BCa1	0.940	0.0070	0.968	0.0201	0.968	0.0365	0.965	0.0554
	BCa2	0.910	0.0068	0.927	0.0187	0.935	0.0337	0.940	0.0505
	SJEL	0.963	0.0083	0.947	0.0201	0.960	0.0359	0.950	0.0534
200	NA1	0.931	0.0051	0.946	0.0140	0.955	0.0252	0.961	0.0376
	NA2	0.932	0.0051	0.944	0.0140	0.955	0.0250	0.942	0.0374
	BT1	0.959	0.0056	0.956	0.0148	0.963	0.0262	0.966	0.0392
	BT2	0.935	0.0056	0.943	0.0148	0.959	0.0262	0.963	0.0392
	BT3	0.940	0.0048	0.945	0.0134	0.937	0.0243	0.940	0.0366
	BT4	0.939	0.0048	0.945	0.0134	0.939	0.0243	0.942	0.0366
	BCa1	0.950	0.0050	0.952	0.0140	0.956	0.0253	0.965	0.0383
	BCa2	0.915	0.0048	0.917	0.0134	0.937	0.0242	0.935	0.0367
	SJEL	0.959	0.0060	0.947	0.0147	0.953	0.0250	0.959	0.0381
500	NA1	0.962	0.0032	0.952	0.0088	0.948	0.0158	0.944	0.0236
	NA2	0.922	0.0032	0.962	0.0088	0.932	0.0158	0.960	0.0236
	BT1	0.968	0.0033	0.961	0.0091	0.951	0.0160	0.952	0.0241
	BT2	0.962	0.0033	0.954	0.0091	0.943	0.0160	0.947	0.0241
	BT3	0.955	0.0031	0.962	0.0086	0.950	0.0156	0.951	0.0234
	BT4	0.947	0.0031	0.951	0.0086	0.955	0.0156	0.953	0.0234
	BCa1	0.961	0.0031	0.955	0.0089	0.948	0.0158	0.952	0.0238
	BCa2	0.940	0.0031	0.950	0.0086	0.955	0.0156	0.957	0.0235
	SJEL	0.961	0.0033	0.955	0.0091	0.954	0.0160	0.959	0.0245

Table 3.9 Coverage probabilities and interval lengths at 95% confidence level for LC with Weibull distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.5 to 0.9

		t=50%		t=60%		t=70%		t=80%		t=90%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.930	0.0719	0.938	0.0893	0.935	0.1053	0.928	0.1140	0.913	0.1071
	NA2	0.931	0.0707	0.930	0.0867	0.937	0.1004	0.937	0.1056	0.860	0.0893
	BT1	0.953	0.0777	0.966	0.0972	0.955	0.1157	0.954	0.1292	0.700	0.1341
	BT2	0.935	0.0777	0.948	0.0972	0.935	0.1157	0.927	0.1292	0.929	0.1341
	BT3	0.947	0.0689	0.932	0.0854	0.940	0.0995	0.935	0.1074	0.925	0.0976
	BT4	0.937	0.0689	0.945	0.0854	0.936	0.0995	0.930	0.1074	0.937	0.0976
	BCa1	0.944	0.0746	0.954	0.0944	0.957	0.1141	0.953	0.1275	0.958	0.1285
	BCa2	0.935	0.0692	0.935	0.0852	0.910	0.0995	0.907	0.1062	0.867	0.0944
	SJEL	0.948	0.0720	0.949	0.0895	0.945	0.1055	0.943	0.1141	0.946	0.1068
200	NA1	0.949	0.0506	0.953	0.0633	0.949	0.0744	0.940	0.0812	0.937	0.0772
	NA2	0.945	0.0500	0.940	0.0625	0.945	0.0731	0.937	0.0788	0.942	0.0736
	BT1	0.955	0.0527	0.957	0.0661	0.962	0.0779	0.957	0.0865	0.971	0.0866
	BT2	0.945	0.0527	0.952	0.0661	0.945	0.0779	0.945	0.0865	0.930	0.0866
	BT3	0.957	0.0493	0.957	0.0618	0.944	0.0726	0.941	0.0784	0.935	0.0725
	BT4	0.940	0.0493	0.945	0.0618	0.942	0.0726	0.938	0.0784	0.938	0.0725
	BCa1	0.950	0.0516	0.949	0.0653	0.953	0.0776	0.963	0.0869	0.966	0.0880
	BCa2	0.927	0.0491	0.925	0.0615	0.922	0.0726	0.887	0.0781	0.882	0.0715
	SJEL	0.953	0.0512	0.956	0.0641	0.953	0.0756	0.948	0.0825	0.944	0.0793
500	NA1	0.944	0.0320	0.940	0.0402	0.941	0.0473	0.944	0.0518	0.941	0.0490
	NA2	0.942	0.0320	0.940	0.0400	0.940	0.0470	0.947	0.0514	0.950	0.0490
	BT1	0.948	0.0325	0.945	0.0408	0.945	0.0483	0.946	0.0531	0.950	0.0513
	BT2	0.942	0.0325	0.945	0.0408	0.946	0.0483	0.941	0.0531	0.937	0.0513
	BT3	0.948	0.0317	0.946	0.0396	0.945	0.0467	0.945	0.0507	0.952	0.0477
	BT4	0.947	0.0317	0.946	0.0396	0.950	0.0467	0.946	0.0507	0.952	0.0477
	BCa1	0.945	0.0324	0.938	0.0407	0.935	0.0482	0.941	0.0534	0.956	0.0523
	BCa2	0.935	0.0317	0.947	0.0397	0.935	0.0467	0.932	0.0507	0.955	0.0473
	SJEL	0.947	0.0323	0.944	0.0407	0.945	0.0481	0.946	0.0533	0.954	0.0514

lege in the 2012 fiscal year. With the initial dataset from the database, we observed that a few individuals with abnormally low wage, which may due to following reasons: (1) 2012 newly-hired professors, who did not work for the whole fiscal year; (2) part-time professors possible either take leave or transfer to another organization during the fiscal year. To filter out these subjects, we first excluded professors who does not have salary record in the 2011 fiscal year, then dropped out those with 2012 income far less than that of 2011 fiscal year and also those salary less than $20,000. Totally, there are 5,921 observations remain in the analysis.

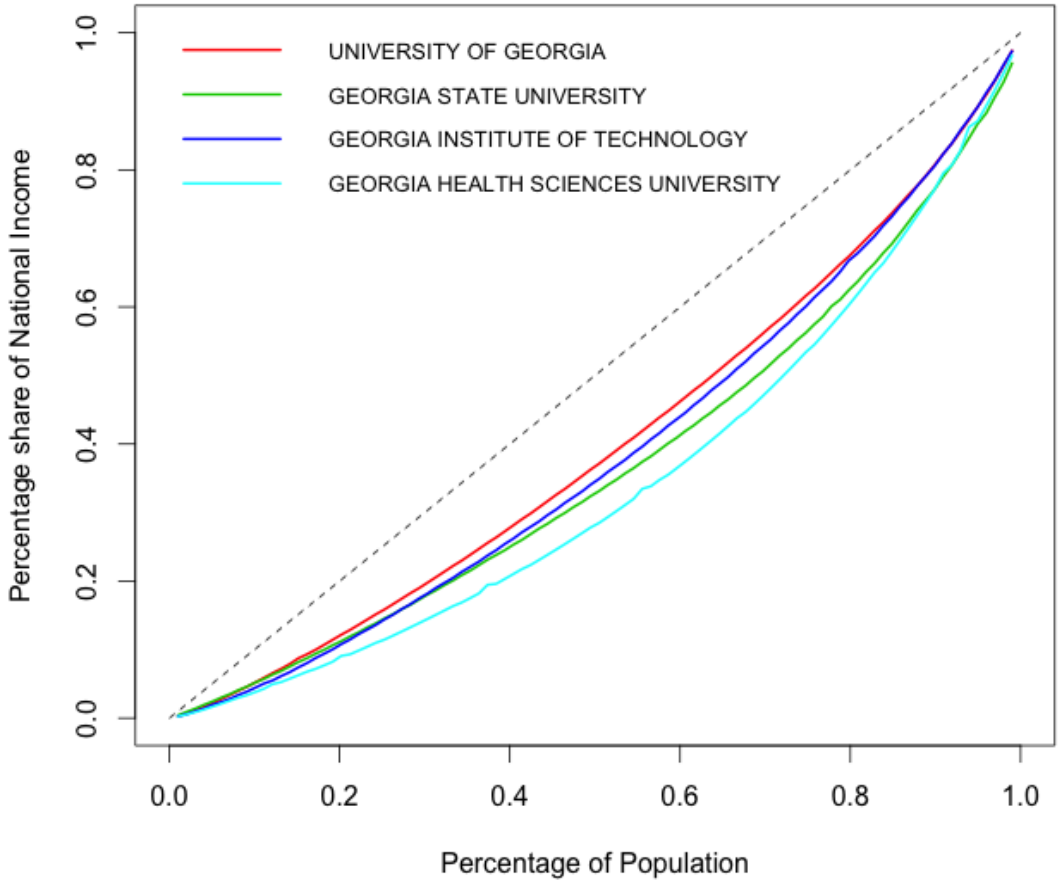

Figure 3.1 Lorenz Curve by School in 2012

We keep in mind that the distribution of the income data has right skewness. Since University of Georgia (UGA), Georgia State University (GSU), Georgia institute of Technology (GIT) and Georgia Health Sciences University (GHSU) have the largest number of recorded full-time professors, we will plot the Lorenz curves for each of the 4 schools in Figure 3.1. We observe that UGA with red color has a Lorenz curve that is closest to the line of equality. So we conclude that UGA has a more evenly-distributed income for professors, while GHSU has a more fluctuated income distribution.

To present the confidence interval, the level of significance is selected at 5%, and t is chosen from 0.5 to 0.8. We evaluate NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL intervals, and then summarized the results in Table (3.10).

It is observed that, at the 95% confidence level, for the least wealthy 80% individuals, the estimated Lorenz ordinate based on the SJEL interval is within (0.6158, 0.6288), which implies that the proportion of the total income earned by the least wealthy 80% individuals is between 61.58% and 62.88%. Meanwhile, the interval is (0.616, 0.628) based on BT3 and BT4 method, which indicates that the proportion of the total amount of income earned by the least wealthy 80% individuals is between 61.6% and 62.8%.

Table 3.10 Georgia Individual Income Example: Confidence interval and interval length at 95% confidence level for LC for professor's real income data are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with t from 0.5 to 0.8.

t	Method	Confidence Interval	Length
0.5	NA1	(0.3194567, 0.3274296)	0.0079729
	NA2	(0.3193395, 0.3272923)	0.0079528
	BT1	(0.3193553, 0.3275310)	0.0081757
	BT2	(0.3193755, 0.3275513)	0.0081757
	BT3	(0.3194435, 0.3271883)	0.0077448
	BT4	(0.3194340, 0.3271789)	0.0077448
	BCa1	(0.3192384, 0.3270876)	0.0078492
	BCa2	(0.3192576, 0.3271813)	0.0079237
	SJEL	(0.3186064, 0.3276064)	0.0090000
0.6	NA1	(0.4061575, 0.4156105)	0.0094530
	NA2	(0.4060734, 0.4155104)	0.0094370
	BT1	(0.4060339, 0.4157340)	0.0097001
	BT2	(0.4062423, 0.4159424)	0.0097001
	BT3	(0.4059776, 0.4156062)	0.0096286
	BT4	(0.4058651, 0.4154936)	0.0096286
	BCa1	(0.4056657, 0.4154564)	0.0097907
	BCa2	(0.4061621, 0.4157431)	0.0095810
	SJEL	(0.4049175, 0.4159175)	0.0110000
0.7	NA1	(0.5038581, 0.5146614)	0.0108033
	NA2	(0.5037835, 0.5145705)	0.0107870
	BT1	(0.5038366, 0.5146830)	0.0108464
	BT2	(0.5039583, 0.5148047)	0.0108464
	BT3	(0.5038569, 0.5144971)	0.0106402
	BT4	(0.5038748, 0.5145150)	0.0106402
	BCa1	(0.5033484, 0.5142673)	0.0109189
	BCa2	(0.5028203, 0.5136433)	0.0108230
	SJEL	(0.5026767, 0.5148765)	0.0121998
0.8	NA1	(0.6167516, 0.6283951)	0.0116435
	NA2	(0.6166880, 0.6283140)	0.0116260
	BT1	(0.6170826, 0.6280641)	0.0109815
	BT2	(0.6172053, 0.6281868)	0.0109815
	BT3	(0.6165209, 0.6284812)	0.0119603
	BT4	(0.6165525, 0.6285128)	0.0119603
	BCa1	(0.6171386, 0.6282821)	0.0111435
	BCa2	(0.6162519, 0.6283271)	0.0120752
	SJEL	(0.6158149, 0.6288149)	0.0130000

3.5 Discussion

There are other measurements to estimate the Lorenz curve. For example, we can also utilize the plug-in empirical likelihood (EL) method by calculating the scale parameters of Chi-square distribution, as proposed by Yang, Qin and Qin (2011). By computing the two estimates, we can get the scale constant to construct the EL-based confidence interval.

In our study, we propose a kernel estimator for Lorenz curve, and propose the cross-validation method to choose bandwidth h for the kernel estimator. Later, by comparing the empirical estimator and the kernel estimator, the kernel estimator is proved to be a good point estimator. Next, for the interval estimation evaluation, we compare 2 normal approximation based confidence intervals (NA1 and NA2), 4 bootstrap based confidence intervals (BT1, BT2, BT3, and BT4)), 2 bootstrap bias correction and acceleration intervals (BCa1 and BCa2) and a smoothed jackknife empirical likelihood (SJEL) confidence interval.

Our simulation studies present that the proposed smoothed jackknife empirical likelihood (SJEL) for Lorenz curve would perform the best, while the proposed bootstrap-based confidence intervals (BT3 and BT4) perform next to the best. Thus, we recommend the proposed SJEL-based confidence interval and the proposed smoothed bootstrap-based confidence intervals for the Lorenz curve.

3.6 Proof

Proof of Theorem 3.1, Theorem 3.2 and Theorem 3.3

Theorem 3.1. Under the conditions in Theorem 3.1, we have

$$\sqrt{n}\{\hat{T}_n(t) - \eta(t)\} \xrightarrow{d} N(0, \sigma^2(t)).$$

Proof:

$$\begin{aligned}
&\sqrt{n}[\hat{T}_n(t) - \eta(t)] \\
&= \sqrt{n}[\frac{1}{n\bar{x}} \sum_{i=1}^{n} X_i K(\frac{t - F_n(X_i)}{h}) - \frac{1}{\mu} \int_0^{\xi_t} x dF(x)] \\
&= \sqrt{n}[\frac{1}{n\bar{x}} \sum_{i=1}^{n} X_i K(\frac{t - F_n(X_i)}{h}) - \frac{1}{n\bar{x}} \sum_{i=1}^{n} X_i K(\frac{t - F(X_i)}{h})] \\
&+ \sqrt{n}[\frac{1}{n\bar{x}} \sum_{i=1}^{n} X_i K(\frac{t - F(X_i)}{h}) - \frac{1}{n\mu} \sum_{i=1}^{n} X_i K(\frac{t - F(X_i)}{h})] \\
&+ \sqrt{n}[\frac{1}{n\mu} \sum_{i=1}^{n} X_i K(\frac{t - F(X_i)}{h}) - \frac{1}{\mu} \int_0^{\xi_t} x dF(x)] \\
&= I_1 + I_2 + I_3. \qquad (3.12)
\end{aligned}$$

Then I_1 of (3.12) can be written as

$$\begin{aligned}
I_1 \quad &= \frac{\sqrt{n}}{n\bar{x}} \sum_{i=1}^{n} X_i [K(\frac{t - F_n(X_i)}{h}) - K(\frac{t - F(X_i)}{h})] \\
&= \frac{1}{\bar{x}} \int_{-\infty}^{\infty} x [K(\frac{t - F_n(x)}{h}) - K(\frac{t - F(x)}{h})] d[\sqrt{n}(F_n(x) - F(x))] \\
&+ \frac{\sqrt{n}}{\bar{x}} \int_{-\infty}^{\infty} x [K(\frac{t - F_n(x)}{h}) - K(\frac{t - F(x)}{h})] dF(x) \\
&= I_{11} + I_{12}. \qquad (3.13)
\end{aligned}$$

By Taylor Series, under condition of Theorem 3.1, I_{12} of (3.13) can be written as

$$\begin{aligned} I_{12} &= \frac{\sqrt{n}}{\bar{x}} \int_{-\infty}^{\infty} x[K(\frac{t-F_n(x)}{h}) - K(\frac{t-F(x)}{h})]dF(x) \\ &= \frac{\sqrt{n}}{\mu} \int_{-\infty}^{\infty} x[\omega(\frac{t-F(x)}{h})\frac{F(x)-F_n(x)}{h} + \frac{1}{2}\omega'(\frac{t-F(x)}{h})(\frac{F(x)-F_n(x)}{h})^2]dF(x) + o_p(1) \\ &= -\frac{\sqrt{n}}{\mu} \int_{-\infty}^{\infty} x[\omega(\frac{t-F(x)}{h})\frac{F_n(x)-F(x)}{h}]dF(x) + O_p(\frac{1}{\sqrt{n}h}) \\ &= -\frac{1}{\mu} \int_{-\infty}^{\infty} x\omega(\frac{t-F(x)}{h})\frac{\sqrt{n}(F_n(x)-F(x))}{h}dF(x) + o_p(1). \end{aligned} \tag{3.14}$$

Let Y=F(X), thus $X = F^{-1}(Y)$, and let

$$U_n(y) = \sqrt{n}[\frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y) - y]$$

.

Clearly Y_i follows uniform[0,1] distribution. Since $h \to 0$, so (3.14) will be equal to

$$\begin{aligned} I_{12} &= -\frac{1}{\mu} \int_{-1}^{1} F^{-1}(y)\omega(\frac{t-y}{h})\frac{1}{h}U_n(y)dy + o_p(1) \\ &= \frac{1}{\mu} \int_{\frac{t-1}{h}}^{\frac{t+1}{h}} F^{-1}(t-hz)\omega(z)U_n(t-hz)dz + o_p(1) \\ &= \frac{1}{\mu} \int_{-a}^{a} F^{-1}(t)\omega(z)U_n(t)dz + o_p(1) \\ &= \frac{1}{\mu} F^{-1}(t)U_n(t) \int_{a}^{b} \omega(z)dz + o_p(1) \\ &= \frac{1}{\mu}\xi_t U_n(t) + o_p(1). \end{aligned} \tag{3.15}$$

where $[a, b]$ is the supporting set of $\omega(x)$. Since $\sqrt{n}[F_n(x) - F(x)] \to B(x)$, which is Gauss Process, so $I_{11} = o_p(1)$.

So we get

$$I_1 = \frac{1}{\mu}\xi_t U_n(t) + o_p(1). \tag{3.16}$$

Secondly, for I_2 of (3.12), we are going to prove

$$\int_{-\infty}^{\infty} xK(\frac{t-F(x)}{h})]dF(x) \longrightarrow \int_0^{\xi_t} xdF(x), as\ h \to 0. \tag{3.17}$$

Since $h \to 0$, it is straight to show

$$\begin{aligned}
&\lim_{h\to 0}\int_{-\infty}^{\infty} xK(\frac{t-F(x)}{h})f(x)dx \\
&= \lim_{h\to 0}\int_{-\infty}^{\infty} x\int_{-\infty}^{\frac{t-F(x)}{h}} \omega(y)dyf(x)dx \\
&= \int_{-\infty}^{\infty} x[\lim_{h\to 0}\int_{-\infty}^{\frac{t-F(x)}{h}} \omega(y)dy]f(x)dx \\
&= \int_{-\infty}^{\infty} x\{0 * I[t < F(x)] + \frac{1}{2} * I[t = F(x)] + 1 * I[t > F(x)]\}f(x)dx \\
&= \int_{-\infty}^{\infty} xI[t > F(x)]f(x)dx \\
&= \int_{-\infty}^{\infty} xI[t > F(x)]dF(x) \\
&= \int_0^{\xi_t} xdF(x) \\
&= \mu\eta(t),
\end{aligned} \tag{3.18}$$

and

$$\begin{aligned}
&\lim_{h\to 0}\int_{-\infty}^{\infty} x^2K^2(\frac{t-F(x)}{h})f(x)dx \\
&=\lim_{h\to 0}\int_{-\infty}^{\infty} x^2(\int_{-\infty}^{\frac{t-F(x)}{h}}\omega(y)dy)^2f(x)dx \\
&=\int_{-\infty}^{\infty} x^2[\lim_{h\to 0}\int_{-\infty}^{\frac{t-F(x)}{h}}\omega(y)dy]^2f(x)dx \\
&=\int_{-\infty}^{\infty} x^2\{0*I[t<F(x)]+\frac{1}{2}*I[t=F(x)]+1*I[t>F(x)]\}^2f(x)dx \\
&=\int_{-\infty}^{\infty} x^2\{I[t>F(x)]\}^2f(x)dx \\
&=\int_{-\infty}^{\infty} x^2I[t>F(x)]dF(x) \\
&=\int_{0}^{\xi_t} x^2dF(x). \qquad (3.19)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\lim_{h\to 0}\int_{-\infty}^{\infty} x^2K(\frac{t-F(x)}{h})f(x)dx \\
&=\lim_{h\to 0}\int_{-\infty}^{\infty} x^2(\int_{-\infty}^{\frac{t-F(x)}{h}}\omega(y)dy)f(x)dx \\
&=\int_{-\infty}^{\infty} x^2[\lim_{h\to 0}\int_{-\infty}^{\frac{t-F(x)}{h}}\omega(y)dy]f(x)dx \\
&=\int_{-\infty}^{\infty} x^2\{0*I[t<F(x)]+\frac{1}{2}*I[t=F(x)]+1*I[t>F(x)]\}f(x)dx \\
&=\int_{-\infty}^{\infty} x^2\{I[t>F(x)]\}f(x)dx \\
&=\int_{-\infty}^{\infty} x^2I[t>F(x)]dF(x) \\
&=\int_{0}^{\xi_t} x^2dF(x). \qquad (3.20)
\end{aligned}$$

Thus,

$$
\begin{aligned}
I_2 &= \sqrt{n}[\frac{1}{n\bar{x}}\sum_{i=1}^{n} X_i K(\frac{t-F(X_i)}{h}) - \frac{1}{n\mu}\sum_{i=1}^{n} X_i K(\frac{t-F(X_i)}{h})] \\
&= \frac{\sqrt{n}(\mu-\bar{x})}{\bar{x}\mu}\frac{1}{n}\sum_{i=1}^{n} X_i K(\frac{t-F(X_i)}{h}) + o_p(1) \\
&= \frac{\sqrt{n}(\mu-\bar{x})}{\bar{x}\mu}[\frac{1}{n}\sum_{i=1}^{n} X_i K(\frac{t-F(X_i)}{h}) - \int_0^{\xi_t} x dF(x)] + \frac{\sqrt{n}(\mu-\bar{x})}{\bar{x}\mu}\int_0^{\xi_t} x dF(x) + o_p(1) \\
&= \frac{\sqrt{n}(\mu-\bar{x})}{\bar{x}\mu}[\int_{-\infty}^{\infty} x K(\frac{t-F(x)}{h}) dF(x) - \int_0^{\xi_t} x dF(x)] \\
&+ \frac{\sqrt{n}(\mu-\bar{x})}{\bar{x}\mu}\int_0^{\xi_t} x dF(x) + o_p(1) \\
&= \frac{\sqrt{n}(\mu-\bar{x})}{\bar{x}\mu}\int_0^{\xi_t} x dF(x) + o_p(1) \\
&= -\frac{1}{\mu^2}\int_0^{\xi_t} x dF(x)\frac{\sqrt{n}}{n}\sum_{i=1}^{n}(X_i-\mu) + o_p(1). \qquad (3.21)
\end{aligned}
$$

Next, let's consider $I_2 + I_3$ of (3.12).

$$
\begin{aligned}
& I_2 + I_3 \\
&= -\frac{1}{\mu^2}\int_0^{\xi_t} x dF(x)\frac{\sqrt{n}}{n}\sum_{i=1}^{n}(X_i-\mu) + \frac{\sqrt{n}}{n\mu}\sum_{i=1}^{n} X_i K(\frac{t-F(X_i)}{h}) - \frac{\sqrt{n}}{\mu}\int_0^{\xi_t} x dF(x) + o_p(1) \\
&= -\frac{\eta(t)}{\mu}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu) + \frac{1}{\sqrt{n}\mu}\sum_{i=1}^{n}[X_i K(\frac{t-F(X_i)}{h}) - EX_i K(\frac{t-F(X_i)}{h})] + o_p(1) \\
&= \frac{1}{\mu\sqrt{n}}\sum_{i=1}^{n}[-\eta(t)(X_i-\mu) + X_i K(\frac{t-F(X_i)}{h}) - EX_i K(\frac{t-F(X_i)}{h})] + o_p(1) \\
&= \frac{1}{\mu\sqrt{n}}\sum_{i=1}^{n}\{[X_i K(\frac{t-F(X_i)}{h}) - \eta(t)X_i] - [EXK(\frac{t-F(X)}{h}) - \eta(t)\mu]\} + o_p(1) \\
&= \frac{1}{\mu\sqrt{n}}\sum_{i=1}^{n}\{\omega_i - E\omega_i\} + o_p(1) \\
&\longrightarrow N(0, \frac{1}{\mu^2}Var(\omega)), \qquad (3.22)
\end{aligned}
$$

where $\omega_i = X_i K(\frac{t-F(X_i)}{h}) - \eta(t)X_i$ and $\omega = XK(\frac{t-F(X)}{h}) - \eta(t)X$.

Since $U_n(t)$ is an ancillary statistics. Based on Basu's theorem, $U_n(t)$ is independent of

$\frac{1}{n}\sum_{i=1}^{n}(\omega_i - E\omega)$.

Based on (3.16), (3.21) and (3.22), we have

$$
\begin{aligned}
& Var[\frac{1}{\mu}\xi_t U_n(t))] + \frac{1}{\mu^2}Var(\omega) \\
&= \frac{1}{\mu^2}\xi_t^2 t(1-t) + \frac{1}{\mu^2}Var[XK(\frac{t-F(X)}{h}) - \eta(t)X] \\
&= \frac{1}{\mu^2}\xi_t^2 t(1-t) + \frac{1}{\mu^2}\{E[X^2(K(\frac{t-F(X)}{h}) - \eta(t))^2] - [E(XK(\frac{t-F(X)}{h}) - \eta(t)X)]^2\} \\
&= \frac{1}{\mu^2}\xi_t^2 t(1-t) + \frac{1}{\mu^2}E[X^2(K(\frac{t-F(X)}{h}) - \eta(t))^2] - \frac{1}{\mu^2}[EXK(\frac{t-F(X)}{h}) - \eta(t)EX]^2 \\
&= \frac{1}{\mu^2}\xi_t^2 t(1-t) + \frac{1}{\mu^2}E[X^2(K(\frac{t-F(X)}{h}) - \eta(t))^2] - \frac{1}{\mu^2}[EXK(\frac{t-F(X)}{h}) - \eta(t)\mu]^2 \\
&= \frac{1}{\mu^2}\xi_t^2 t(1-t) + \frac{1}{\mu^2}E[X^2(K(\frac{t-F(X)}{h}) - \eta(t))^2] + o_p(1) \\
&= \frac{1}{\mu^2}\{\xi_t^2 t(1-t) + EX^2K^2(\frac{t-F(X)}{h}) + \eta^2(t)EX^2 - 2\eta(t)EX^2K(\frac{t-F(X)}{h})\} + o_p(1) \\
&\longrightarrow \frac{1}{\mu^2}\{\xi_t^2 t(1-t) + (1-2\eta(t))\int_0^{\xi_t} x^2 dF(x) + \eta^2(t)\int_{-\infty}^{\infty} x^2 dF(x)\} \\
&\equiv \sigma^2(t). \qquad (3.23)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
I_1 + I_2 + I_3 \quad &= \frac{1}{\mu}\xi_t U_n(t) + \frac{1}{\mu\sqrt{n}}\sum_{i=1}^{n}\{\omega_i - E\omega\} + o_p(1) \\
&= \frac{1}{\mu}\xi_t U_n(t) + \frac{\sqrt{n}}{\mu}\frac{1}{n}\sum_{i=1}^{n}\{\omega_i - E\omega\} + o_p(1) \\
&\xrightarrow{d} N(0, \sigma^2(t)). \qquad (3.24)
\end{aligned}
$$

We need Lemma 1 and Lemma 2 to prove Theorem 3.2

Lemma 1. Under the conditions in Theorem 3.1, we have

$$\sqrt{n}\{\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k(t)-\eta(t)\} \xrightarrow{d} N(0,\sigma^2(t)). \tag{3.25}$$

Proof: Note $\frac{1}{n}\sum_{k=1}^{n}V_k(t)$ can be decomposed into

$$\frac{1}{n}\sum_{k=1}^{n}V_k(t)=\frac{n-1}{n}\sum_{k=1}^{n}(\hat{T}_n-\hat{T}_{n-1,k})+\hat{T}_n, \tag{3.26}$$

We have

$$\begin{aligned}
&\hat{T}_n-\hat{T}_{n-1,k}\\
&=\frac{1}{n\bar{X}}\sum_{i=1}^{n}X_iK(\frac{t-F_n(X_i)}{h})-\frac{1}{(n-1)\bar{X}_{n-1,k}}\sum_{j\neq k}X_jK(\frac{t-F_{n,k}(X_j)}{h})\\
&=\frac{1}{n\bar{X}}\sum_{i=1}^{n}X_iK(\frac{t-F_n(X_i)}{h})-\frac{1}{n\bar{X}}\sum_{j\neq k}^{n}X_jK(\frac{t-F_{n,k}(X_j)}{h})\\
&+\frac{1}{n\bar{X}}\sum_{j\neq k}^{n}X_jK(\frac{t-F_{n,k}(X_j)}{h})-\frac{1}{(n-1)\bar{X}_{n-1,k}}\sum_{j\neq k}X_jK(\frac{t-F_{n,k}(X_j)}{h}),
\end{aligned} \tag{3.27}$$

where $\bar{X}_{n-1,k}=\frac{1}{n-1}\sum_{j\neq k}^{n}X_j$.

So

$$
\begin{aligned}
&\sum_{k=1}^{n}(\hat{T}_n - \hat{T}_{n-1,k}) \\
&= \frac{1}{n\bar{X}}\{\sum_{k=1}^{n}\sum_{i=1}^{n} X_i K(\frac{t-F_n(X_i)}{h}) - \sum_{k=1}^{n}\sum_{j\neq k} X_j[K(\frac{t-F_{n,k}(X_j)}{h})]\} \\
&+\sum_{k=1}^{n}[\{\frac{1}{n\bar{X}}\sum_{j\neq k} X_j K(\frac{t-F_{n,k}(X_j)}{h}) - \frac{1}{(n-1)\bar{X}_{n-1,k}}\sum_{j\neq k} X_j K(\frac{t-F_{n,k}(X_j)}{h})\} \\
&= \frac{1}{n\bar{X}}\sum_{k=1}^{n}\sum_{i=1}^{n} X_i[K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})] + \frac{1}{n\bar{X}}\sum_{k=1}^{n} X_k K(\frac{t-F_{n,k}(X_k)}{h}) \\
&+\frac{1}{n\bar{X}}\sum_{k=1}^{n}\sum_{j\neq k} X_j K(\frac{t-F_{n,k}(X_j)}{h}) - \frac{1}{(n-1)\bar{X}_{n-1,k}}\sum_{k=1}^{n}\sum_{j\neq k} X_j K(\frac{t-F_{n,k}(X_j)}{h}) \\
&= \frac{1}{n\bar{X}}\sum_{k=1}^{n}\sum_{i=1}^{n} X_i[K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})] \\
&+\{\frac{1}{n\bar{X}}\sum_{k=1}^{n}\sum_{j=1}^{n} X_j K(\frac{t-F_{n,k}(X_j)}{h}) - \frac{1}{(n-1)\bar{X}_{n-1,k}}\sum_{k=1}^{n}\sum_{j=1}^{n} X_j K(\frac{t-F_{n,k}(X_j)}{h}) \\
&+\frac{1}{(n-1)\bar{X}_{n-1,k}}\sum_{k=1}^{n} X_k K(\frac{t-F_{n,k}(X_k)}{h})\} \\
&= I_1 + I_2. \qquad (3.28)
\end{aligned}
$$

By Talyor series, I_1 of (3.28) can be written as

$$
\begin{aligned}
&\frac{1}{n\bar{X}}\sum_{k=1}^{n}\sum_{i=1}^{n} X_i[K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})] \\
&= \frac{1}{n\bar{X}}\sum_{k=1}^{n}\sum_{i=1}^{n} X_i\{-\omega(\frac{t-F_n(X_i)}{h})\frac{F_{n,k}(X_i)-F_n(X_i)}{h} \\
&-\frac{1}{2}\omega'(\frac{t-\xi_{n,k,i}}{h})(\frac{F_{n,k}(X_i)-F_n(X_i)}{h})^2\} \\
&= \frac{1}{n\bar{X}}\sum_{i=1}^{n} X_i\{-\omega(\frac{t-F_n(X_i)}{h})\sum_{k=1}^{n}\frac{F_{n,k}(X_i)-F_n(X_i)}{h} \\
&-\frac{1}{2}\sum_{k=1}^{n}\omega'(\frac{t-\xi_{n,k,i}}{h})(\frac{F_{n,k}(X_i)-F_n(X_i)}{h})^2\} \\
&= -\frac{1}{2}\frac{1}{n\bar{X}}\sum_{i=1}^{n}\sum_{k=1}^{n} X_i\omega'(\frac{t-\xi_{n,k,i}}{h})(\frac{F_{n,k}(X_i)-F_n(X_i)}{h})^2. \qquad (3.29)
\end{aligned}
$$

where $\xi_{n,k,i}$ is a random variable between $F_n(X_i)$ and $F_{n,k}(X_i)$. Since

$$\begin{aligned} F_n(X) - F_{n,k}(X) &= \frac{1}{n}\sum_{i=1}^{n} I(X_i \le X) - \frac{1}{n-1}\sum_{j\neq k} I(X_i \le X) \\ &= \frac{1}{n}\sum_{i=1}^{n} I(X_i \le X) - \frac{1}{n-1}\sum_{i=1}^{n} I(X_i \le X) + \frac{1}{n-1} I(X_k \le X) \\ &= -\frac{1}{n-1}\frac{1}{n}\sum_{i=1}^{n} I(X_i \le X) + \frac{1}{n-1} I(X_k \le X) \\ &= \frac{1}{n-1}\{I(X_k \le X) - F_n(X)\} \\ &= O_p(\frac{1}{n-1}) \\ &= O_p(\frac{1}{n}). \end{aligned} \tag{3.30}$$

Under conditions of Theorem 3.1, ω' is bounded. We further assume that the product of x and ω' is bounded at its supporting set. Thus, (3.29) can be decomposed into

$$\begin{aligned} & I_1 \\ &= -\frac{1}{2\bar{X}}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} |X_i\omega'(\frac{t-\xi_{n,k,i}}{h})|(\frac{F_{n,k}(X_i) - F_n(X_i)}{h})^2 \\ &= -\frac{1}{2\bar{X}}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} |X_i\omega'(\frac{t-\xi_{n,k,i}}{h})|O_p(\frac{1}{(n-1)^2h^2}) \\ &= -\frac{1}{2\bar{X}}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} |X_i\omega'(\frac{t-\xi_{n,k,i}}{h})|O_p(\frac{1}{(n-1)^2h^2}) \\ &= -\frac{1}{2}O_p(n)O_p(\frac{1}{(n-1)^2h^2}) \\ &= O_p(\frac{1}{nh^2}). \end{aligned} \tag{3.31}$$

Then, based on (3.31), I_2 of (3.28) can be written to

$$
\begin{aligned}
& I_2 \\
&= \frac{1}{n\bar{X}} \sum_{k=1}^{n} \sum_{j=1}^{n} X_j [K(\frac{t - F_{n,k}(X_j)}{h}) - K(\frac{t - F_n(X_j)}{h})] + \frac{1}{n\bar{X}} \sum_{k=1}^{n} \sum_{j=1}^{n} X_j K(\frac{t - F_n(X_j)}{h}) \\
&- \frac{1}{(n-1)\bar{X}_{n-1,k}} \sum_{k=1}^{n} \sum_{j=1}^{n} X_j [K(\frac{t - F_{n,k}(X_j)}{h}) - K(\frac{t - F_n(X_j)}{h})] \\
&+ \frac{1}{(n-1)\bar{X}_{n-1,k}} \sum_{k=1}^{n} X_k [K(\frac{t - F_{n,k}(X_k)}{h}) - K(\frac{t - F_n(X_k)}{h})] \\
&- \frac{1}{(n-1)\bar{X}_{n-1,k}} \sum_{k=1}^{n} \sum_{j=1}^{n} X_j K(\frac{t - F_n(X_j)}{h}) + \frac{1}{(n-1)\bar{X}_{n-1,k}} \sum_{k=1}^{n} X_k K(\frac{t - F_n(X_k)}{h}) \\
&= O_p(\frac{1}{nh^2}) + \frac{1}{n\bar{X}} \sum_{k=1}^{n} \sum_{j=1}^{n} X_j K(\frac{t - F_n(X_j)}{h}) \\
&+ O_p(\frac{1}{nh^2}) + O_p(\frac{1}{n}) \\
&- \frac{1}{(n-1)\bar{X}_{n-1,k}} \sum_{k=1}^{n} \sum_{j=1}^{n} X_j K(\frac{t - F_n(X_j)}{h}) + \frac{1}{(n-1)\bar{X}_{n-1,k}} \sum_{k=1}^{n} X_k K(\frac{t - F_n(X_k)}{h}) \\
&= \sum_{k=1}^{n} \{(\frac{1}{n\bar{X}} - \frac{1}{(n-1)\bar{X}_{n-1,k}}) \sum_{j=1}^{n} X_j K(\frac{t - F_n(X_j)}{h}) + \frac{X_k}{(n-1)\bar{X}_{n-1,k}} K(\frac{t - F_n(X_k)}{h})\} \\
&+ O_p(\frac{1}{nh^2}) + O_p(\frac{1}{n}) \\
&= \sum_{k=1}^{n} \{\frac{-X_k}{\sum_{i=1}^{n} X_i \sum_{l \neq k}^{n} X_l} \sum_{j=1}^{n} X_j K(\frac{t - F_n(X_j)}{h}) + \frac{X_k}{\sum_{l \neq k} X_l} K(\frac{t - F_n(X_k)}{h})\} + O_p(\frac{1}{n}) \\
&= \sum_{k=1}^{n} \{\frac{X_k}{\sum_{l \neq k}^{n} X_l} [K(\frac{t - F_n(X_k)}{h}) - \frac{\sum_{j=1}^{n} X_j K(\frac{t - F_n(X_j)}{h})}{\sum_{i=1}^{n} X_i}]\} + O_p(\frac{1}{n}) \\
&= \sum_{k=1}^{n} \{\frac{X_k}{\sum_{l \neq k}^{n} X_l} [\frac{\sum_{i=1}^{n} X_i K(\frac{t - F_n(X_i)}{h})}{\sum_{i=1}^{n} X_i} - \frac{\sum_{j=1}^{n} X_j K(\frac{t - F_n(X_j)}{h})}{\sum_{i=1}^{n} X_i}]\} + O_p(\frac{1}{n}) \\
&= \sum_{k=1}^{n} \{\frac{X_k}{\sum_{i=1}^{n} X_i \sum_{l \neq k}^{n} X_l} [\sum_{i=1}^{n} X_i K(\frac{t - F_n(X_i)}{h}) - \sum_{j=1}^{n} X_j K(\frac{t - F_n(X_j)}{h})]\} + O_p(\frac{1}{n}) \\
&= O_p(\frac{1}{n}). \qquad (3.32)
\end{aligned}
$$

From (3.31) and (3.32), we get (3.26) as follows:

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n} V_k(t) \\
&= \frac{n-1}{n}\sum_{k=1}^{n}(\hat{T}_n - \hat{T}_{n-1,k}) + \hat{T}_n(t) \\
&= \frac{n-1}{n} O_p(\frac{1}{nh^2}) + \hat{T}_n(t) \\
&= \hat{T}_n(t) + O_p(\frac{1}{nh^2}),
\end{aligned} \tag{3.33}$$

and hence,

$$\begin{aligned}
&\sqrt{n}\{\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k(t) - \eta(t)\} \\
&= \sqrt{n}\{[\hat{T}_n(t) - \eta(t)] + O_p(\frac{1}{\sqrt{n}h^2})\} \\
&\xrightarrow{d} N(0, \sigma^2(t)).
\end{aligned} \tag{3.34}$$

Thus, Lemma 1 is proved.

Lemma 2. Under the conditions in Theorem 3.1, we have

$$\frac{1}{n}\sum_{k=1}^{n}\{\hat{V}_k(t)-\eta(t)\}^2 \xrightarrow{p} \sigma^2(t). \tag{3.35}$$

Proof:

Lemma 2 can also be proved using the similar method to that of the proof for Lemma 2 in Part 4 for generalized Lorenz curve. Meanwhile, the sample Lorenz curve parameters are examples of smooth L-statistics or functions of smooth L-statistics (Shao (1994)). Shao (1994) also proved that the smooth L-statistics are asymptotically normal under weak conditions and the asymptotic variances can be consistently estimated by jackknifing. Moreover, the textbook by Jun, S. & Tu, D. S. (1995) shows that the jackknife variance estimate of smooth L-statistics is consistent according to Theorem 3.3 . So without proof, Lemma 2 is established.

Proof of Theorem 3.2. It follows immediately from Lemma 1 and Lemma 2.

Proof of Theorem 3.2. Similar to Gong, Peng and Li(2010), define $g(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{V}_i(t)-\eta}{1+\lambda(\hat{V}_i(t)-\eta)}$ It is easy to check that

$$\begin{aligned}
0 = |g(\lambda)| &= \frac{1}{n}|\sum_{i=1}^{n}(\hat{V}_i(t)-\eta) - \lambda\sum_{i=1}^{n}\frac{(\hat{V}_i(t)-\eta)^2}{1+\lambda(\hat{V}_i(t)-\eta)}| \\
&\geq |\frac{\lambda}{n}\sum_{i=1}^{n}\frac{(\hat{V}_i(t)-\eta)^2}{1+\lambda(\hat{V}_i(t)-\eta)}| - |\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\eta| \\
&\geq \frac{|\lambda|S_n}{1+|\lambda|Z_n} - |\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\eta)|, \qquad (3.36)
\end{aligned}$$

where $S_n = \frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\eta)^2$ and $Z_n = max_{1\leq i\leq n}|\hat{V}_i(t)-\eta|$.

From Lemma 1 and Lemma 2, we have $|\lambda| = O_p\{n^{-\frac{1}{2}}\}$.

Put $\gamma_i = \lambda(\hat{V}_i(t)-\eta)$, then we have $max_{1\leq i\leq n}|\gamma_i| = O_p(1)$.

$$\begin{aligned}
0 = g(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\eta)(1-\gamma_i+\frac{\gamma_i^2}{1+\gamma_i}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\eta) - S_n\lambda + \frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{V}_i(t)-\eta)\gamma_i^2}{1+\gamma_i} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\eta) - S_n\lambda + O_p(\frac{1}{n}), \qquad (3.37)
\end{aligned}$$

which implies that $\lambda = S_n^{-1}\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\eta) + \beta_n$, where $\beta_n = O_p(\frac{1}{n})$.

So

$$
\begin{aligned}
& l_n(\eta(t)) \\
&= -2\ log\ L_n(\eta(t)) \\
&= 2\sum_{i=1}^{n} log\ \{1+\lambda(\hat{V}_i(t)-\eta)\} \\
&= 2\sum_{i=1}^{n}\gamma_i - \sum_{i=1}^{n}\gamma_i^2 + 2\sum_{i=1}^{n}\eta_i \\
&= 2n\lambda\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\eta) - nS_n\lambda^2 + 2\sum_{i=1}^{n}\eta_i \\
&= \frac{n\{\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\eta)\}^2}{S_n} - nS_n\beta_n^2 + 2\sum_{i=1}^{n}\eta_i \\
&= \frac{n\{\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\eta)\}^2}{S_n} + O_p(1) \\
&\xrightarrow{d} \chi^2(1). \qquad (3.38)
\end{aligned}
$$

Theorem 3.2 holds.

PART 4

GENERALIZED LORENZ CURVE

This part is organized as follows. In Section 4.1, we propose a kernel estimator for the generalized Lorenz curve, and then apply cross-validation method to choose bandwidth h. Then, the empirical estimator and the kernel estimator for generalized Lorenz curve (GLC) are evaluated in terms of Mean Square Error (MSE) and Asymptotic Relative Efficiency (ARE). In Section 4.2, the jackknife pseudo-values for generalized Lorenz curve are defined, and then the jackknife empirical likelihood properties are derived. In section 4.3, multiple confidence intervals are developed and presented. In section 4.4, extensive simulation are conducted to compare the two point estimators, and the proposed confidence intervals. Later, the proposed methods are illustrated through a real example. Proof will be provided at the end of this part.

4.1 Estimation of a Generalized Lorenz Curve

4.1.1 Empirical Estimator for Generalized Lorenz Curve

Shorrocks (1983) and Kakwani (1984) defined the generalized Lorenz curve as follows

$$\theta(t) = \int_0^{\xi_t} x dF(x), \quad 0 \leq t \leq 1, \tag{4.1}$$

Let $X_1, X_2, ..., X_n$ be a simple random sample drawn from the population X with c.d.f. $F(x)$.

For a fixed $t \in (0, 1)$, the generalized Lorenz curve θ(t) satisfies

$$E[XI(X \leq \xi_t)] - \theta(t) = 0.$$

An empirical estimate for $\theta(t)$ can be found from the following estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} X_i I(F_n(X_i) \leq t) - \theta(t) = 0,$$

where

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

Therefore, the empirical estimator for generalized Lorenz curve $\theta(t)$ is

$$\hat{\theta}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i I(X_i \leq F_n^{-1}(t)) = \frac{1}{n}\sum_{i=1}^{n} X_i I(t \geq F_n(X_i)).$$

4.1.2 A Kernel Estimator for a Generalized Lorenz Curve

As we mentioned in Part 2 and Part 3, we should use the kernel estimator for the generalized Lorenz curve $\theta(t)$, because $\theta(t)$ is a smoothing function in many applications. To find a smoothing estimator for $\theta(t)$, we utilize the kernel method.

Define the kernel function as $K(x) = \int_{-\infty}^{x} \omega(y)dy$, where ω is a probability density function. The smoothed estimator for the generalized Lorenz curve $\theta(t)$ is defined as follows:

$$\hat{T}_n(t) = \frac{1}{n}\sum_{i=1}^{n} X_i K(\frac{t - F_n(X_i)}{h}). \tag{4.2}$$

The asymptotic normality of the smoothed estimator $\hat{T}_n(t)$ is shown in the following Theorem 4.1.

Theorem 4.1. *Assume ω is a probability density function with bounded support and its first derivative exists on its supporting set. If $h = h(n) \to 0$, $\sqrt{n}h^2 \to \infty$ as $n \to \infty$, then*

$$\sqrt{n}\{\hat{T}_n(t) - \theta(t)\} \xrightarrow{d} N(0, \sigma^2(t)),$$

where $\sigma^2(t) = \int_0^{\xi_t} x^2 dF(x) - \theta^2(t) + \xi_t^2 t(1-t)$.

4.1.3 Bandwidth Selection for the Kernel Estimator by Cross-validation Method

One of the difficulties to study kernel estimator is the choice of bandwidth h for the generalized Lorenz curve. In this study, the 2-fold cross-validation method is used to choose bandwidth. The bandwidth h is chosen to be $h = cn^{-1/3}$, based on our simulation experience. Clearly, the constant c will control the choice of bandwidth h. Here and thereafter, we denote $\hat{T}_{n,c}(t) = \hat{T}_n(t)$.

First of all, at a given t, the constant c is chosen by minimizing the Mean Squared Error(MSE).

$$MSE(c) = E[\hat{T}_{n,c}(t) - \theta(t)]^2.$$

By randomly split the sample into two parts, we get a training sample and a validation sample. Based on the training sample, the smoothed Lorenz ordinate $\hat{T}^{(1)}_{n,c}(t)$ is constructed; based on the validation sample, the empirical estimate $\hat{\theta}^{(2)}(t)$ is constructed. By repeating this process a lot of times, the cross-validation estimate of the MSE is constructed as below

$$CV_c = \frac{1}{L}\sum_{l=1}^{L}[\hat{T}^{(1,l)}_{n,c}(t) - \hat{\theta}^{(2,l)}(t)]^2,$$

where L is the number of random splits, and c is chosen as the constant that minimize CV_c.

Alternatively, if we focus on the overall performance of the smoothed estimator for the generalized Lorenz curve across all t, we can use a similar cross-validation procedure for selecting c by minimizing the Average Mean Squared Error (AMSE)

$$AMSE(c) = E\frac{1}{K}\sum_{k=1}^{K}[\hat{T}_{n,c}(t_k) - \theta(t_k)]^2, k = 1, 2, ..., K$$

where t_k is in a fine grid of (0,1), K is an integer.

And the cross-validation estimate of the AMSE is

$$ACV_c = \frac{1}{L}\frac{1}{K}\sum_{l=1}^{L}\sum_{k=1}^{K}[\hat{T}_{n,c}^{(1,l)}(t_k) - \hat{\theta}^{(2,l)}(t_k)]^2.$$

4.1.4 Point Estimator Evaluation

We will evaluate the proposed estimators in terms of Mean Square Error (MSE) and Asymptotic Relative Efficiency (ARE).

The MSE of the empirical estimator $\hat{\theta}(t)$ is

$$MSE_{\hat{\theta}} = E[\hat{\theta}(t) - \theta(t)]^2,$$

and the MSE of the smoothed estimator $\hat{T}_n(t)$ is

$$MSE_{\hat{T}_n(t)} = E[\hat{T}_n(t) - \theta(t)]^2.$$

Note that based on Beach and Davidson (1983), and Zheng (2001), the empirical estimator $\hat{\theta}(t)$ satisfies

$$\sqrt{n}\left(\hat{\theta}(t) - \theta(t)\right) \xrightarrow{d} N(0, \sigma_v^2),$$

where $\sigma_v^2 = \int_0^{\xi_{t_0}}(x - \xi_{t_0})^2 dF(x) - (\theta - t_0\xi_{t_0})^2$.

While the smoothed estimator $\hat{T}_n(t)$ satisfies

$$\sqrt{n}(\hat{T}_n(t) - \theta(t)) \xrightarrow{d} N(0, \sigma^2(t)).$$

Then the ARE of $\hat{T}_n(t)$ with respect to $\hat{\theta}(t)$ is

$$ARE(\hat{T}_n(t), \hat{\theta}(t)) = \frac{\sigma_v^2}{\sigma^2(t)}.$$

$ARE > 1$ implies that the kernel estimator is more efficient than the empirical estimator.

4.2 Smoothed Jackknife Empirical Likelihood for a Generalized Lorenz Curve

Tukey (1958) used the jackknife method to estimate the variance. Based on his method, we define the jackknife pseudo-values as

$$\hat{V}_i(t) = n\hat{T}_n(t) - (n-1)\hat{T}_{n-1,i}(t). \tag{4.3}$$

where $\hat{T}_{n-1,i}(t) = \frac{1}{(n-1)}\sum_{j\neq i} X_j K(\frac{t-F_{n,i}(X_j)}{h})$ is the given statistics T_{n-1} but computed on $n-1$ observations $X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_n$, $F_{n,i}(t) = \frac{1}{n-1}\sum_{j\neq i} I(X_j \leq t)$ is the sample distribution function based on $n-1$ observations.

Thus, the jackknife empirical likelihood for $\theta = \theta(t)$ is defined as below

$$L(t,\theta) = \sup\left\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} np_i = 1, \sum_{i=1}^{n} p_i\hat{V}_i(t) = \theta\right\}. \tag{4.4}$$

By using the Lagrange multiplier method, we obtain the maximization for (4.4) at

$$p_i = \frac{1}{n}\{1 + \lambda[\hat{V}_i(t) - \theta]\}^{-1}, \tag{4.5}$$

where $\lambda = \lambda(t,\theta)$ is the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{V}_i(t) - \theta}{1 + \lambda(\hat{V}_i(t) - \theta)} = 0. \tag{4.6}$$

Since $\prod_{i=1}^{n} p_i$ is subject to $\sum_{i=1}^{n} p_i = 1$, $p_i \geq 0$, i=1,2,..., n, $L(t,\theta)$ will attain its maximization n^{-n} at $p_i = n^{-1}$. Thus, the jackknife empirical likelihood ratio for θ can be defined as

$$L_n(\theta(t)) = \prod_{i=1}^{n}(np_i) = \prod_{i=1}^{n}\{1 + \lambda(\hat{V}_i(t) - \theta)\}^{-1}, \tag{4.7}$$

which gives the log empirical likelihood ratio as

$$l_n(\theta(t)) = -2\log L_n(\theta(t)) = 2\sum_{i=1}^{n}\log\{1 + \lambda(\hat{V}_i(t) - \theta)\}. \tag{4.8}$$

Thus, we derive the following two theorems

Theorem 4.2. *Under conditions of Theorem 4.1, we have*

$$\upsilon_{JACK}(t) \xrightarrow{p} \sigma^2(t), \tag{4.9}$$

where $\sigma^2(t)$ is defined in Theorem 4.1.

Theorem 4.3. *Under the conditions of Theorem 4.1, we have*

$$l_n(\theta(t)) \xrightarrow{d} \chi^2(1). \tag{4.10}$$

where $\upsilon_{JACK}(t)$ is constructed as

$$\begin{aligned} \upsilon_{JACK}(t) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} [\hat{V}_i(t) - \frac{1}{n} \sum_{j=1}^{n} \hat{V}_j(t)]^2 \\ &= \frac{n-1}{n} \sum_{i=1}^{n} [\hat{T}_{n-1,i}(t) - \frac{1}{n} \sum_{j=1}^{n} \hat{T}_{n-1,j}(t)]^2. \end{aligned}$$

4.3 Confidence Intervals for a Generalized Lorenz Curve

4.3.1 Normal Approximation-based Confidence Intervals

In this section, we construct confidence intervals for generalized Lorenz curve (GLC) $\theta(t)$ by normal approximation method. Since we already have two appropriate estimators for $\theta(t)$, we will construct two normal approximation-based confidence intervals for the GLC based on the two estimators.

It has been showed that the estimate $\hat{\theta}(t)$ for generalized Lorenz curve $\theta(t)$ is asymptotically normal with variances σ_v^2 , i.e.,

$$\sqrt{n} \left(\hat{\theta}(t) - \theta(t) \right) \longrightarrow N(0, \sigma_v^2),$$

where

$$\sigma_v^2 = \int_0^{\xi_{t_0}} (x - \xi_{t_0})^2 dF(x) - (\theta - t_0\xi_{t_0})^2.$$

Therefore, a $(1-\alpha)$ level normal approximate (NA1)-based confidence interval for $\theta(t)$ can be constructed as

$$(l_1, u_1) = (\hat{\theta}(t) - \frac{z_{1-\frac{\alpha}{2}}\hat{\sigma}_v}{\sqrt{n}}, \hat{\theta}(t) + \frac{z_{1-\frac{\alpha}{2}}\hat{\sigma}_v}{\sqrt{n}}),$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2}) - th$ quantile of the standard normal distribution, and

$$\hat{\sigma}_v^2 = \int_0^{\hat{\xi}_{t_0}} (x - \hat{\xi}_{t_0})^2 dF_n(x) - (\hat{\theta} - t_0\hat{\xi}_{t_0})^2$$

is a consistent estimate for σ_v^2.

Based on Theorem 4.1, the smoothed estimator $\hat{T}_n(t)$ for the generalized Lorenz curve $\theta(t)$ is asymptotically normal with variances $\sigma^2(t)$, i.e.,

$$\sqrt{n}(\hat{T}_n(t) - \theta(t)) \longrightarrow N(0, \sigma^2(t)).$$

Then we construct another $(1-\alpha)$ level normal approximate (NA2)-based confidence interval for $\theta(t)$ as

$$(l_2, u_2) = (\hat{T}_n(t) - \frac{z_{1-\frac{\alpha}{2}}\hat{\sigma}(t)}{\sqrt{n}}, \hat{T}_n(t) + \frac{z_{1-\frac{\alpha}{2}}\hat{\sigma}(t)}{\sqrt{n}}),$$

where $\hat{\sigma}(t)$ is the standard deviation of the variance estimate for the kernel estimator

$$\hat{\sigma}^2(t) = \int_0^{\hat{\xi}_t} x^2 dF_n(x) - \hat{T}_n^2(t) + \hat{\xi}_t^2 t(1-t) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 I(X_i \leq \hat{\xi}_t) - \hat{T}_n^2(t) + \hat{\xi}_t^2 t(1-t).$$

4.3.2 Bootstrap-based Confidence Intervals

In this section, we apply bootstrap methods to construct confidence intervals for generalized Lorenz curve.

By drawing bootstrap resample $\{X_1^*, X_2^*, X_3^*, ..., X_n^*\}$ with replacement from the original data $\{X_1, X_2, X_3, ..., X_n\}$, the bootstrap versions of the empirical estimator for the generalized Lorenz ordinate $\hat{\theta}(t)$ can be defined as

$$\hat{\theta}^*(t) = \frac{1}{n}\sum_{i=1}^{n} X_i^* I(X_i^* \leq \xi_{t_0}^*).$$

We repeat this bootstrap procedure for B times. Thus, B bootstrap copies of $\hat{\theta}$ are obtained, denoted them as $\{\hat{\theta}_b^*, b = 1, 2, ..., B\}$.

Then, the bootstrap sample variance of $\hat{\theta}_b^*$'s,

$$V_G^* = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\theta}^*)^2,$$

where $\bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$, is used to estimate the asymptotic variance of $\hat{\theta}$.

Two bootstrap confidence intervals based on the empirical estimator for the generalized Lorenz ordinate $\theta(t)$ are constructed as follows:

1. BT1 interval:

$$(l_3, u_3) = (\hat{\theta} - z_{1-\alpha/2}\sqrt{V_G^*}, \hat{\theta} + z_{1-\alpha/2}\sqrt{V_G^*}),$$

2. BT2 interval:

$$(l_4, u_4) = (\bar{\theta}^* - z_{1-\alpha/2}\sqrt{V_G^*}, \bar{\theta}^* + z_{1-\alpha/2}\sqrt{V_G^*}).$$

Lastly, we apply the bootstrap bias correction and acceleration (BCa1) method to construct a confidence interval for $\theta(t)$, which does not need a variance estimation.

3. BCa1 interval:

$$(l_5, u_5) = (\hat{\theta}^*_{([B\beta_1])}, \hat{\theta}^*_{([B\beta_2])}).$$

where

$$\beta_1 = \Phi(b + \frac{b + z_{\alpha/2}}{1 - a(b + z_{\alpha/2})}), \beta_2 = \Phi(b + \frac{b + z_{1-\alpha/2}}{1 - a(b + z_{1-\alpha/2})})$$

with correction constants a and b defined by

$$a = \frac{1}{6}\sum_{i=1}^{n} \varphi_i^3 / (\sum_{i=1}^{n} \varphi_i^2)^{\frac{3}{2}}, b = \Phi^{-1}(\frac{1}{B}\sum_{b=1}^{B} I(\hat{\theta}^*_b \leq \hat{\theta})),$$

where $\varphi_i = \hat{\theta}_{(.)} - \hat{\theta}_{(-i)}$, and $\hat{\theta}_{(-i)}$ is the $\hat{\theta}$ computed by deleting the $i - th$ observation in original data, and $\hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(-i)}$

Similarly, based on the kernel estimator $\hat{T}_n(t)$, three corresponding confidence intervals can also be built. We draw a bootstrap resample $\{X_1^*, X_2^*, X_3^*, ..., X_n^*\}$ with replacement from the original data $\{X_1, X_2, X_3, ..., X_n\}$. The bootstrap version of $\hat{T}_n(t)$ is

$$\hat{T}^*(t) = \frac{1}{n}\sum_{i=1}^{n} x_i^* K(\frac{t - F_n(x_i^*)}{h}).$$

After repeating this bootstrap procedure for B times, B bootstrap copies of $\hat{T}_n$ are obtained, denoted as $\{\hat{T}_b^*, b = 1, 2, ..., B\}$.

The sample variance of $\hat{T}_b^*$'s

$$V_{GT}^* = \frac{1}{B-1}\sum_{b=1}^{B} (\hat{T}_b^* - \bar{T}^*)^2,$$

where $\bar{T}^* = \frac{1}{B}\sum_{b=1}^{B} \hat{T}_b^*$, is used to estimate the asymptotic variance of the kernel estimator $\hat{T}_n(t)$.

Thus, two bootstrap confidence intervals for the generalized Lorenz curve $\theta(t)$ are constructed as follows:

4. BT3 interval:

$$(l_6, u_6) = (\hat{T}_n - z_{1-\alpha/2}\sqrt{V^*_{GT}}, \hat{T}_n + z_{1-\alpha/2}\sqrt{V^*_{GT}}),$$

5. BT4 interval:

$$(l_7, u_7) = (\bar{T}^* - z_{1-\alpha/2}\sqrt{V^*_{GT}}, \bar{T}^* + z_{1-\alpha/2}\sqrt{V^*_{GT}}).$$

Another non-parametric method to construct confidence interval is the bootstrap bias correction and acceleration (BCa2) method, which does not need variance estimation.

6. BCa2 interval:

$$(l_8, u_8) = (\hat{T}^*_{([B\beta_1])}, \hat{T}^*_{([B\beta_2])}).$$

where

$$\beta_1 = \Phi(b + \frac{b + z_{\alpha/2}}{1 - a(b + z_{\alpha/2})}), \beta_2 = \Phi(b + \frac{b + z_{1-\alpha/2}}{1 - a(b + z_{1-\alpha/2})})$$

with correction constants a and b defined by

$$a = \frac{1}{6}\sum_{i=1}^{n}\varphi_i^3/(\sum_{i=1}^{n}\varphi_i^2)^{\frac{3}{2}}, b = \Phi^{-1}(\frac{1}{B}\sum_{b=1}^{B} I(\hat{T}^*_b \leq \hat{T}_n))$$

where $\varphi_i = \hat{T}_{(.)} - \hat{T}_{(-i)}$, and $\hat{T}_{(-i)}$ is the $\hat{T}_n$ computed by deleting the $i-th$ observation in original data, and $\hat{T}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{T}_{(-i)}$.

4.3.3 Smoothed Jackknife Empirical Likelihood-based Confidence Interval

The smoothed version of the jackknife empirical likelihood for the generalized Lorenz curve $\theta(t)$ is derived in Section 4.2. We can construct confidence intervals based on this smoothed jackknife empirical likelihood theory. Based on Theorem 4.3, the SJEL-based confidence interval for the generalized Lorenz curve $\theta(t)$ can be constructed as

$$(l_e, u_e) = \{\theta : l_n(\theta(t)) \leq \chi^2_{1,1-\alpha}\}.$$

Table 4.1 MSE, bias and the percentage of $ARE > 1$ generated from Chi-square distribution(df=3) are compared for empirical estimator and the proposed smoothed estimator for generalized Lorenz curve with t from 0.2 to 0.8

Sample Size	t	$Bias_{\hat{\theta}}$	$Bias_{\hat{T}_n(t)}$	$MSE_{\hat{\theta}}$	$MSE_{\hat{T}_n(t)}$	ARE> 1
100	0.2	0.0031995	0.0008617	0.0004583	0.0004189	97.9%
	0.3	0.0041156	0.0002344	0.0012882	0.0012099	99.7%
	0.4	0.0046863	0.0016318	0.0027291	0.0026133	100%
	0.5	0.0048895	0.0034679	0.0050855	0.0049301	100%
	0.6	0.0050529	0.0055697	0.0088246	0.0086053	100%
	0.7	0.0040055	0.0089308	0.0143087	0.0140426	100%
	0.8	0.0019127	0.0132925	0.0222319	0.0219004	100%
200	0.2	0.0017584	0.0009334	0.0002210	0.0002115	94.6%
	0.3	0.0020155	0.0001939	0.0006501	0.0006302	99.9%
	0.4	0.0021752	0.0005515	0.0014164	0.0013863	100%
	0.5	0.0031560	0.0006204	0.0026262	0.0025793	100%
	0.6	0.0043088	0.0004650	0.0045186	0.0044449	100%
	0.7	0.0057898	0.0000555	0.0074270	0.0073101	100%
	0.8	0.0075163	0.0007350	0.0117703	0.0115827	100%
500	0.2	0.0003534	0.0002641	0.0000898	0.0000884	70.8%
	0.3	0.0005503	0.0000742	0.0002535	0.0002510	98.4%
	0.4	0.0009014	0.0000674	0.0005631	0.0005585	100%
	0.5	0.0012734	0.0000807	0.0010774	0.0010683	100%
	0.6	0.0014360	0.0000924	0.0018319	0.0018216	99.9%
	0.7	0.0017368	0.0001215	0.0029869	0.0029688	99.9%
	0.8	0.0021810	0.0001425	0.0046657	0.0046403	100%

4.4 Numerical Studies and a Real Example

4.4.1 Simulation Studies

In this section, we first compare the empirical estimator and the kernel estimator. Then, the coverage probabilities and interval lengths of the proposed intervals are evaluated by extensive simulation studies. The proposed intervals are also illustrated by a real application.

Recall that ARE of $\hat{T}_n(t)$ with respect to $\hat{\theta}(t)$ is

$$ARE(\hat{T}_n(t), \hat{\theta}(t)) = \frac{\sigma_v^2}{\sigma^2(t)}$$

The evaluation methods are selected to be MSE and ARE. For each setting, 1,000 random sample is generated from the Chi-square distribution with degree of freedom 3, and the sample sizes are chosen to be 100, 200, and 500; t will range from 0.2 to 0.8. Table (4.1) presents the comparisons results. $Bias_{\hat{\theta}}$ presents the bias for the empirical estimator $\hat{\theta}(t)$, and $Bias_{\hat{T}_n(t)}$ presents the bias for the kernel estimator $\hat{T}_n(t)$. We observe that the MSE of the smoothed estimator is much less than the MSE of empirical estimator, although the bias of smoothed estimator is larger than the bias of empirical estimator. Meanwhile, the percentage of $\{ARE > 1\}$ are all larger than 50%. It implies that, most of time, the variance of the smoothed estimator is less than that of the empirical estimator.

After the evaluation for point estimators, we will continue to present results for the coverage probabilities and the average interval lengths of the normal approximation-based confidence intervals, along with the proposed smoothed bootstrap-based confidence intervals and the smoothed jackknife empirical likelihood (SJEL)-based confidence interval.

We generate data from Chi-square distribution with degree of freedom 3, and Wellbull distribution with shape=1 and scale=2, respectively. The sample sizes are chosen to be 100, 200, and 500, separately, and 1,000 random samples are generated from the above distributions. We construct confidence intervals at 95% and 90% confidence level with t=0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. The Quartic/Triweight kernel density function $\omega(x) = \frac{35}{32}(1-t^2)^2 I(|t| \leq 1)$ is selected for the kernel estimator of the generalized Lorenz curve, and the bandwidth $h = cn^{-1/3}$ is chosen via the proposed cross-validation method, where c are valued differently based on different t. For the bootstrap variance estimates, 500 bootstrap re-samples are drawn from the original sample based on $F(x)$.

The coverage probabilities and average lengths of the 90% and 95% confidence levels for generalized Lorenz curve are presented in Table 4.2 to Table 4.5 for Chi-square distribution, and in Table 4.6 to Table 4.9 for Weibull distribution.

We observe that SJEL interval performs better than any other confidence intervals, while BT3 and BT4 intervals perform the second to the best. As sample size increases, all the methods perform better. The average lengths increase as t increases. Therefore, we

recommend that the smoothed jackknife empirical likelihood-based (SJEL) interval and the proposed smoothed bootstrap-based intervals (BT3 and BT4) for the generalized Lorenz curve when income data is right skewed.

4.4.2 A Real Example

More than ten thousands income data for professors in Georgia is extracted from the historical records of the Georgia Public Institutes for 2012 fiscal year. After filter by several criterions, we get 5,921 income data for full-time professors. The salary information for professors who do not provide full-time services in 2012 will not be included. The lower bound and upper bound of the proposed NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL intervals are calculated, and results are summarized in Table (4.10). Based on the SJEL interval, the least wealthy 80% professors have an average annual salary from \$60,973.24 to \$62,282.91. Based on the BT3 interval, it is observed that the mean income of the least wealthy 50% professors is between (31696.16, 32287.59), which indicates that the average annual salary for the least wealthy 50% professors is from \$31,696.16 to \$32,287.59 in 2012.

Table 4.2 Coverage probabilities and interval lengths at 90% confidence level for GLC with Chi-square distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.1 to 0.4

		t=10%		t=20%		t=30%		t=40%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.859	0.0301	0.866	0.0704	0.885	0.1193	0.879	0.1752
	NA2	0.825	0.0298	0.825	0.0706	0.910	0.1180	0.880	0.1740
	BT1	0.926	0.0356	0.902	0.0779	0.909	0.1279	0.909	0.1877
	BT2	0.874	0.0356	0.883	0.0779	0.886	0.1279	0.899	0.1877
	BT3	0.867	0.0276	0.882	0.0681	0.892	0.1143	0.875	0.1684
	BT4	0.872	0.0276	0.890	0.0681	0.887	0.1143	0.877	0.1684
	BCa1	0.884	0.0301	0.884	0.0716	0.884	0.1208	0.896	0.1788
	BCa2	0.810	0.0298	0.860	0.0712	0.867	0.1175	0.877	0.1730
	SJEL	0.884	0.0301	0.882	0.0708	0.887	0.1205	0.896	0.1777
200	NA1	0.905	0.0212	0.906	0.0502	0.904	0.0848	0.903	0.1248
	NA2	0.890	0.0212	0.887	0.0496	0.865	0.0840	0.880	0.1237
	BT1	0.938	0.0232	0.925	0.0530	0.917	0.0878	0.916	0.1295
	BT2	0.904	0.0232	0.902	0.0530	0.907	0.0878	0.907	0.1295
	BT3	0.872	0.0201	0.877	0.0487	0.882	0.0824	0.875	0.1216
	BT4	0.880	0.0201	0.885	0.0487	0.885	0.0824	0.880	0.1216
	BCa1	0.922	0.0213	0.917	0.0510	0.900	0.0854	0.914	0.1267
	BCa2	0.847	0.0208	0.850	0.0499	0.862	0.0836	0.880	0.1232
	SJEL	0.919	0.0219	0.903	0.0504	0.901	0.0849	0.902	0.1247
500	NA1	0.897	0.0131	0.892	0.0313	0.882	0.0533	0.885	0.0784
	NA2	0.932	0.0132	0.875	0.0316	0.885	0.0532	0.885	0.0786
	BT1	0.903	0.0136	0.895	0.0320	0.889	0.0539	0.892	0.0796
	BT2	0.893	0.0136	0.893	0.0320	0.880	0.0539	0.882	0.0796
	BT3	0.895	0.0129	0.880	0.0312	0.875	0.0526	0.885	0.0779
	BT4	0.900	0.0129	0.880	0.0312	0.877	0.0526	0.875	0.0779
	BCa1	0.897	0.0132	0.897	0.0315	0.886	0.0534	0.890	0.0791
	BCa2	0.862	0.0130	0.867	0.0316	0.870	0.0528	0.875	0.0783
	SJEL	0.904	0.0132	0.899	0.0321	0.894	0.0540	0.886	0.0791

Table 4.3 Coverage probabilities and interval lengths at 90% confidence level for GLC with Chi-square distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.5 to 0.9

		t=50%		t=60%		t=70%		t=80%		t=90%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.896	0.2404	0.898	0.3139	0.899	0.3976	0.895	0.4946	0.885	0.6146
	NA2	0.880	0.2342	0.890	0.3053	0.895	0.3916	0.895	0.4904	0.880	0.6022
	BT1	0.917	0.2561	0.916	0.3328	0.919	0.4239	0.916	0.5287	0.915	0.6695
	BT2	0.904	0.2561	0.908	0.3328	0.905	0.4239	0.907	0.5287	0.916	0.6695
	BT3	0.882	0.2322	0.887	0.3037	0.890	0.3850	0.892	0.4824	0.882	0.5877
	BT4	0.875	0.2322	0.887	0.3037	0.880	0.3850	0.887	0.4824	0.875	0.5877
	BCa1	0.912	0.2460	0.907	0.3224	0.912	0.4120	0.909	0.5142	0.904	0.6483
	BCa2	0.887	0.2401	0.890	0.3141	0.877	0.3943	0.880	0.4973	0.872	0.6078
	SJEL	0.891	0.2373	0.890	0.3101	0.897	0.3928	0.894	0.4890	0.892	0.6377
200	NA1	0.913	0.1697	0.907	0.2215	0.900	0.2807	0.904	0.3497	0.898	0.4358
	NA2	0.875	0.1682	0.885	0.2199	0.885	0.2787	0.885	0.3469	0.895	0.4335
	BT1	0.920	0.1750	0.915	0.2275	0.909	0.2897	0.910	0.3608	0.909	0.4545
	BT2	0.907	0.1750	0.907	0.2275	0.902	0.2897	0.903	0.3608	0.906	0.4545
	BT3	0.885	0.1671	0.895	0.2181	0.897	0.2767	0.892	0.3448	0.890	0.4281
	BT4	0.890	0.1671	0.888	0.2181	0.895	0.2767	0.891	0.3448	0.887	0.4281
	BCa1	0.905	0.1716	0.913	0.2238	0.909	0.2861	0.900	0.3562	0.901	0.4480
	BCa2	0.865	0.1704	0.867	0.2217	0.865	0.2803	0.890	0.3486	0.872	0.4344
	SJEL	0.902	0.1692	0.906	0.2208	0.900	0.2796	0.901	0.3482	0.898	0.4340
500	NA1	0.917	0.1069	0.920	0.1397	0.907	0.1777	0.908	0.2219	0.894	0.2752
	NA2	0.905	0.1065	0.925	0.1391	0.932	0.1767	0.930	0.2210	0.910	0.2762
	BT1	0.919	0.1082	0.922	0.1412	0.909	0.1801	0.913	0.2246	0.899	0.2797
	BT2	0.922	0.1082	0.924	0.1412	0.913	0.1801	0.910	0.2246	0.900	0.2797
	BT3	0.907	0.1058	0.905	0.1391	0.902	0.1761	0.903	0.2200	0.905	0.2746
	BT4	0.902	0.1058	0.907	0.1391	0.905	0.1761	0.912	0.2200	0.907	0.2746
	BCa1	0.920	0.1074	0.913	0.1402	0.913	0.1793	0.908	0.2232	0.899	0.2783
	BCa2	0.912	0.1063	0.927	0.1402	0.902	0.1765	0.927	0.2209	0.887	0.2768
	SJEL	0.914	0.1076	0.920	0.1401	0.908	0.1781	0.908	0.2223	0.898	0.2755

Table 4.4 Coverage probabilities and interval lengths at 95% confidence level for GLC with Chi-square distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.1 to 0.4

		t=10%		t=20%		t=30%		t=40%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.928	0.0359	0.924	0.0839	0.935	0.1422	0.937	0.2088
	NA2	0.915	0.0355	0.925	0.0842	0.917	0.1406	0.962	0.2074
	BT1	0.971	0.0425	0.951	0.0929	0.949	0.1524	0.957	0.2241
	BT2	0.936	0.0425	0.940	0.0929	0.942	0.1524	0.950	0.2241
	BT3	0.917	0.0329	0.947	0.0815	0.945	0.1361	0.950	0.2009
	BT4	0.917	0.0329	0.945	0.0815	0.952	0.1361	0.950	0.2009
	BCa1	0.934	0.0350	0.943	0.0851	0.942	0.1439	0.950	0.2136
	BCa2	0.870	0.0352	0.907	0.0848	0.922	0.1393	0.930	0.2055
	SJEL	0.929	0.0359	0.935	0.0841	0.937	0.1429	0.953	0.2155
200	NA1	0.958	0.0252	0.955	0.0598	0.957	0.1011	0.953	0.1487
	NA2	0.910	0.0253	0.902	0.0592	0.937	0.1001	0.925	0.1474
	BT1	0.973	0.0276	0.966	0.0629	0.968	0.1046	0.962	0.1542
	BT2	0.955	0.0276	0.959	0.0629	0.954	0.1046	0.961	0.1542
	BT3	0.927	0.0239	0.927	0.0578	0.927	0.0978	0.912	0.1445
	BT4	0.930	0.0239	0.930	0.0578	0.930	0.0978	0.907	0.1445
	BCa1	0.962	0.0253	0.967	0.0605	0.961	0.1017	0.953	0.1511
	BCa2	0.892	0.0249	0.917	0.0590	0.932	0.0987	0.920	0.1456
	SJEL	0.956	0.0265	0.949	0.0599	0.960	0.1009	0.948	0.1483
500	NA1	0.945	0.0156	0.940	0.0373	0.940	0.0635	0.937	0.0935
	NA2	0.940	0.0157	0.937	0.0376	0.942	0.0634	0.937	0.0937
	BT1	0.957	0.0162	0.948	0.0381	0.946	0.0642	0.944	0.0950
	BT2	0.949	0.0162	0.942	0.0381	0.948	0.0642	0.945	0.0950
	BT3	0.937	0.0153	0.937	0.0372	0.915	0.0627	0.930	0.0930
	BT4	0.942	0.0153	0.935	0.0372	0.920	0.0627	0.930	0.0930
	BCa1	0.943	0.0157	0.948	0.0375	0.941	0.0634	0.939	0.0941
	BCa2	0.930	0.0156	0.927	0.0375	0.915	0.0630	0.920	0.0934
	SJEL	0.952	0.0164	0.944	0.0381	0.942	0.0641	0.949	0.0948

Table 4.5 Coverage probabilities and interval lengths at 95% confidence level for GLC with Chi-square distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.5 to 0.9

		t=50%		t=60%		t=70%		t=80%		t=90%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.946	0.2864	0.946	0.3740	0.945	0.4738	0.946	0.5894	0.935	0.7324
	NA2	0.947	0.2791	0.945	0.3638	0.942	0.4667	0.945	0.5843	0.937	0.7175
	BT1	0.964	0.3053	0.959	0.3957	0.959	0.5060	0.965	0.6301	0.951	0.7961
	BT2	0.954	0.3053	0.950	0.3957	0.956	0.5060	0.965	0.6301	0.952	0.7961
	BT3	0.930	0.2762	0.940	0.3623	0.935	0.4587	0.945	0.5738	0.942	0.7004
	BT4	0.932	0.2762	0.932	0.3623	0.932	0.4587	0.937	0.5738	0.927	0.7004
	BCa1	0.953	0.2926	0.953	0.3827	0.959	0.4917	0.964	0.6142	0.955	0.7728
	BCa2	0.932	0.2841	0.935	0.3735	0.940	0.4690	0.942	0.5886	0.922	0.7189
	SJEL	0.948	0.2822	0.948	0.3718	0.946	0.4673	0.946	0.5819	0.954	0.7736
200	NA1	0.954	0.2022	0.948	0.2639	0.952	0.3345	0.951	0.4167	0.949	0.5193
	NA2	0.952	0.2004	0.937	0.2621	0.940	0.3321	0.947	0.4133	0.940	0.5166
	BT1	0.959	0.2084	0.957	0.2714	0.955	0.3458	0.956	0.4305	0.962	0.5419
	BT2	0.956	0.2084	0.952	0.2714	0.959	0.3458	0.965	0.4305	0.954	0.5419
	BT3	0.927	0.1986	0.922	0.2595	0.942	0.3287	0.947	0.4102	0.937	0.5093
	BT4	0.922	0.1986	0.922	0.2595	0.935	0.3287	0.940	0.4102	0.940	0.5093
	BCa1	0.956	0.2043	0.950	0.2672	0.957	0.3413	0.958	0.4250	0.957	0.5349
	BCa2	0.927	0.2012	0.935	0.2631	0.942	0.3336	0.945	0.4147	0.935	0.5175
	SJEL	0.952	0.2014	0.951	0.2628	0.950	0.3327	0.949	0.4145	0.947	0.5169
500	NA1	0.960	0.1274	0.962	0.1665	0.954	0.2117	0.945	0.2644	0.940	0.3279
	NA2	0.962	0.1269	0.960	0.1657	0.957	0.2106	0.960	0.2634	0.937	0.3292
	BT1	0.965	0.1292	0.963	0.1683	0.960	0.2144	0.949	0.2678	0.943	0.3331
	BT2	0.964	0.1292	0.964	0.1683	0.960	0.2144	0.950	0.2678	0.941	0.3331
	BT3	0.955	0.1266	0.952	0.1649	0.955	0.2090	0.960	0.2621	0.950	0.3267
	BT4	0.957	0.1266	0.955	0.1649	0.952	0.2090	0.957	0.2621	0.947	0.3267
	BCa1	0.960	0.1284	0.953	0.1671	0.953	0.2131	0.947	0.2662	0.942	0.3314
	BCa2	0.962	0.1274	0.945	0.1654	0.970	0.2098	0.957	0.2625	0.940	0.3297
	SJEL	0.959	0.1279	0.961	0.1669	0.955	0.2121	0.950	0.2646	0.948	0.3321

Table 4.6 Coverage probabilities and interval lengths at 90% confidence level for GLC with Weibull distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.1 to 0.4

		t=10%		t=20%		t=30%		t=40%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.889	0.0132	0.878	0.0367	0.865	0.0685	0.867	0.1083
	NA2	0.880	0.0132	0.877	0.0368	0.865	0.0691	0.860	0.1086
	BT1	0.935	0.0156	0.923	0.0405	0.901	0.0739	0.895	0.1168
	BT2	0.906	0.0156	0.886	0.0405	0.887	0.0739	0.891	0.1168
	BT3	0.882	0.0121	0.885	0.0347	0.895	0.0664	0.895	0.1056
	BT4	0.890	0.0121	0.892	0.0347	0.895	0.0664	0.891	0.1056
	BCa1	0.891	0.0124	0.900	0.0361	0.884	0.0683	0.888	0.1098
	BCa2	0.810	0.0115	0.857	0.0355	0.855	0.0682	0.875	0.1088
	SJEL	0.896	0.0135	0.890	0.0363	0.884	0.0684	0.884	0.1098
200	NA1	0.898	0.0089	0.911	0.0258	0.902	0.0485	0.895	0.0768
	NA2	0.870	0.0091	0.885	0.0256	0.915	0.0483	0.908	0.0761
	BT1	0.933	0.0098	0.928	0.0272	0.915	0.0502	0.904	0.0796
	BT2	0.915	0.0098	0.919	0.0272	0.899	0.0502	0.902	0.0796
	BT3	0.877	0.0086	0.882	0.0248	0.907	0.0469	0.902	0.0748
	BT4	0.885	0.0086	0.890	0.0248	0.904	0.0469	0.903	0.0748
	BCa1	0.914	0.0087	0.917	0.0257	0.908	0.0484	0.904	0.0773
	BCa2	0.845	0.0083	0.852	0.0250	0.887	0.0475	0.895	0.0756
	SJEL	0.904	0.0097	0.909	0.0263	0.902	0.0489	0.895	0.0768
500	NA1	0.884	0.0055	0.892	0.0162	0.902	0.0306	0.899	0.0485
	NA2	0.917	0.0056	0.885	0.0162	0.910	0.0306	0.885	0.0485
	BT1	0.900	0.0058	0.901	0.0166	0.901	0.0310	0.904	0.0493
	BT2	0.895	0.0058	0.896	0.0166	0.895	0.0310	0.896	0.0493
	BT3	0.899	0.0055	0.898	0.0160	0.902	0.0303	0.897	0.0482
	BT4	0.895	0.0055	0.896	0.0160	0.901	0.0303	0.902	0.0482
	BCa1	0.896	0.0055	0.898	0.0162	0.901	0.0306	0.900	0.0488
	BCa2	0.875	0.0053	0.882	0.0160	0.885	0.0304	0.882	0.0484
	SJEL	0.908	0.0060	0.902	0.0168	0.903	0.0307	0.908	0.0493

Table 4.7 Coverage probabilities and interval lengths at 90% confidence level for GLC with Weibull distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL with different sample sizes and t from 0.5 to 0.9

		t=50%		t=60%		t=70%		t=80%		t=90%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.897	0.1589	0.889	0.2160	0.883	0.2852	0.887	0.3685	0.893	0.4777
	NA2	0.867	0.1575	0.870	0.2150	0.867	0.2823	0.865	0.3666	0.897	0.4689
	BT1	0.919	0.1691	0.907	0.2294	0.904	0.3032	0.906	0.3952	0.916	0.5226
	BT2	0.899	0.1691	0.898	0.2294	0.900	0.3032	0.902	0.3952	0.916	0.5226
	BT3	0.885	0.1540	0.891	0.2104	0.896	0.2787	0.904	0.3601	0.890	0.4562
	BT4	0.887	0.1540	0.898	0.2104	0.893	0.2787	0.902	0.3601	0.892	0.4562
	BCa1	0.907	0.1614	0.904	0.2205	0.898	0.2928	0.899	0.3827	0.912	0.5037
	BCa2	0.852	0.1587	0.852	0.2156	0.842	0.2862	0.847	0.3722	0.872	0.4735
	SJEL	0.896	0.1565	0.890	0.2130	0.891	0.2812	0.893	0.3640	0.890	0.5017
200	NA1	0.910	0.1115	0.904	0.1526	0.906	0.2015	0.907	0.2616	0.892	0.3370
	NA2	0.865	0.1110	0.865	0.1518	0.870	0.2012	0.882	0.2613	0.885	0.3344
	BT1	0.923	0.1149	0.918	0.1574	0.912	0.2082	0.918	0.2709	0.906	0.3506
	BT2	0.915	0.1149	0.914	0.1574	0.903	0.2082	0.919	0.2709	0.903	0.3506
	BT3	0.894	0.1096	0.885	0.1502	0.894	0.1997	0.882	0.2581	0.892	0.3296
	BT4	0.890	0.1096	0.892	0.1502	0.906	0.1997	0.890	0.2581	0.905	0.3296
	BCa1	0.919	0.1122	0.908	0.1544	0.911	0.2048	0.909	0.2672	0.898	0.3449
	BCa2	0.897	0.1111	0.877	0.1517	0.885	0.2023	0.877	0.2621	0.877	0.3342
	SJEL	0.905	0.1111	0.909	0.1522	0.905	0.2009	0.909	0.2605	0.892	0.3354
500	NA1	0.887	0.0702	0.884	0.0962	0.890	0.1275	0.892	0.1656	0.894	0.2138
	NA2	0.920	0.0695	0.912	0.0954	0.910	0.1264	0.920	0.1644	0.892	0.2143
	BT1	0.885	0.0709	0.890	0.0974	0.899	0.1292	0.894	0.1676	0.899	0.2173
	BT2	0.891	0.0709	0.882	0.0974	0.890	0.1292	0.897	0.1676	0.897	0.2173
	BT3	0.910	0.0692	0.905	0.0951	0.901	0.1260	0.904	0.1632	0.900	0.2126
	BT4	0.912	0.0692	0.900	0.0951	0.905	0.1260	0.902	0.1632	0.901	0.2126
	BCa1	0.889	0.0703	0.884	0.0966	0.889	0.1283	0.892	0.1665	0.894	0.2163
	BCa2	0.915	0.0695	0.910	0.0960	0.905	0.1267	0.917	0.1642	0.891	0.2143
	SJEL	0.888	0.0709	0.887	0.0969	0.896	0.1281	0.891	0.1660	0.898	0.2142

Table 4.8 Coverage probabilities and interval lengths at 95% confidence level for GLC with Weibull distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with different sample sizes and t from 0.1 to 0.4

		t=10%		t=20%		t=30%		t=40%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.922	0.0157	0.932	0.0437	0.921	0.0816	0.924	0.1291
	NA2	0.920	0.0157	0.937	0.0439	0.925	0.0823	0.920	0.1294
	BT1	0.963	0.0186	0.955	0.0484	0.949	0.0882	0.952	0.1388
	BT2	0.951	0.0186	0.948	0.0484	0.943	0.0882	0.937	0.1388
	BT3	0.962	0.0144	0.954	0.0415	0.951	0.0791	0.953	0.1262
	BT4	0.957	0.0144	0.945	0.0415	0.945	0.0791	0.940	0.1262
	BCa1	0.934	0.0146	0.944	0.0430	0.934	0.0815	0.946	0.1307
	BCa2	0.892	0.0136	0.907	0.0423	0.922	0.0810	0.927	0.1288
	SJEL	0.939	0.0164	0.948	0.0442	0.942	0.0824	0.944	0.1373
200	NA1	0.942	0.0106	0.951	0.0308	0.946	0.0578	0.943	0.0915
	NA2	0.937	0.0109	0.932	0.0305	0.945	0.0575	0.947	0.0907
	BT1	0.965	0.0117	0.963	0.0324	0.955	0.0599	0.955	0.0949
	BT2	0.958	0.0117	0.959	0.0324	0.957	0.0599	0.951	0.0949
	BT3	0.962	0.0103	0.967	0.0295	0.955	0.0559	0.945	0.0891
	BT4	0.957	0.0103	0.957	0.0295	0.955	0.0559	0.950	0.0891
	BCa1	0.952	0.0103	0.958	0.0306	0.950	0.0579	0.950	0.0921
	BCa2	0.962	0.0098	0.955	0.0294	0.940	0.0565	0.950	0.0900
	SJEL	0.946	0.0115	0.945	0.0312	0.948	0.0580	0.947	0.0913
500	NA1	0.945	0.0066	0.945	0.0193	0.951	0.0365	0.951	0.0578
	NA2	0.940	0.0067	0.940	0.0193	0.945	0.0360	0.950	0.0578
	BT1	0.955	0.0069	0.950	0.0198	0.957	0.0370	0.951	0.0588
	BT2	0.948	0.0069	0.942	0.0198	0.955	0.0370	0.946	0.0588
	BT3	0.957	0.0065	0.947	0.0190	0.947	0.0362	0.950	0.0572
	BT4	0940	0.0065	0.947	0.0190	0.947	0.0362	0.950	0.0572
	BCa1	0.954	0.0066	0.949	0.0193	0.948	0.0364	0.949	0.0581
	BCa2	0.935	0.0063	0.927	0.0189	0.937	0.0363	0.930	0.0572
	SJEL	0.960	0.0067	0.953	0.0202	0.951	0.0373	0.950	0.0586

Table 4.9 Coverage probabilities and interval lengths at 95% confidence level for GLC with Weibull distribution are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL with different sample sizes and t from 0.5 to 0.9

		t=50%		t=60%		t=70%		t=80%		t=90%	
Size	Method	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	NA1	0.951	0.1893	0.949	0.2574	0.942	0.3399	0.938	0.4391	0.938	0.5693
	NA2	0.927	0.1877	0.925	0.2562	0.947	0.3364	0.922	0.4368	0.930	0.5588
	BT1	0.956	0.2013	0.960	0.2735	0.959	0.3622	0.956	0.4701	0.956	0.6221
	BT2	0.956	0.2013	0.950	0.2735	0.950	0.3622	0.955	0.4701	0.947	0.6221
	BT3	0.957	0.1835	0.956	0.2510	0.957	0.3311	0.960	0.4297	0.957	0.5447
	BT4	0.961	0.1835	0.950	0.2510	0.952	0.3311	0.943	0.4297	0.946	0.5447
	BCa1	0.957	0.1916	0.958	0.2629	0.952	0.3500	0.948	0.4569	0.95	0.6033
	BCa2	0.952	0.1891	0.947	0.2576	0.958	0.3419	0.956	0.4437	0.922	0.5648
	SJEL	0.948	0.1860	0.949	0.2532	0.946	0.3344	0.943	0.4329	0.951	0.6015
200	NA1	0.953	0.1328	0.953	0.1819	0.947	0.2401	0.946	0.3118	0.938	0.4015
	NA2	0.947	0.1323	0.945	0.1809	0.942	0.2398	0.947	0.3113	0.932	0.3984
	BT1	0.960	0.1367	0.956	0.1879	0.955	0.2477	0.953	0.3218	0.943	0.4185
	BT2	0.954	0.1367	0.960	0.1879	0.959	0.2477	0.956	0.3218	0.947	0.4185
	BT3	0.946	0.1309	0.944	0.1793	0.945	0.2367	0.943	0.3083	0.945	0.3928
	BT4	0.942	0.1309	0.960	0.1793	0.946	0.2367	0.957	0.3083	0.946	0.3928
	BCa1	0.950	0.1333	0.952	0.1843	0.959	0.2435	0.957	0.3173	0.942	0.4122
	BCa2	0.947	0.1329	0.942	0.1808	0.930	0.2404	0.930	0.3146	0.922	0.3966
	SJEL	0.952	0.1321	0.951	0.1810	0.947	0.2390	0.950	0.3100	0.946	0.4134
500	NA1	0.941	0.0837	0.941	0.1146	0.944	0.1519	0.946	0.1974	0.943	0.2548
	NA2	0.930	0.0828	0.942	0.1137	0.955	0.1506	0.962	0.1960	0.945	0.2554
	BT1	0.945	0.0846	0.946	0.1160	0.941	0.1538	0.951	0.1998	0.945	0.2585
	BT2	0.942	0.0846	0.946	0.1160	0.942	0.1538	0.946	0.1998	0.948	0.2585
	BT3	0.950	0.0826	0.950	0.1133	0.950	0.1496	0.950	0.1942	0.947	0.2525
	BT4	0.947	0.0826	0.948	0.1133	0.952	0.1496	0.949	0.1942	0.945	0.2525
	BCa1	0.947	0.0837	0.946	0.1151	0.946	0.1526	0.943	0.1986	0.943	0.2571
	BCa2	0.950	0.0828	0.950	0.1137	0.945	0.1499	0.965	0.1949	0.930	0.2531
	SJEL	0.947	0.0843	0.945	0.1152	0.948	0.1524	0.949	0.1976	0.944	0.2550

Table 4.10 Georgia Individual Income Example: 95% confidence interval and interval length for GLC for professor's real income data are reported based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2 and SJEL methods with t from 0.5 to 0.8.

quantile	Method	Confidence Interval	Length
0.5	NA1	(32004.46, 32004.46)	0.0000
	NA2	(30651.73, 33332.02)	2680.2
	BT1	(31694.22, 32314.70)	620.48
	BT2	(31695.82, 32316.30)	620.48
	BT3	(31696.16, 32287.59)	591.43
	BT4	(31690.49, 32281.93)	591.43
	BCa1	(31664.83, 32306.79)	641.96
	BCa2	(31686.21, 32306.68)	641.96
	SJEL	(31717.33, 32332.47)	615.14
0.6	NA1	(40656.66, 40656.67)	0.0100
	NA2	(39202.91, 42092.21)	2889.3
	BT1	(40274.97, 41038.37)	763.40
	BT2	(40281.74, 41045.14)	763.40
	BT3	(40244.99, 41050.12)	805.13
	BT4	(40230.78, 41035.92)	805.13
	BCa1	(40271.12, 41030.95)	759.83
	BCa2	(40272.49, 41105.52)	759.83
	SJEL	(40227.26, 41023.32)	796.06
0.7	NA1	(50390.88, 50390.89)	0.0100
	NA2	(48871.04, 51894.35)	3023.31
	BT1	(49913.24, 50868.52)	955.28
	BT2	(49904.21, 50859.49)	955.28
	BT3	(49899.60, 50865.78)	966.18
	BT4	(49883.95, 50850.13)	966.18
	BCa1	(49956.38, 50911.49)	955.11
	BCa2	(49955.02, 51020.14)	1065.12
	SJEL	(49989.57, 50924.68)	935.11
0.8	NA1	(61603.17, 61603.18)	0.0100
	NA2	(60049.61, 63142.43)	3092.82
	BT1	(60936.02, 62270.33)	1334.31
	BT2	(60969.15, 62303.46)	1334.31
	BT3	(60961.48, 62230.56)	1269.08
	BT4	(60932.65, 62201.74)	1269.08
	BCa1	(60857.96, 62272.20)	1414.24
	BCa2	(60979.48, 62249.00)	1269.52
	SJEL	(60973.24, 62282.91)	1309.67

4.5 Discussion

A new kernel estimator has been proposed for the generalized Lorenz curve, and is compared with the empirical estimator. Then, we illustrate the 2-fold cross-validation method to choose bandwidth h for the kernel estimator.

Meanwhile, we also propose the smoothed jackknife empirical likelihood-based confidence interval (SJEL) and the smoothed bootstrap-based confidence intervals (BT3 and BT4). These intervals are compared with the normal approximation-based confidence intervals. SJEL, BT3 and BT4 intervals are observed to have the best performance than any other intervals in most cases. The proposed smoothed jackknife empirical likelihood-based method combines the powers of both jackknife and empirical likelihood methods. Based on this study, we recommend the smoothed estimator for the generalized Lorenz curve, as well as the proposed confidence intervals.

4.6 Proof

Proof of Theorem 4.1, Theorem 4.2 and Theorem 4.3

Theorem 4.1. Under the conditions in Theorem 4.1, we have

$$\sqrt{n}\{\hat{T}_n(t) - \theta(t)\} \xrightarrow{d} N(0, \sigma^2(t)).$$

Proof:

We have the following decomposition

$$\begin{aligned}
&\sqrt{n}[\hat{T}_n(t) - \theta(t)] \\
&= \sqrt{n}[\frac{1}{n}\sum_{i=1}^{n} X_i K(\frac{t - F_n(X_i)}{h}) - \int_0^{\xi_t} x dF(x)] \\
&= \sqrt{n}[\frac{1}{n}\sum_{i=1}^{n} X_i K(\frac{t - F_n(X_i)}{h}) - \frac{1}{n}\sum_{i=1}^{n} X_i K(\frac{t - F(X_i)}{h})] \\
&+\sqrt{n}[\frac{1}{n}\sum_{i=1}^{n} X_i K(\frac{t - F(X_i)}{h}) - \int_0^{\xi_t} x dF(x)] \\
&\equiv I_1 + I_2. \qquad (4.11)
\end{aligned}$$

Then I_1 of (4.11) can be written as

$$\begin{aligned}
I_1 \quad &= \frac{\sqrt{n}}{n}\sum_{i=1}^{n} X_i [K(\frac{t - F_n(X_i)}{h}) - K(\frac{t - F(X_i)}{h})] \\
&= \int_{-\infty}^{\infty} x[K(\frac{t - F_n(x)}{h}) - K(\frac{t - F(x)}{h})] d[\sqrt{n}(F_n(x) - F(x))] \\
&+\sqrt{n}\int_{-\infty}^{\infty} x[K(\frac{t - F_n(x)}{h}) - K(\frac{t - F(x)}{h})] dF(x) \\
&\equiv I_{11} + I_{12}. \qquad (4.12)
\end{aligned}$$

By Taylor Series, under condition of Theorem 4.1, I_{12} of (4.12) can be written as

$$
\begin{aligned}
I_{12} \quad &= \sqrt{n}\int_{-\infty}^{\infty} x[K(\frac{t-F_n(x)}{h}) - K(\frac{t-F(x)}{h})]dF(x) \\
&= \sqrt{n}\int_{-\infty}^{\infty} x[\omega(\frac{t-F(x)}{h})\frac{F(x)-F_n(x)}{h} + \frac{1}{2}\omega'(\frac{t-F(x)}{h})(\frac{F(x)-F_n(x)}{h})^2]dF(x) \\
&= -\sqrt{n}\int_{-\infty}^{\infty} x[\omega(\frac{t-F(x)}{h})\frac{F_n(x)-F(x)}{h}]dF(x) + O_p(\frac{1}{\sqrt{n}h^2}) \\
&= -\int_{-\infty}^{\infty} x\omega(\frac{t-F(x)}{h})\frac{\sqrt{n}(F_n(x)-F(x))}{h}dF(x) + O_p(\frac{1}{\sqrt{n}h^2}). \qquad (4.13)
\end{aligned}
$$

Let Y=F(X), thus $X = F^{-1}(Y)$, and let

$$U_n(y) = \sqrt{n}[\frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y) - y].$$

Clearly Y_i follows uniform[0,1] distribution. Since $h \to 0$, the support of $\omega(x)$ is bounded by $(0, \infty)$, so (4.13) will be equal to

$$
\begin{aligned}
I_{12} \quad &= -\int_{-1}^{1} F^{-1}(y)\omega(\frac{t-y}{h})\frac{1}{h}U_n(y)dy + o_p(1) \\
&= \int_{\frac{t-1}{h}}^{\frac{t+1}{h}} F^{-1}(t-hz)\omega(z)U_n(t-hz)dz + o_p(1) \\
&= \int_{-a}^{a} F^{-1}(t)\omega(z)U_n(t)dz + o_p(1) \\
&= F^{-1}(t)U_n(t)\int_{-a}^{a}\omega(z)dz + o_p(1) \\
&= \xi_t U_n(t) + o_p(1). \qquad (4.14)
\end{aligned}
$$

Since $\sqrt{n}[F_n(x) - F(x)] \to B(x)$, which is Gaussian Process, so $I_{11} = o_p(1)$. Therefore, $I_1 = \xi_t U_n(t) + o_p(1)$.

Next, let's consider I_2 of (4.11). We are going to prove

$$EXK(\frac{t-F(x)}{h}) = \int_{-\infty}^{\infty} xK(\frac{t-F(x)}{h})]dF(x) \longrightarrow \int_{0}^{\xi_t} xdF(x) = \theta(t), as\ h \to 0. \qquad (4.15)$$

and

$$EX^2K^2(\frac{t-F(x)}{h}) = \int_{-\infty}^{\infty} x^2K^2(\frac{t-F(x)}{h})]dF(x) \longrightarrow \int_0^{\xi_t} x^2dF(x), as\ h \to 0. \tag{4.16}$$

Notice that

$$\begin{aligned}
&\lim_{h\to 0}\int_{-\infty}^{\infty} xK(\frac{t-F(x)}{h})f(x)dx \\
&= \lim_{h\to 0}\int_{-\infty}^{\infty} x\int_{-\infty}^{\frac{t-F(x)}{h}} \omega(y)dyf(x)dx \\
&= \int_{-\infty}^{\infty} x[\lim_{h\to 0}\int_{-\infty}^{\frac{t-F(x)}{h}} \omega(y)dy]f(x)dx \\
&= \int_{-\infty}^{\infty} x\{0 * I[t < F(x)] + \frac{1}{2} * I[t = F(x)] + 1 * I[t > F(x)]\}f(x)dx \\
&= \int_{-\infty}^{\infty} xI[t > F(x)]f(x)dx \\
&= \int_{-\infty}^{\infty} xI[t > F(x)]dF(x) \\
&= \int_0^{\xi_t} xdF(x) \\
&= \theta(t).
\end{aligned} \tag{4.17}$$

Similarly, we have

$$\begin{aligned}
&\lim_{h\to 0}\int_{-\infty}^{\infty} x^2K^2(\frac{t-F(x)}{h})f(x)dx \\
&=\lim_{h\to 0}\int_{-\infty}^{\infty} x^2(\int_{-\infty}^{\frac{t-F(x)}{h}}\omega(y)dy)^2f(x)dx \\
&=\int_{-\infty}^{\infty} x^2[\lim_{h\to 0}\int_{-\infty}^{\frac{t-F(x)}{h}}\omega(y)dy]^2f(x)dx \\
&=\int_{-\infty}^{\infty} x^2\{0*I[t<F(x)]+\frac{1}{2}*I[t=F(x)]+1*I[t>F(x)]\}^2f(x)dx \\
&=\int_{-\infty}^{\infty} x^2\{I[t>F(x)]\}^2f(x)dx \\
&=\int_{-\infty}^{\infty} x^2I[t>F(x)]dF(x) \\
&=\int_{0}^{\xi_t} x^2dF(x). \qquad (4.18)
\end{aligned}$$

Define $\omega_i = X_iK(\frac{t-F(X_i)}{h})$ and $\omega = xK(\frac{t-F(x)}{h})$. So I_2 of (4.11) can be rewritten as

$$\begin{aligned}
I_2 \quad &= \sqrt{n}[\frac{1}{n}\sum_{i=1}^{n}X_iK(\frac{t-F(X_i)}{h}) - EXK(\frac{t-F(X)}{h})] + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\omega_i - E\omega) + o_p(1). \qquad (4.19)
\end{aligned}$$

Since $U_n(t)$ is an ancillary statistics. Based on Basu's Lemma in Shao(2003), $U_n(t)$ is independent of $\frac{1}{n}\sum_{i=1}^{n}(\omega_i - E\omega)$.

Also,

$$
\begin{aligned}
& Var[\xi_t U_n(t)) + Var(\omega)] \\
& = \xi_t^2 t(1-t) + Var[XK(\frac{t-F(X)}{h})] \\
& = \xi_t^2 t(1-t) + EX^2K^2(\frac{t-F(X)}{h}) - [EXK(\frac{t-F(X)}{h})]^2 \\
& \longrightarrow \xi_t^2 t(1-t) + \int_0^{\xi_t} x^2 dF(x) - [\int_0^{\xi_t} x dF(x)]^2 \\
& = \xi_t^2 t(1-t) + \int_0^{\xi_t} x^2 dF(x) - \theta^2(t) \\
& \equiv \sigma^2(t). \qquad (4.20)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& I_1 + I_2 \\
& = \xi_t U_n(t) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_i - E\omega) + o_p(1) \\
& \xrightarrow{d} N(0, \sigma^2(t)). \qquad (4.21)
\end{aligned}
$$

We need Lemma 1 and Lemma 2 to prove Theorem 4.2.

Lemma 1. Under the conditions in Theorem 4.1, we have

$$\sqrt{n}\{\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k(t)-\theta(t)\} \xrightarrow{d} N(0,\sigma^2(t)), \tag{4.22}$$

where $\sigma^2(t)$ is defined in Theorem 4.1.

Proof: Note that $\frac{1}{n}\sum_{k=1}^{n}V_k(t)$ can be decomposed into

$$\frac{1}{n}\sum_{k=1}^{n}V_k(t)=\frac{n-1}{n}\sum_{k=1}^{n}(\hat{T}_n-\hat{T}_{n-1,k})+\hat{T}_n, \tag{4.23}$$

while

$$\begin{aligned}
&\hat{T}_n-\hat{T}_{n-1,k}\\
&=\frac{1}{n}\sum_{i=1}^{n}X_iK(\frac{t-F_n(X_i)}{h})-\frac{1}{(n-1)}\sum_{j\neq k}X_jK(\frac{t-F_{n,k}(X_j)}{h})\\
&=\frac{1}{n}\sum_{i=1}^{n}X_iK(\frac{t-F_n(X_i)}{h})-\frac{1}{n}\sum_{j\neq k}^{n}X_jK(\frac{t-F_{n,k}(X_j)}{h})\\
&+\frac{1}{n}\sum_{j\neq k}^{n}X_jK(\frac{t-F_{n,k}(X_j)}{h})-\frac{1}{(n-1)}\sum_{j\neq k}X_jK(\frac{t-F_{n,k}(X_j)}{h}).
\end{aligned} \tag{4.24}$$

So

$$
\begin{aligned}
&\sum_{k=1}^{n}(\hat{T}_n - \hat{T}_{n-1,k}) \\
&= \frac{1}{n}\{\sum_{k=1}^{n}\sum_{i=1}^{n} X_i K(\frac{t-F_n(X_i)}{h}) - \sum_{k=1}^{n}\sum_{j\neq k} X_j[K(\frac{t-F_{n,k}(X_j)}{h})]\} \\
&+ \sum_{k=1}^{n}[\{\frac{1}{n}\sum_{j\neq k} X_j K(\frac{t-F_{n,k}(X_j)}{h}) - \frac{1}{n-1}\sum_{j\neq k} X_j K(\frac{t-F_{n,k}(X_j)}{h})\} \\
&= \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n} X_i[K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})] + \frac{1}{n}\sum_{k=1}^{n} X_k K(\frac{t-F_{n,k}(X_k)}{h}) \\
&+ \frac{1}{n}\sum_{k=1}^{n}\sum_{j\neq k} X_j K(\frac{t-F_{n,k}(X_j)}{h}) - \frac{1}{n-1}\sum_{k=1}^{n}\sum_{j\neq k} X_j K(\frac{t-F_{n,k}(X_j)}{h}) \\
&= \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n} X_i[K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})] \\
&+ \{\frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{n} X_j K(\frac{t-F_{n,k}(X_j)}{h}) - \frac{1}{n-1}\sum_{k=1}^{n}\sum_{j=1}^{n} X_j K(\frac{t-F_{n,k}(X_j)}{h}) \\
&+ \frac{1}{n-1}\sum_{k=1}^{n} X_k K(\frac{t-F_{n,k}(X_k)}{h})\} \\
&\equiv I_1 + I_2. \qquad (4.25)
\end{aligned}
$$

Using Talyor Series, I_1 of (4.25) can be written as

$$
\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n} X_i[K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})] \\
&= \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n} X_i\{\omega(\frac{t-F_n(X_i)}{h})\frac{F_{n,k}(X_i)-F_n(X_i)}{h} \\
&-\frac{1}{2}\omega'(\frac{t-\xi_{n,k,i}}{h})(\frac{F_{n,k}(X_i)-F_n(X_i)}{h})^2\} \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i\{\omega(\frac{t-F_n(X_i)}{h})\sum_{k=1}^{n}\frac{F_{n,k}(X_i)-F_n(X_i)}{h} \\
&-\frac{1}{2}\sum_{k=1}^{n}\omega'(\frac{t-\xi_{n,k,i}}{h})(\frac{F_{n,k}(X_i)-F_n(X_i)}{h})^2\} \\
&= -\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} X_i\omega'(\frac{t-\xi_{n,k,i}}{h})(\frac{F_{n,k}(X_i)-F_n(X_i)}{h})^2, \qquad (4.26)
\end{aligned}
$$

where $\xi_{n,k,i}$ is a random variable between $F_n(X_i)$ and $F_{n,k}(X_i)$. Since

$$
\begin{aligned}
& F_n(X) - F_{n,k}(X) \\
&= \frac{1}{n}\sum_{i=1}^{n} I(X_i \le X) - \frac{1}{n-1}\sum_{j\neq k} I(X_i \le X) \\
&= \frac{1}{n}\sum_{i=1}^{n} I(X_i \le X) - \frac{1}{n-1}\sum_{i=1}^{n} I(X_i \le X) + \frac{1}{n-1} I(X_k \le X) \\
&= -\frac{1}{n-1}\frac{1}{n}\sum_{i=1}^{n} I(X_i \le X) + \frac{1}{n-1} I(X_k \le X) \\
&= \frac{1}{n-1}\{I(X_k \le X) - F_n(X)\} \\
&= O_p(\frac{1}{n-1}) = O_p(\frac{1}{n}).
\end{aligned} \tag{4.27}
$$

Under conditions of Theorem 4.1, we have

$$
\begin{aligned}
& \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} |X_i\omega'(\frac{t-\xi_{n,k,i}}{h})|(\frac{F_{n,k}(X_i) - F_n(X_i)}{h})^2 \\
&= \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} |X_i\omega'(\frac{t-\xi_{n,k,i}}{h})| O_p(\frac{1}{(n-1)^2h^2}) \\
&= \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} |X_i\omega'(\frac{t-\xi_{n,k,i}}{h})| O_p(\frac{1}{(n-1)^2h^2}) \\
&= \frac{1}{2} O_p(n) O_p(\frac{1}{(n-1)^2h^2}) \\
&= O_p(\frac{n}{(n-1)^2h^2}) \\
&= O_p(\frac{1}{nh^2}).
\end{aligned} \tag{4.28}
$$

Meanwhile, I_2 from (4.25) can be written to

$$\sum_{k=1}^{n}\{(\frac{1}{n}-\frac{1}{(n-1)})\sum_{j=1}^{n}X_jK(\frac{t-F_{n,k}(X_j)}{h})+\frac{X_k}{(n-1)}K(\frac{t-F_{n,k}(X_k)}{h})\}$$
$$=\sum_{k=1}^{n}\{\frac{-1}{n(n-1)}\sum_{j=1}^{n}X_jK(\frac{t-F_{n,k}(X_j)}{h})+\frac{1}{n-1}X_kK(\frac{t-F_{n,k}(X_k)}{h})\}$$
$$=\frac{-1}{n-1}\sum_{k=1}^{n}\{\frac{1}{n}\sum_{j=1}^{n}X_jK(\frac{t-F_{n,k}(X_j)}{h})-X_kK(\frac{t-F_{n,k}(X_k)}{h})\}$$
$$=\frac{-1}{n-1}\sum_{k=1}^{n}\{\frac{1}{n}\sum_{j=1}^{n}X_jK(\frac{t-F_{n,k}(X_j)}{h})-\frac{1}{n}\sum_{j=1}^{n}X_jK(\frac{t-F_n(X_j)}{h})$$
$$+\frac{1}{n}\sum_{j=1}^{n}X_jK(\frac{t-F_n(X_j)}{h})\}-\frac{-1}{n-1}\sum_{k=1}^{n}\{X_kK(\frac{t-F_{n,k}(X_k)}{h})-X_kK(\frac{t-F_n(X_k)}{h})$$
$$+X_kK(\frac{t-F_n(X_k)}{h})\}$$
$$=\frac{-1}{n(n-1)}\sum_{k=1}^{n}\sum_{j=1}^{n}[X_jK(\frac{t-F_{n,k}(X_j)}{h})-X_jK(\frac{t-F_n(X_j)}{h})]$$
$$+\frac{-1}{n(n-1)}\sum_{k=1}^{n}\sum_{j=1}^{n}X_jK(\frac{t-F_n(X_j)}{h})+\frac{1}{n-1}\sum_{k=1}^{n}[X_kK(\frac{t-F_{n,k}(X_k)}{h})$$
$$-X_kK(\frac{t-F_n(X_k)}{h})]+\frac{1}{n-1}\sum_{k=1}^{n}X_kK(\frac{t-F_n(X_k)}{h})$$
$$=\frac{1}{n-1}O_p(\frac{1}{nh^2})-\frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{j=1}^{n}X_jK(\frac{t-F_n(X_j)}{h})+O_p(\frac{1}{n})$$
$$+\frac{1}{n-1}\sum_{k=1}^{n}X_kK(\frac{t-F_n(X_k)}{h})$$
$$=-\frac{1}{n-1}\sum_{j=1}^{n}X_jK(\frac{t-F_n(X_j)}{h})+\frac{1}{n-1}\sum_{k=1}^{n}X_kK(\frac{t-F_n(X_k)}{h})+O_p(\frac{1}{n})+O_p(\frac{1}{n^2h^2})$$
$$=O_p(\frac{1}{n}). \tag{4.29}$$

From (4.28) and (4.29), we get (4.23) as follows:

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n} V_k(t) \\
&= \frac{n-1}{n}\sum_{k=1}^{n}(\hat{T}_n - \hat{T}_{n-1,k}) + \hat{T}_n(t) \\
&= \frac{n-1}{n} O_p(\frac{1}{nh^2}) + \hat{T}_n(t) \\
&= \hat{T}_n(t) + O_p(\frac{1}{nh^2}).
\end{aligned} \tag{4.30}$$

Therefore

$$\begin{aligned}
&\sqrt{n}\{\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k(t) - \theta(t)\} \\
&= \sqrt{n}[\hat{T}_n(t) - \theta(t)] + O_p(\frac{1}{\sqrt{n}h^2}) \\
&\xrightarrow{d} N(0, \sigma^2(t)).
\end{aligned} \tag{4.31}$$

Thus, Lemma 1 holds.

Lemma 2. Under the conditions in Theorem 4.1, we have

$$\frac{1}{n}\sum_{k=1}^{n}\{\hat{V}_k(t)-\theta(t)\}^2 \xrightarrow{p} \sigma^2(t). \tag{4.32}$$

Proof:

We define our jackknife pseudovalue as:

$$\hat{V}_k(t) = n\hat{T}_n(t) - (n-1)\hat{T}_{n-1,k}(t). \tag{4.33}$$

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\{\hat{V}_k(t)-\theta(t)\}^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\hat{V}_k^2(t) - 2\theta(t)\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k(t) + \frac{1}{n}\sum_{k=1}^{n}\theta^2(t),
\end{aligned} \tag{4.34}$$

where

$$\begin{aligned}
&\hat{V}_k(t) \\
&= \sum_{i=1}^{n} X_i K(\frac{t-F_n(X_i)}{h}) - \sum_{j\neq k}^{n} X_j K(\frac{t-F_{n,k}(X_j)}{h}) \\
&= \sum_{i=1}^{n} X_i [K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})] + X_k K(\frac{t-F_{n,k}(X_k)}{h}).
\end{aligned} \tag{4.35}$$

So we can get that

$$\begin{aligned}
&\hat{V}_k^2(t) \\
&= \{\sum_{i=1}^{n} X_i [K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})]\}^2 \\
&+ [X_k K(\frac{t-F_{n,k}(X_k)}{h})]^2 \\
&+ 2X_k K(\frac{t-F_{n,k}(X_k)}{h}) \sum_{i=1}^{n} X_i [K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})].
\end{aligned} \tag{4.36}$$

Then

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\hat{V}_k^2(t)\\
&=\frac{1}{n}\sum_{k=1}^{n}\{\sum_{i=1}^{n}X_i[K(\frac{t-F_n(X_i)}{h})-K(\frac{t-F_{n,k}(X_i)}{h})]\}^2\\
&+\frac{1}{n}\sum_{k=1}^{n}X_k^2K^2(\frac{t-F_{n,k}(X_k)}{h})\\
&+\frac{2}{n}\sum_{k=1}^{n}X_kK(\frac{t-F_{n,k}(X_k)}{h})\sum_{i=1}^{n}X_i[K(\frac{t-F_n(X_i)}{h})-K(\frac{t-F_{n,k}(X_i)}{h})]\\
&\equiv J_1+J_2+J_3. \qquad (4.37)
\end{aligned}$$

Note that $\sqrt{n}(F_n(x)-F(x))\to B(x)$, which is a Gaussian Process. Also, based on Lemma 1, $F_{n,k}(x)-F_n(x)=O_p(\frac{1}{n})$. Therefore, based on Taylor series, J_1 of (4.37) can be written as:

$$\begin{aligned}
J_1&=\frac{1}{n}\sum_{k=1}^{n}\{\sum_{i=1}^{n}X_i[K(\frac{t-F_n(X_i)}{h})-K(\frac{t-F_{n,k}(X_i)}{h})]\}^2\\
&=\frac{1}{n}\sum_{k=1}^{n}\{\sum_{i=1}^{n}X_i\omega(\frac{t-F_n(X_i)}{h})\frac{F_{n,k}(X_i)-F_n(X_i)}{h}\\
&-\frac{1}{2}\sum_{i=1}^{n}X_i\omega'(\frac{t-\xi_{n,k,i}}{h})(\frac{F_{n,k}(X_i)-F_n(X_i)}{h})^2\}^2\\
&=\frac{1}{n}\sum_{k=1}^{n}\{\sum_{i=1}^{n}X_i\omega(\frac{t-F_n(X_i)}{h})\frac{F_{n,k}(X_i)-F_n(X_i)}{h}\}^2+O_p(\frac{1}{(nh)^2})\\
&=\frac{1}{n}\sum_{k=1}^{n}\{n\int_{-\infty}^{\infty}x\omega(\frac{t-F_n(x)}{h})\frac{F_{n,k}(x)-F_n(x)}{h}dF_n(x)\}^2+o_p(1)\\
&=\frac{1}{n}\sum_{k=1}^{n}\{\frac{n}{\sqrt{n}}\int_{-\infty}^{\infty}x\omega(\frac{t-F_n(x)}{h})\frac{F_{n,k}(x)-F_n(x)}{h}d[\sqrt{n}(F_n(x)-F(x))]\\
&+\int_{-\infty}^{\infty}nx\omega(\frac{t-F_n(x)}{h})\frac{F_{n,k}(x)-F_n(x)}{h}dF(x)\}^2+o_p(1)\\
&=\frac{1}{n}\sum_{k=1}^{n}\{\int_{-\infty}^{\infty}nx\omega(\frac{t-F_n(x)}{h})\frac{F_{n,k}(x)-F_n(x)}{h}dF(x)\}^2+o_p(1). \qquad (4.38)
\end{aligned}$$

where $\xi_{n,k,i}$ is a random variable between $F_{n,k}(X_i)$ and $F_n(X_i)$. Let $y_1=F_n(x_1), F_{n,k}(x_1)=$

$\frac{1}{n-1}\sum_{i\neq k}^{n} I(X_i \leq x_1)$. Exactly similar to Gong(2010), J_1 can be written as

$$
\begin{aligned}
&J_1 \\
&= \frac{n^2}{n}\sum_{k=1}^{n}\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2(\frac{F_{n,k}(x_1)-F_n(x_1)}{h})(\frac{F_{n,k}(x_2)-F_n(x_2)}{h}) \\
&\omega(\frac{t-F_n(x_1)}{h})\omega(\frac{t-F_n(x_2)}{h})dF_n(x_1)dF_n(x_2)\} + o_p(1) \\
&= \frac{n^2}{nh^2}\sum_{k=1}^{n}\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2[\frac{1}{n-1}\sum_{i\neq k}^{n} I(X_i \leq x_1) - \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x_1)] \\
&[\frac{1}{n-1}\sum_{i\neq k}^{n} I(X_i \leq x_2) - \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x_2)]\omega(\frac{t-F_n(x_1)}{h})\omega(\frac{t-F_n(x_2)}{h})dF_n(x_1)dF_n(x_2)\} \\
&+o_p(1) \\
&= \frac{n^2}{nh^2}\sum_{k=1}^{n}\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2[\frac{1}{n-1}\sum_{i=1}^{n} I(X_i \leq x_1) - \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x_1) - \frac{1}{n-1}I(X_k \leq x_1)] \\
&[\frac{1}{n-1}\sum_{i=1}^{n} I(X_i \leq x_2) - \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x_2) - \frac{1}{n-1}I(X_k \leq x_2)] \\
&\omega(\frac{t-F_n(x_1)}{h})\omega(\frac{t-F_n(x_2)}{h})dF_n(x_1)dF_n(x_2)\} + o_p(1) \\
&= \frac{n^2}{nh^2}\sum_{k=1}^{n}\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2[\frac{1}{n(n-1)}\sum_{i=1}^{n} I(X_i \leq x_1) - \frac{1}{n-1}I(X_k \leq x_1)] \\
&[\frac{1}{n(n-1)}\sum_{i=1}^{n} I(X_i \leq x_2) \\
&-\frac{1}{n-1}I(X_k \leq x_2)]\omega(\frac{t-F_n(x_1)}{h})\omega(\frac{t-F_n(x_2)}{h})dF_n(x_1)dF_n(x_2)\} + o_p(1) \\
&= \frac{n^2}{nh^2}\sum_{k=1}^{n}\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2\frac{1}{n-1}[\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x_1) - I(X_k \leq x_1)] \\
&\frac{1}{n-1}[\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x_2) - I(X_k \leq x_2)]\omega(\frac{t-F_n(x_1)}{h})\omega(\frac{t-F_n(x_2)}{h})dF_n(x_1)dF_n(x_2)\} \\
&+o_p(1) \\
&= \frac{n^2}{n(n-1)^2h^2}\sum_{k=1}^{n}\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2[F_n(x_1) - I(X_k \leq x_1)][F_n(x_2) - I(X_k \leq x_2)] \\
&\omega(\frac{t-F_n(x_1)}{h})\omega(\frac{t-F_n(x_x)}{h}))dF_n(x_1)dF_n(x_2)\} + o_p(1)
\end{aligned}
$$

$$= \frac{n}{(n-1)^2h^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2\sum_{k=1}^{n}[F_n(x_1)F_n(x_2) + I(X_k \le x_1)I(X_k \le x_2)$$

$$-I(X_k \le x_1)F_n(x_2) - I(X_k \le x_2)F_n(x_1)]\omega(\frac{t-F_n(x_1)}{h})\omega(\frac{t-F_n(x_2)}{h}))dF_n(x_1)dF_n(x_2)\}$$

$$+o_p(1)$$

$$= \frac{n^2}{(n-1)^2h^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2[F_n(x_1 \wedge x_2) - F_n(x_1)F_n(x_2)]$$

$$\omega(\frac{t-F_n(x_1)}{h})\omega(\frac{t-F_n(x_2)}{h}))dF_n(x_1)dF_n(x_2)\} + o_p(1)$$

$$= \frac{n^2}{(n-1)^2h^2}\int_{-1}^{1}\int_{-1}^{1} F_n^{-1}(y_1)F_n^{-1}(y_2)\{F_n[F_n^{-1}(y_1) \wedge F_n^{-1}(y_2)] - y_1y_2\}$$

$$\omega(\frac{t-y_1}{h})\omega(\frac{t-y_2}{h}))dy_1dy_2\} + o_p(1)$$

$$= \frac{n^2}{(n-1)^2h^2}\int_{-1}^{1}\int_{-1}^{1} F^{-1}(t-u_1h)F^{-1}(t-u_2h)\{F[F^{-1}(t-u_1h) \wedge F^{-1}(t-u_2h)]$$

$$-(t-u_1h)(t-u_2h)\}\omega(u_1)\omega(u_2)d(t-u_1h)d(t-u_2h) + o_p(1)$$

$$= \frac{n^2h^2}{(n-1)^2h^2}\int_{-1}^{1}\int_{-1}^{1} F^{-1}(t)F^{-1}(t)\{t \wedge t - t^2\}\omega(u_1)\omega(u_2)du_1du_2 + o_p(1)$$

$$= \frac{n^2}{(n-1)^2}\xi_t^2[t \wedge t - t^2]\int_{-1}^{1}\omega(u_1)du_1\int_{-1}^{1}\omega(u_2)du_2 + o_p(1)$$

$$\xrightarrow{p} \xi_t^2 t(1-t). \tag{4.39}$$

Based on (4.18), $EX^2K^2(\frac{t-F(X)}{h}) \longrightarrow \int_0^{\xi_t} x^2dF(x)$, and we have

$$\frac{1}{n}\sum_{k=1}^{n} K^2(\frac{t-F(X_k)}{h}) \xrightarrow{p} t.$$

By Taylor series, J_2 of (4.37) is

$$
\begin{aligned}
&J_2\\
&= \frac{1}{n}\sum_{k=1}^{n} X_k^2 K^2(\frac{t-F_{n,k}(X_k)}{h})\\
&= \frac{1}{n}\sum_{k=1}^{n} X_k^2 K^2(\frac{t-F_{n,k}(X_k)}{h}) - \frac{1}{n}\sum_{k=1}^{n} X_k^2 K^2(\frac{t-F(X_k)}{h}) + \frac{1}{n}\sum_{k=1}^{n} X_k^2 K^2(\frac{t-F(X_k)}{h})\\
&= \frac{1}{n}\sum_{k=1}^{n} X_k^2 [K^2(\frac{t-F_{n,k}(X_k)}{h}) - K^2(\frac{t-F(X_k)}{h})] + \frac{1}{n}\sum_{k=1}^{n} X_k^2 K^2(\frac{t-F(X_k)}{h})\\
&= \frac{1}{n}\sum_{k=1}^{n} X_k^2 \{2K(\frac{t-F(X_k)}{h})\omega(\frac{t-F(X_k)}{h})\frac{F_{n,k}(X_k)-F(X_k)}{h}\\
&+\frac{1}{2}[2\omega^2(\frac{t-F(X_k)}{h})(\frac{F_{n,k}(X_k)-F(X_k)}{h})^2\\
&+2K(\frac{t-F(X_k)}{h})\omega'(\frac{t-\xi_{n,k,i})}{h})(\frac{F_{n,k}(X_k)-F(X_k)}{h})^2]\} + \frac{1}{n}\sum_{k=1}^{n} X_k^2 K^2(\frac{t-F(X_k)}{h})\\
&= \frac{1}{n}\sum_{k=1}^{n} X_k^2 K^2(\frac{t-F(X_k)}{h}) + o_p(1)\\
&= EX^2K^2(\frac{t-F(X)}{h}) + o_p(1)\\
&= \int_0^{\xi_t} x^2 dF(x). \qquad (4.40)
\end{aligned}
$$

Based on Lemma 1, we have $\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n} X_i[K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})] = O_p(\frac{n}{(n-1)^2h^2})$. Under conditions of Theorem 4.1, we already proved that

$$EX^2K^2(\frac{t-F(x)}{h}) = \int_{-\infty}^{\infty} x^2K^2(\frac{t-F(x)}{h})]dF(x) \longrightarrow \int_0^{\xi_t} x^2 dF(x), as\ h \to 0. \qquad (4.41)$$

Thus, by Chebyshev's Inequality, we have $P(|X_kK(\frac{t-F(X_k)}{h})| \geq M) = P(|XK(\frac{t-F(X)}{h})| \geq M) \leq \frac{1}{M^2}E[X^2K^2(\frac{t-F(X}{h})] = O(\frac{\int_0^{\xi_t} x^2dF(x)}{M^2}) = O(\frac{1}{M^2}) \to 0, as\ M \to \infty$, so $X_kK(\frac{t-F_{n,k}(X_k)}{h}) = O_p(1)$ uniformly for $k = 1, 2, ..., n$.

Then based on (4.28), J_3 from (4.37) can be written as

$$
\begin{aligned}
& J_3 \\
&= \frac{2}{n}\sum_{k=1}^{n} X_k K(\frac{t-F_{n,k}(X_k)}{h}) \sum_{i=1}^{n} X_i[K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})] \\
&\leq \frac{2}{n}\sum_{k=1}^{n} X_k K(\frac{t-F_{n,k}(X_k)}{h}) \sum_{i=1}^{n} |X_i[K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})]| \\
&= O_p(1)\frac{2}{n}\sum_{k=1}^{n}\sum_{i=1}^{n} |X_i[K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})]| \\
&= O_p(1)\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n} |X_i[K(\frac{t-F_n(X_i)}{h}) - K(\frac{t-F_{n,k}(X_i)}{h})]| \\
&= O_p(1)O_p(\frac{1}{nh^2}) \\
&= O_p(\frac{1}{nh^2}).
\end{aligned} \tag{4.42}
$$

Based on (4.39), (4.40) and (4.42), we have

$$
\frac{1}{n}\sum_{k=1}^{n} \hat{V}_k^2(t) \xrightarrow{p} \xi_t^2 t(1-t) + \int_0^{\xi_t} x^2 dF(x). \tag{4.43}
$$

In sum,

$$
\begin{aligned}
& \frac{1}{n}\sum_{k=1}^{n} \{\hat{V}_k(t) - \theta(t)\}^2 \\
& \xrightarrow{p} \xi_t^2 t(1-t) + \int_0^{\xi_t} x^2 dF(x) - \theta^2(t) \\
& = \sigma^2(t).
\end{aligned} \tag{4.44}
$$

Thus, Lemma 2 is proved.

Proof of Theorem 4.2. It follows immediately from Lemma 1 and Lemma 2.

Proof of Theorem 4.3. According to Gong, Peng and Li(2010), define $g(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{V}_i(t)-\theta}{1+\lambda(\hat{V}_i(t)-\theta)}$ It is easy to check that

$$\begin{aligned}
0 = |g(\lambda)| &= \frac{1}{n}|\sum_{i=1}^{n}(\hat{V}_i(t)-\theta) - \lambda\sum_{i=1}^{n}\frac{(\hat{V}_i(t)-\theta)^2}{1+\lambda(\hat{V}_i(t)-\theta)}| \\
&\geq |\frac{\lambda}{n}\sum_{i=1}^{n}\frac{(\hat{V}_i(t)-\theta)^2}{1+\lambda(\hat{V}_i(t)-\theta)}| - |\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\theta| \\
&\geq \frac{|\lambda|S_n}{1+|\lambda|Z_n} - |\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\theta)| \qquad (4.45)
\end{aligned}$$

where $S_n = \frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\theta)^2$ and $Z_n = max_{1\leq i\leq n}|\hat{V}_i(t)-\theta|$

From Lemma 1 and Lemma 2, we have $|\lambda| = O_p\{n^{-\frac{1}{2}}\}$.

Put $\gamma_i = \lambda(\hat{V}_i(t)-\theta)$, then we have $max_{1\leq i\leq n}|\gamma_i| = O_p(1)$

$$\begin{aligned}
0 = g(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\theta)(1-\gamma_i+\frac{\gamma_i^2}{1+\gamma_i}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\theta) - S_n\lambda + \frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{V}_i(t)-\theta)\gamma_i^2}{1+\gamma_i} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\theta) - S_n\lambda + O_p(\frac{1}{n}) \qquad (4.46)
\end{aligned}$$

which implies that $\lambda = S_n^{-1}\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\theta) + \beta_n$, where $\beta_n = O_p(\frac{1}{n})$

So

$$
\begin{aligned}
& l_n(\theta(t)) \\
&= -2\ log\ L_n(\theta(t)) \\
&= 2\sum_{i=1}^{n} log\ \{1+\lambda(\hat{V}_i(t)-\theta)\} \\
&= 2\sum_{i=1}^{n}\gamma_i - \sum_{i=1}^{n}\gamma_i^2 + 2\sum_{i=1}^{n}\eta_i \\
&= 2n\lambda\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\theta) - nS_n\lambda^2 + 2\sum_{i=1}^{n}\eta_i \\
&= \frac{n\{\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\theta)\}^2}{S_n} - nS_n\beta_n^2 + 2\sum_{i=1}^{n}\eta_i \\
&= \frac{n\{\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_i(t)-\theta)\}^2}{S_n} + O_p(1) \\
&\xrightarrow{d} \chi^2(1)
\end{aligned}
\tag{4.47}
$$

Theorem 4.3 holds.

PART 5

CONCLUSION AND FUTURE WORK

Income distributions provide useful information to measure a society's economics status. Developing accurate and robust estimates of income distributions are increasingly important. In this dissertation, we studied various indexes for measuring income distributions including low income proportion, Lorenz curve and generalized Lorenz curve.

New kernel estimators for these indexes of income distributions were proposed, and proved to follow asymptotic normal distribution. We accessed the performance of the proposed estimators and showed that the kernel estimators outperform the empirical estimators, in terms of Mean Square Error (MSE) and Asymptotic Relative Efficiency (ARE). Then, confidence intervals were constructed and compared based on normal approximation, Bootstrap, and Jackknife Empirical likelihood methods. Inferences based on NA1, NA2, BT1, BT2, BT3, BT4, BCa1, BCa2, and SJEL are compared. The comparison are illustrated through extensive simulation studies and a real example. Simulation studies indicate that the proposed jackknife empirical likelihood confidence interval based on the kernel estimator outperforms all the other intervals; the Bootstrap confidence intervals based on the proposed smoothed estimators perform the second to the best. Based on this study, we recommend the use of the smoothed Jackknife Empirical Likelihood-based confidence interval (SJEL) and the smoothed bootstrap-based confidence intervals (BT3 and BT4) for low income proportion, Lorenz curve and generalized Lorenz curve.

Missing data is a common problem in many statistical applications. The conventional way is to use completed observations from the dataset and omit those observations with missing value. However, this practice may not achieve desirable statistical results by disregarding the pattern of missing data. For our future work, we will discuss the case when the response variable is missing at random (MAR). That is, missingness depends on the response

and covariate. The response is actually the income data, while the covariate can be the individual demographic characteristics. There are also other two types of missing including missing completely at random(MCAR) and missing not at random(MNAR), which maybe meaningful to investigate as well.

Empirical Likelihood (EL) method allows researchers to incorporate auxiliary information without giving parametric assumption. In recent years, there are increasing studies on imputation for missing value by EL method. Wang and Rao (2002) discussed the EL-based inference for the mean of response variable with missing value under kernel regression. Qin (2009) gave a general discussion on Empirical Likelihood for missing data. Wu (2003) used a calibration-type Empirical Likelihood method for estimating population totals and other related quantities in survey sampling, whereas Qin and Zhang (2007) considered a calibration-type Empirical Likelihood method in the context of the estimation of the mean of a response variable. Zhou, Wan and Wang (2008) utilized the EL combining with estimating equations (EE) for the kernel regression for missing data. Chen, Leung, and Qin (2003) used a two sample Empirical Likelihood method to combine the complete and incomplete observations, but their method only works under MCAR. Liu, Liu and Zhou (2011) proposed the auxiliary information with missing data and reformulate the EE through a semi-parametric procedure.

It would be meaningful to apply Empirical Likelihood (EL) based estimation method for Lorenz Curve and Generalized Lorenz Curve. EL method can effectively combine auxiliary information contained in the covariate and remove the selection bias due to missing values. Also, regression estimation method and several other estimation method can be studied and compared together with the EL estimation method. MSE and bias can be used to evaluate different point estimators. Moreover, Jackknife Empirical Likelihood method and Bootstrap method can be used to construct confidence intervals.

Meanwhile, Zheng, Zhao and Yu recently proposed a new empirical likelihood method built on the influence functions of the parameters, and they proved that the limiting distribution of the log empirical likelihood ratio follows the Wilks theorem. Based on the properties

of their proposed method, it would be very interesting to incorporate their development into our inference framework for the income distributions.

REFERENCES

[1] Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*. **2**, 244-263.

[2] Bahadur, R. (1966). A Note on Quantiles in Large Samples. *Annals of Mathematical Statistics*. **37**, 577-580.

[3] Beach, C. M. & Davidson, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economics and Statistics*. **50**, 723-735.

[4] Beach, C. M. & Richmond, J. (1985). Joint confidence intervals for income shares and Lorenz curves. *International Economic Review*. **26**, 439-450.

[5] Bezzina, E. (2012). Statistics in Focus. *Eurostat*.

[6] Bishop, J., Chakraborty, S. & Thistle, P. D. (1989). Asymptotic distribution-free statistical inference for Generalized Lorenz curves. *Review of Economics and Statistics*. **71**, 725-727.

[7] Campbell, M. K. & Torgerson, D. J. (1999). Bootstrapping: estimating confidence intervals for cost-effectiveness ratios. *QJM: An International Journal of Medicine*. Volume 92, Issue 3, 177-182

[8] Casella, G. & Berger, R. L. (2002). *Statistical Inference*. Duxbury Press.

[9] Chang, R.K.R. & Halfon, N. (1997). Graphical distribution of pediatricians in the United States: an analysis of the fifty states and Washington, DC. *Pediatrics*. **100**, 172-179.

[10] Chen, J. H. & Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*. **80**, 107-116.

[11] Chen, S. X., Leung, H. Y. & Qin, J. (2003). Information recovery in a study with surrogate endpoints. *Journal of the American Statistical Association*. **98**, 1052-1062.

[12] Claeskens, G., Jing B., Peng, L., & Zhou, W. (2003). Empirical likelihood confidence regions for comparison distributions and ROC curves. *The Canadian Journal of Statistics*. **31**, 2, 173-190.

[13] Csorgo, M., Csorgo, S. & Horvath, L. (1986). Asymptotic theory for empirical reliability and concentration process. Vol 33. Springer, New York.

[14] Cowell, Frank. A. (1998). Statistical inference for Lorenz curves with censored data. Discussion Paper. No. DARY/35. London School of Economics and Political Science.

[15] Damgaard, Christian. & Weiner, Jacob (2000). Describing inequality in plant size or fecundity. *Ecology*. **81**, 1139-1142.

[16] Dempster, A., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm *Journal of Royal Statistical Society*(Ser. B). **39**, 1-38.

[17] DiCiccio, J, T. & Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science*. Vol. 11, No. 3, 189-228.

[18] Efron, B. (1981). Nonparametric Standard Errors and Confidence intervals (with discussion). *Canadian Journal of Statsitics*. **9**, 139-172.

[19] Efron, Bradley. (1982). The Jackknife, the Boostrap, and Other Resampling Plans. CBMS 38, SIAM-NSF.

[20] Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*. Vol. 82, No. 397, 171-185.

[21] Efron, Bradley & Tibshirani, Robert (1993). An Introduction to the Bootstrap. Chapman & Hall/CRC. ISBN 9780412042317.

[22] Eurostat(2000). Low-Wage Employees in EU Countries. In "*Statistics in Focus: Population and Social Conditions*", Luxembourg: Office for Official Publications of EC.

[23] Falk, M. (1983). Relative e ciency and deciency of kernel type estimator of smooth distribution functions. *Statist. Neerlandica*. **37**, 7383.

[24] Falk, M. (1985). Asymptotic normality of the kernel quantile estimator. *Ann. Statist.*. **13**, 428433.

[25] Foster, James E. & Shorrocks, Anthony F. (1988). Poverty ordering. *Econometrica*. **56**, 173-177.

[26] Gail, M. H. & Gastwirth, J. L. (1978). A scale-free goodness-of-fit test for the exponential distribution based on the Lorenz curve. *Journal of the American Statistical Association*. **73**, 229-243.

[27] Goldie, C. M. (1971). A general definition of Lorenz curve. *Econometrica*. **39**, 1037-1039.

[28] Goldie, C. M. (1977). Convergence theorems for empirical Lorenz curves and their inverses. *Advances in Applied Probability*. **9**, 765-791.

[29] Gong, Y., Peng, L., Qi Y. C (2010). Smoothed jackknife empirical likelihood method for ROC curve *Journal of Multivariate Analysis*. **101**, 1520-1531.

[30] Hall, P. & La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *International Statist. Review*. **58**, 2, 109-127.

[31] Hallas, J. & Stovring, H. (2006). Templates for analysis of individual-level prescription data. *Basic and Clinical Pharmacology and Toxicology*. **98**, 260-265

[32] Hansen, Bruce. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*. **24**, 726-748

[33] Hart, P. E. (1975). Moment distributions in economics: an exposition. *J. Roy. Statist. Soc. Serv. A*. **138**, 423-434.

[34] Harvey, Deborah J., Gange, Alan C., Hawes, Colin J. & Rink, Markus (2011). Bionomics and distribution of the stag beetle, Lucanus cervus (L.) across Europe*. *Insect Conservation and Diversity*. **4**, 23-38.

[35] Hasegawa, H. & Kozumi, H. (2003). Estimation of Lorenz curves: A Baysian nonparametric approach. *Journal of Econometrics*. **115**, 277-291.

[36] Haukoos, J. S. & Lewis, R. J. (2005). Advanced Statistics: Bootstrapping Condence Intervals for Statistics with Difcult Distributions *Academic Emergency Medicine*. **12**, 360365.

[37] Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*. **47**, 133-140.

[38] Hsieh, F. & Turnbull, B.W. (1996a). Non-parametric and semi-parametric estimation of the receiver operating characteristic curve. *Ann. Statist*. **24**, 25 40.

[39] Hsieh, F. & Turnbull, B.W. (1996b). Nonparametric methods for evaluating diagnostic tests. *Statist. Sinica*. **6**, 47-62.

[40] Huang, X., Qin, G. S., Yuan, Y. & Zhou X. H. (2012). Confidence intervals for the difference between two partial AUCs. *Australian & New Zealand Journal of Statistics*. **54**, 63-79.

[41] Jing, B., Yuan, J., & Zhou, W. (2009). Jackknife empirical likelihood. *Journal of American Statistical Association*. **104**, 1224-1232.

[42] Kakwani, N. (1984). Welfare ranking of income distributions. *Advances in Econometrics*. **3**, 191-213.

[43] Kaplan, E. L., & Meier, P.. (1958). Nonparametric estimation from incomplete observations. *Journal of American Statistical Association*. **53**, 457-481.

[44] Kleiber, C., & Kramer, W. (2003). Efficiency, equity and generalized Lorenz dominance. *Estadistica*. **55** (Special Issue on Income Distribuiton, Inequality and Poverty, ed. C. Dagum), 173-186.

[45] Kobayashi, Y. & Takaki, H. (1992). Graphical distribution of physician in Japan. *The Lancet*. **340**, 1391-1393.

[46] Lee, W. C. (1999). Probabilistic analysis of global performance of diagnostic test: interpreting the Lorenz curve-based summary measures. *Statistics in Medicine*. **18**, 455-471.

[47] Little, R. J. A, & Rubin, D. B. (2002). Statistical analysis with missing data. New York: John Wiley.

[48] Liu, Y., Zou, C & Zhang, R. (2008). Empirical likelihood for the two-sample mean problem. *Statistics & Probability Letters*. **78**, 548-556.

[49] Liu, X., Liu, P. X, & Zhou, Y. (2011). Distribution estimation with auxiliary information for missing data. *Journal of Statistical Planning and Inference*. **141**, 711-724.

[50] Lloyd, C.J. & Yong, Z (1999). Kernel estimators for the ROC curve are better than empirical. *Statistics and Probability Letters*. **44**, 221-228..

[51] Lorenz, M. C. (1905). Methods of measuring the concentration of wealth. *J. Amer. Statist. Assoc.*. **9**, 209-219.

[52] Owen, A. (1988). Empirical likelihood ratio confidence intervals for single functional. *Biometrika*. **75**, 237-249.

[53] Owen, A. (1990). Empirical likelihood ratio confidence regions. *Annals of Statistics*. **18**, 90-120.

[54] Owen, A. (2001). *Empirical likelihood*. Chapman and Hall/CRC: New York.

[55] Preston, I. (1995). Sampling distributions of relative poverty statistics. *Applied Statistics*. **44**, 91-99; correction, 45, (1996), 399.

[56] Rongve, I. (1997). Statistical inference for poverty indices with fixed poverty lines. *Applied Economics*. **29**, 387-392.

[57] Qin, G. S. & Tsao, M. (2002). Empirical likelihood ratio confidence interval for the trimmed mean. *Communication in Statistics, Theory and Methods*. **31**, 2197-2208.

[58] Qin, G. S. & Zhou, X.H. (2006). Empirical likelihood inference for the area under the ROC curve . .

[59] Qin, J., & Zhang, B. (2007). Empirical likelihood-based inference in missing response problems and its application in observational studies. *JRSSB*. **69**, 101-122.

[60] Qin, G. S., Yang, B.Y & Qin, J. (2010). Empirical likelihood-based Inferences for the Lorenz curve *Biomerics*. **62**, 613-622.

[61] Quenouille, M. (1956). Note on bias in estimation. *Biometrika* **43**, 353-360.

[62] Sadras, Victor, & Bonglovanni, Rodolfo. (2004). Use of Lorenz curves and Gini coefficients to assess yield inequality within paddocks *Field Crops Research*. **90**, 303-310.

[63] Shao, J. (1994). L-statistics in complex survey problems. *The Annals of Statistics*. **22**, 2, 946-967.

[64] Shao, J., & Tu, D. (1995). *The Jackknife and Boostrap*. Springer.

[65] Shao, J. (2003). *Mathematical Statistics*. Springer.

[66] Shorrocks, Anthony F. (1983). Ranking income distributions. *Economica*. **50**, 3-17.

[67] Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*. **6**, 177-184.

[68] Slottje, D. J. (1989). *The Structure of Earnings and the Measurement of Income Inequality in the US*. North-Holland, Amsterdam.

[69] Smeeding, T. M., Rainwater, L., Rein, M., Hauser, R.& Schaber, G. (1990). Income poverty in seven countries: Initial estimates from the LIS database, in "*Poverty, Inequality and Income Distribution in Comparative Perspective: The Luxembourg Income Study*", Smeeding, T. M., Higgins, M. O and Rainwater, L., editors, Harvester Wheatsheaf, New York, pp 57-76.

[70] Smith, George C. JR. (1947). Lorenz curve analysis of industrial decentralization. *Journal of the American Statistical Association*. **42**, 591-596.

[71] Thistle, P.D. (1989a). Duality between generalized Lorenz curves and distribution functions. *Economic Studies Quarterly*. **40**, 183-187.

[72] Thistle, P.D. (1989b). Ranking distribution with generalized Lorenz curves. *Southern Economic Journal*. **56**, 1-12.

[73] Tukey, J. W. (1958). Bias and Confidence in Not-quite large Sample (Abstract). *Annals of Mathematical Statisitics*. **29**, 614.

[74] Wang, Q., & Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *Journal of Computational and Graphical Statistics*. , 365-385.

[75] Wood, A.T.A., Do, K.A., & Broom, N.M. (1996). Sequential linearization of empirical likelihood constraints with application to u-statistics. *Journal of Computational and Graphical Statistics*. , 365-385.

[76] Yang, B.Y., Qin, G.S., & Qin, J. (2011). Empirical likelihood-based inferences for a low income proportion. *The Canadian Joural of Statistics*. Vol.39, No.1, 1-16.

[77] Yves, G.B. , & Chris, J. S. (2003). Empirical likelihood-based inferences for a low income proportion. *Applied Statistics*. **52**, 457-468.

[78] Zheng, B. (2002). Testing Lorenz curves with non-simple random samples. *Econometrica*. **70**, 1235-1243.

[79] Zhou, X. H., Qin, G. S., Lin, H. Z. & Li, G. (2006). Inferences in censored cost regression models with empirical likelihood. *Statistica Sinica*. **16**, 1213-1232.

[80] Zhou, Y., Wan, A. T. K., & Wang, X. (2008). Estimating equation inference with missing data. *Journal of American Statistical Association*. **103**, 11871199.

Appendix A

LOW WAGE FOR EU MEMBERS IN 2006 AND 2010 BY GENDER

The 2010 data for Low Wage will be given for EU members, together with data in 2006. Meanwhile, the Low Wage by gender will also be listed for each county. The top five countries that are label with highest proportions of low-wage earners are Latvia (27.8 %), Lithuania (27.2 %), Romania (25.6 %), Poland (24.2 %), and Estonia (23.8 %), while the top five countries that own lowest proportions of low-wage earners are Sweden (2.5 %), Finland (5.9 %), France (6.1 %), Belgium (6.4 %) and Denmark (7.7 %). So there would be more low-wage earners in Latvia than in Sweden. Sweden relatively has a fairly distributed income. Low wage definition and the data source is based on Eurostat. The ":" will be used to denote missing data.

Table A.1 Low income proportion for EU members in 2006 and 2010 by gender

EU member	2006	2010	2006 M	2010 M	2006 F	2010 F
EU (27)	16.82	16.96	12.55	13.27	21.87	21.15
Euro area (17)	14.42	14.76	10.31	11.04	19.63	19.2
Belgium	7.63	6.37	5.23	3.31	11.06	10.33
Bulgaria	18.9	22.01	18.31	22.46	19.51	21.55
Czech Republic	17.05	18.18	10.9	12.93	25.08	24.53
Denmark	9.04	7.7	6.43	5.39	12.35	9.84
Germany	20.3	22.24	14.71	17.03	27.44	28.72
Estonia	23.19	23.76	14.97	15.47	29.84	30.1
Ireland	21.41	20.66	15.94	17.56	26.67	23.57
Greece	15.73	:	12.45	:	20.16	:
Spain	13.37	14.66	7.97	9.23	21.24	21.02
France	7.13	6.08	5.35	4.53	9.33	7.87
Italy	10.27	12.36	7.52	10.25	14	15.13
Cyprus	22.65	22.69	12.26	14.89	34.18	31.44
Latvia	30.9	27.81	29.46	26.66	32.05	28.67
Lithuania	29.12	27.24	27.69	24.53	30.42	29.44
Luxembourg	13.18	13.06	7.89	9.27	22.79	20.16
Hungary	21.87	19.75	22.68	18.13	21.05	21.46
Malta	14.43	18.33	13.17	15.62	16.59	22.4
Netherlands	17.74	18.13	15.47	15.31	20.36	21.17
Austria	14.19	15.02	6.84	8.19	25.32	24.76
Poland	24.72	24.16	21.8	21.79	27.98	26.78
Portugal	20.72	16.08	15.37	10.24	26.4	22.13
Romania	26.85	25.6	25.97	25.45	27.9	25.77
Slovenia	19.24	17.14	15.57	15.27	23.49	19.29
Slovakia	18.3	19.03	12.14	14.6	24.92	23.66
Finland	4.75	5.85	2.49	3.31	6.81	8.02
Sweden	1.77	2.51	1.35	1.85	2.17	3.12
United Kingdom	21.77	22.05	15.03	16.68	28.47	27.56
Iceland	11.24	9.14	6.9	5.73	14.89	11.99
Norway	6.48	7.27	4.85	6.02	8.78	8.59
Switzerland	:	11.03	:	6.14	:	16.92
Croatia	:	18.17	:	15.7	:	20.74
Yugoslav Republic of Macedonia	:	28.25	:	26.37	:	30.33
Turkey	0.24	0.19	0.24	0.18	0.26	0.21

Appendix B

BASIC STATISTICS FOR GEORGIA PUBLIC UNIVERSITY INCOME IN 2012

The basic statistics including median, mean and maximum value will be listed in Table (B.1). Out of the 5,921 selected individuals, University of Georgia (UGA) contains 18% of professors who provide full-time service around the 2012 fiscal year. Georgia State University (GSU) and Georgia Institute of Technology (GIT) include 10.9% and 10.7% professors, while Georgia Health Sciences University (GHSU) has 8.17% recorded professors. Out of the 37 selected public universities and colleges in Georgia, GSU has the largest maximum salary of $949,419.33, while GHSU has the largest median salary of $137,596.625 and the largest mean salary of $169,371.5048.

Table B.1 Basic Statistics for 2012 Annual Income by School in Georgia

Organization	N	Median	Mean	Maximum
UNIVERSITY OF GEORGIA	1066	94480.92	102028.34	315672.5
GEORGIA STATE UNIVERSITY	648	93282.12	115344.81	949419.33
GEORGIA INSTITUTE OF TECHNOLOGY	636	132044.33	145819	445395.96
GEORGIA HEALTH SCIENCES UNIVERSITY	484	137596.63	169371.5	633260.27
KENNESAW STATE UNIVERSITY	378	75005.84	83185.99	236777.9
GEORGIA SOUTHERN UNIVERSITY	322	74409.56	81143.88	186384.92
VALDOSTA STATE UNIVERSITY	203	69810	72299.71	147794.56
COLUMBUS STATE UNIVERSITY	178	74206.1	78193.53	156847.72
UNIVERSITY OF WEST GEORGIA	167	71470.4	79061.09	183013.08
GEORGIA PERIMETER COLLEGE	157	62244.2	62945.92	111588.37
GEORGIA COLLEGE&STATE UNIVERSITY	154	74188.8	78206.65	135147.06
NORTH GEORGIA COLLEGE&STATE UNIVERSITY	130	72778.9	73651.18	126528.75
ARMSTRONG ATLANTIC STATE UNIVERSITY	120	72622.87	76309.79	158262.31
GAINESVILLE STATE COLLEGE	113	58918	61418.9	108062.04
CLAYTON STATE UNIVERSITY	111	64778.04	70533.48	126973.44
GEORGIA GWINNETT COLLEGE	106	69564.5	76645.85	165067.5
AUGUSTA STATE UNIVERSITY	96	66886.41	73044.82	170517
SOUTHERN POLYTECHNIC STATE UNIVERSITY	93	77234.42	79588.04	142815.28
ALBANY STATE UNIVERSITY	82	75064	77109.71	119188.45
GORDON COLLEGE	59	64170	63739.29	90623
ABRAHAM BALDWIN AGRICULTURAL COLLEGE	56	59068.8	60070.11	83757.94
MACON STATE COLLEGE	56	59598.49	65800.94	259706.28
SAVANNAH STATE UNIVERSITY	56	71163.25	75591.09	121817.7
DALTON STATE COLLEGE	55	56195.1	62351.98	135627.9
GEORGIA SOUTHWESTERN STATE UNIVERSITY	50	80606	82649.15	130974
DARTON STATE COLLEGE	47	67625	72708.66	146362.05
BAINBRIDGE COLLEGE	45	61044	62125.71	88081.1
COLLEGE OF COASTAL GEORGIA	45	62970	66926.25	102832.7
FORT VALLEY STATE UNIVERSITY	45	66813.35	70412.93	154323.41
MIDDLE GEORGIA COLLEGE	41	55122.9	57742.84	81096.87
GEORGIA HIGHLANDS COLLEGE	37	56428	60623.52	88265.8
EAST GEORGIA STATE COLLEGE	28	54771.4	56376.73	71543.5
GEORGIA MILITARY COLLEGE	22	29912.5	34415.45	57776.6
ATLANTA METROPOLITAN STATE COLLEGE	17	62300	63694.23	82847.48
SOUTH GEORGIA COLLEGE	9	52610	51997.78	61260
WAYCROSS COLLEGE	6	55030	56137.83	70292
SKIDAWAY INSTITUTE OF OCEANOGRAPHY	3	69998.54	54874.81	72528.62