5-8-2020

Jackknife Empirical Likelihood For The Correlation Coefficient With Additive Distortion Errors

Da Chen

JACKKNIFE EMPIRICAL LIKELIHOOD FOR THE CORRELATION COEFFICIENT WITH ADDITIVE DISTORTION MEASUREMENT ERRORS

by

DA CHEN

Under the Direction of Yichuan Zhao, PhD

ABSTRACT

The calculation of correlation coefficient can be inaccurate with the existence of distortion measurement errors. Such measurement errors could act in an additive or multiplicative fashion. To study the additive model, previous research has shown residual-based estimation of correlation coefficients. The powerful tool of empirical likelihood has been used to construct the confidence interval for the correlation coefficient. However, the methods so far only perform well when sample sizes are large. With small sample size situations, the coverage of EL can be below 90%. On the basis of previous research, this article proposes new methods of interval estimation for the correlation coefficient using jackknife empirical likelihood, mean jackknife empirical likelihood and adjusted jackknife empirical likelihood. For better performance with small sample sizes, we also propose adjusted mean jackknife empirical likelihood and mean adjusted empirical likelihood. The simulation results show the best performance with mean adjusted jackknife empirical likelihood when the sample sizes are as small as 25. Real data analyses are used to illustrate the proposed approach.

INDEX WORDS: Correlation coefficient, Distortion errors, Jackknife empirical likelihood, Adjusted jackknife empirical likelihood, Mean jackknife empirical likelihood, Adjusted mean jackknife empirical likelihood, Mean adjusted jackknife empirical likelihood

JACKKNIFE EMPIRICAL LIKELIHOOD FOR THE CORRELATION COEFFICIENT WITH ADDITIVE DISTORTION MEASUREMENT ERRORS

by

DA CHEN

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

Master of Science

in the College of Arts and Sciences

Georgia State University

2020

JACKKNIFE EMPIRICAL LIKELIHOOD FOR CORRELATION COEFFICIENT WITH ADDITIVE DISTORTION MEASUREMENT ERRORS

by

DA CHEN

Committee Chair: Yichuan Zhao

Committee: Jun Kong

Jing Zhang

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2020

DEDICATION

To those who supported me for all time.

ACKNOWLEDGEMENTS

First of all, I want to demonstrate my greatest acknowledgement to my mentor for the past years. Dr. Yichuan Zhao has been very kind and tolerant to me and all the inexcusable mistakes I have made. His immeasurable knowledge of empirical likelihood as well as biostatistics taught me not only the essential skills to be researcher but also how to think as a researcher.

Next, I would like to express my appreciation to my parents and my wife. I think I am confident enough to say that they did not raise me as a failure. I am not a smart person so I have to take more time than others to truly discover where my interest really lies. I am glad I have never given up along the way and so are my parents. Also, I feel lucky to have my wife being supportive all these years. She may not understand a single word of what I have done, but she is always willing to keep our small family organized with her greatest love. The past four years have been a tough period for her as well, and I am thankful that she has not abandoned me.

The last but not least, I would like to show my gratitude to my grandfather. He is the most respected among many successful people I have known, as a mentor and a family senior. Ever since I have can remember, he has always believed in me and been supportive. He showed me his trust and encouragement when I decided to pursue higher degree. It was my grandfather who taught me to hang in there because the sun would always shine after hard storms. What I have learned from him the most is his ambition to striving for a better future for all mankind. The research projects I have been working on also seem to be meaningless, considering the gigantic knowledge base of statistics and biology. However, I believe my graduation would not be an end, but a new start of an adventurous career.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

- AJEL - Adjusted empirical likelihood
- AL - Average length
- AMJEL - Adjusted mean jackknife empirical likelihood
- CI - Conficence interval
- CP - Coverage probability
- EL - Empirical likelihood
- JEL - Jackknife empirical likelihood
- LB - Lower bound
- MAJEL - Mean adjusted jackknife empirical likelihood
- MJEL - Mean jackknife empirical likelihood
- UB - Upper bound

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Measurement errors are inevitable during research and experiments. The existence of such errors could lead to inflated bias and variation, while affecting corresponding coefficients in linear regression. When studying the correlation coefficient between two variables, a confounding variable may exist in an additive way to the variables being studied, which could either underestimate or overestimate the true correlation coefficient between the variables of interest. The model of additive distortion errors is first introduced by Senturk and Muller (2005) and can be modeled as:

$$\begin{cases} \tilde{X} = X + \psi(U), \\ \tilde{Y} = Y + \phi(U), \end{cases} \tag{1.1}$$

where $(\tilde{X}, \tilde{Y})$ are the observable variables while (X, Y) are the corresponding unobservable true values of interest. $\phi(U)$ and $\psi(U)$ are unknown functions of an observed confounding variable U with identifiability condition $E[\phi(U)] = E[\psi(U)] = 0$. Senturk and Muller (2005) also proposed a model for multiplicative errors. The multiplicative error model suggests that $\tilde{X} = \psi(U)X$ and $\tilde{Y} = \psi(U)Y$ where $E[\phi(U)] = E[\psi(U)] = 1$. Different models of $\phi(\cdot)$ and $\psi(\cdot)$ have been studied under the scenario of multiplicative distortion measurement errors. Senturk and Muller (2008) investigated linear and generalized linear models. Zhang et al. (2019) studied multiplicative regression models with distortion measurement errors. Three kinds of estimators are proposed in Zhang's research. By taking the logarithm of the response variable, a least squares estimator and a moment-based estimator can be calculated. Without the logarithmic transformation, a least product relative error estimator is proposed. As to the additive fashion of distortion measurement errors, Feng et al. (2018) studied linear regression models. In this study, a residual based least squares estimator is proposed under restricted and unrestricted conditions. Then, a hypothesis testing method is proposed by introducing a test statistic according to the normalized difference of residual sums of squares under null and alternative hypotheses. In addition, an estimation and hypothesis testing method is also studied on partial linear models by Zhang et al. (2017b) under additive fashion of distortion measurement errors.

To study the correlation coefficient between X and Y, Zhang et al. (2017a) proposed a direct plug-in estimator and a residual-based estimator. Under the method of direct plug-in estimator, estimators of $\phi(U)$ and $\psi(U)$, denoted as $\hat{\phi}(U)$ and $\hat{\psi}(U)$, are subtracted from the observed response and predictor to obtain calibrated X, Y and construct the estimation using the calibrated values. The method of residual-based estimator first obtains the residuals by:

$$\begin{cases} e_{\tilde{X}U} = \tilde{X} - E[\tilde{X}|U], \\ e_{\tilde{Y}U} = \tilde{Y} - E[\tilde{Y}|U], \end{cases} \tag{1.2}$$

and then we calculate $\rho(X, Y)$ based on the fact that $\rho(X, Y) = \rho(e_{\tilde{X}U}, e_{\tilde{Y}U})$. An empirical likelihood statistic of residual-based estimation is then used to construct the confidence interval of $\rho(X, Y)$.

First introduced by Owen (1988, 1990), empirical likelihood has shown its advantage as a non-parametric tool that does not need a distribution assumption. Huang and Zhao (2017) published an article using empirical likelihood for bivariate survival function under univariate censoring. Cheng et al. (2012) introduced empirical likelihood inference for semiparametric additive isotonic regression. The method of adjusted empirical likelihood has been shown by Chen (2008). Adjusted empirical likelihood has been applied to a variety of research fields since then. Yu and Zhao (2019) proposed empirical likelihood inference and adjusted empirical likelihood for semi-parametric transformation models with length-biased sampling. Wang et al. (2019) proposed the method of penalized empirical likelihood when dealing with the sparse Cox model. However, in situations with small sample sizes, the coverage probability and the confidence region could be inaccurate using empirical likelihood. Thus, Liang et al. (2019) proposed mean empirical likelihood by constructing a pseudo dataset through the means of observed values. It has been shown that the method of MEL satisfies Wilk's theorem.

When dealing with complicated statistics, the calculation of empirical likelihood can still be redundant. Thus, jackknife empirical likelihood was introduced by Jing et al. (2009) to simplify the application of empirical likelihood to complicated statistics. Various research articles have been published using jackknife empirical likelihood. Sang et al. (2019) proposed JEL method for estimating Gini correlations. Lin et al. (2017) published the method for the error variance using JEL in linear regression models. JEL is also introduced for the accelerated failure time model by Yu and Zhao (2019). The method of JEL can also be applied to Bayesian inference, which was published by Cheng and Zhao (2019). When measuring the spread of data using mean absolute deviation, Zhao et al. (2015) have shown the JEL inference for mean absolute deviation. To compare two correlated Gini indices, Alemdjrodo and Zhao (2019) proposed a new method to reduce the computation in jackknife empirical likelihood. According to Zhao et al. (2018), Jackknife empirical likelihood can also be used for skewness and kurtosis.

From the basis of JEL, the method of adjusted jackknife empirical likelihood was proposed by Zhao et al. (2015) and Chen and Ning (2016). AJEL preserves the property of JEL while providing better coverage probability with slightly longer confidence intervals according to Zhao et al. (2015). AJEL is also widely applied as a supplement to JEL method. Yang and Zhao (2017) applied both methods to obtain the quantile difference using smoothed non-parametric estimating equation.

In this paper, we propose applications of mean and jackknife empirical likelihood into the estimation of $\rho(e_{\tilde{X}U}, e_{\tilde{Y}U})$. The residual based estimator, $\rho(e_{\tilde{X}U}, e_{\tilde{Y}U})$, is first calculated according to Zhang et al. (2017a). The JEL method is then applied to calculate the jackknife estimator of $\rho(X, Y)$. The confidence intervals based on mean and jackknife empirical likelihood are proposed in this paper. Furthermore, the confidence interval from the adjusted jackknife empirical likelihood is constructed. Simulation studies use normal distribution, beta distribution and Weibull distribution to generate the confounding variable U. The additive error terms are linear functions of U. In the simulation, the true values of $\rho(X, Y)$ are -0.9, -0.5, 0, 0.5 and 0.9. Sample sizes vary from 25 to 100 to show the performance of proposed methods in small sample size cases. In a real data analysis, the well-known Boston House Price data set is used where the location and house price are considered as X and Y while the education level is used as the confounding variable.

The organization of the thesis is as follows. In Chapter 2, we first review the method of residual-based estimation and construction of empirical likelihood proposed by Zhang et al. (2017a). In Chapter 3, application of jackknife empirical likelihood is proposed. The applications of mean jackknife empirical likelihood and adjusted jackknife empirical likelihood are proposed in Chapter 4. In Chapter 5, we propose AMJEL and MAJEL to improve the performance for small sample size situations. In Chapter 6, a simulation study is conducted using Normal, Beta and Weibull distributions. In Chapter 7, the real data analyses using the new methods are performed for illustrative purpose. The well-known 1993 new car dataset and Boston housing price dataset are used for the analysis. In Chapter 8, we make a conclusion that jackknife empirical likelihood performs better than empirical likelihood in small sample size situations.

CHAPTER 2

RESIDUAL-BASED ESTIMATOR AND EMPIRICAL LIKELIHOOD FOR CORRELATION COEFFICIENTS WITH ADDITIVE ERRORS

In this chapter, we review the EL method for correlation coefficients with additive errors, which is developed by Zhang et al. (2017a). We use similar notations which are used in Zhang et al. (2017a). The residuals of X and Y are defined as:

$$\begin{cases} \hat{e}_{i\tilde{X}U} = \tilde{X}_i - \hat{E}_h(\tilde{X}_i|U = U_i), \\ \hat{e}_{i\tilde{Y}U} = \tilde{Y}_i - \hat{E}_h(\tilde{Y}_i|U = U_i), \end{cases} \tag{2.1}$$

where

$$\begin{aligned} \hat{E}_h(\tilde{X}|U = u) &= \frac{n^{-1}\sum_{j=1}^n K_h(U_j - u)\tilde{X}_j}{n^{-1}\sum_{j=1}^n K_h(U_j - u)}, \\ \hat{E}_h(\tilde{Y}|U = u) &= \frac{n^{-1}\sum_{j=1}^n K_h(U_j - u)\tilde{Y}_j}{n^{-1}\sum_{j=1}^n K_h(U_j - u)}, \end{aligned} \tag{2.2}$$

with kernel function $K_h(\cdot) = h^{-1}K(\cdot/h)$, where $h = \hat{\sigma}_U n^{-1/3}$ and $\hat{\sigma}_U$ is the sample standard deviation of U.

Then, the residual based estimator is calculated as:

$$\begin{aligned} \hat{\rho}_{(e_{\tilde{Y}U}, e_{\tilde{X}U})} &= \frac{\widehat{Cov}(e_{\tilde{Y}U}, e_{\tilde{X}U})}{\sqrt{\hat{\sigma}^2_{e_{\tilde{Y}U}}\hat{\sigma}^2_{e_{\tilde{X}U}}}}, \\ \widehat{Cov}(e_{\tilde{Y}U}, e_{\tilde{X}U}) &= n^{-1}\sum_{i=1}^n \hat{e}_{i\tilde{X}U}\hat{e}_{i\tilde{Y}U} - \bar{\hat{e}}_{\tilde{X}U}\bar{\hat{e}}_{\tilde{Y}U}, \\ \hat{\sigma}^2_{e_{\tilde{X}U}} &= n^{-1}\sum_{i=1}^n \hat{e}^2_{i\tilde{X}U} - [\bar{\hat{e}}_{\tilde{X}U}]^2, \\ \hat{\sigma}^2_{e_{\tilde{Y}U}} &= n^{-1}\sum_{i=1}^n \hat{e}^2_{i\tilde{Y}U} - [\bar{\hat{e}}_{\tilde{Y}U}]^2, \end{aligned} \tag{2.3}$$

where $\bar{\hat{e}}_{\tilde{X}U} = n^{-1}\sum_{i=1}^n \hat{e}_{i\tilde{X}U}$ and $\bar{\hat{e}}_{\tilde{Y}U} = n^{-1}\sum_{i=1}^n \hat{e}_{i\tilde{Y}U}$.

The empirical likelihood is constructed based on the fact that

$$E\left[\frac{1}{2}\left(\frac{e_{\tilde{X}U}}{\sqrt{\sigma^2_{e_{\tilde{X}U}}}} - \rho(X,Y)\frac{e_{\tilde{Y}U}}{\sqrt{\sigma^2_{e_{\tilde{Y}U}}}}\right)\frac{e_{\tilde{Y}U}}{\sqrt{\sigma^2_{e_{\tilde{Y}U}}}}\right] + E\left[\frac{1}{2}\left(\frac{e_{\tilde{Y}U}}{\sqrt{\sigma^2_{e_{\tilde{Y}U}}}} - \rho(X,Y)\frac{e_{\tilde{X}U}}{\sqrt{\sigma^2_{e_{\tilde{X}U}}}}\right)\frac{e_{\tilde{X}U}}{\sqrt{\sigma^2_{e_{\tilde{X}U}}}}\right] = 0. \tag{2.4}$$

The empirical log-likelihood ratio function is then defined as:

$$\hat{l}(\rho(X,Y)) = -2\sup\left\{\sum_{i=1}^{n}\log(np_i); p_i > 0; \sum_{i=1}^{n} p_i = 1; \sum_{i=1}^{n} p_i\hat{\varsigma}_{n,i}(\rho(X,Y)) = 0\right\}, \tag{2.5}$$

where

$$\hat{\varsigma}_{n,i} = \frac{1}{2}\left[\left(\frac{e_{i\tilde{X}U}}{\sqrt{\sigma^2_{e_{\tilde{X}U}}}} - \rho(X,Y)\frac{e_{i\tilde{Y}U}}{\sqrt{\sigma^2_{e_{\tilde{Y}U}}}}\right)\frac{e_{i\tilde{Y}U}}{\sqrt{\sigma^2_{e_{\tilde{Y}U}}}} + \left(\frac{e_{i\tilde{Y}U}}{\sqrt{\sigma^2_{e_{\tilde{Y}U}}}} - \rho(X,Y)\frac{e_{i\tilde{X}U}}{\sqrt{\sigma^2_{e_{\tilde{X}U}}}}\right)\frac{e_{\tilde{X}U}}{\sqrt{\sigma^2_{e_{\tilde{X}U}}}}\right].$$

Using the method of Lagrange multiplier, $\hat{l}_n(\rho(X,Y))$ can be obtained as:

$$\hat{l}_n(\rho(X,Y)) = 2\sum_{i=1}^{n}\log\{1 + \lambda\hat{\varsigma}_{n,i}(\rho(X,Y))\},$$

where λ is a solution of the following equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\varsigma}_{n,i}(\rho(X,Y))}{1 + \lambda\hat{\varsigma}_{n,i}(\rho(X,Y))} = 0.$$

The EL confidence interval proposed by Zhang et al. (2016) can be then constructed as

$$I_{\rho(X,Y)} = \{\rho(X,Y) : \hat{l}_n(\rho(X,Y)) \leq c_\kappa\},$$

where c_κ is the κ quantile of χ^2_1 distribution.

CHAPTER 3

JACKKNIFE EMPIRICAL LIKELIHOOD FOR CORRELATION COEFFICIENTS WITH ADDITIVE ERRORS

In this chapter, we develop JEL methods for correlation coefficients with additive errors. Let $\hat{\rho}_{i,(e_{\tilde{Y}U},e_{\tilde{X}U})}$ denote the residual based estimator of $\rho(X,Y)$ calculated with the i^{th} observation deleted, where $i = 1, ..., n$. Let $\hat{V}_i$ denote the jackknife pseudo-value, which is obtained by

$$\hat{V}_i = n\hat{\rho}_{(e_{\tilde{Y}U},e_{\tilde{X}U})} - (n-1)\hat{\rho}_{i,(e_{\tilde{Y}U},e_{\tilde{X}U})};\ i = 1, ..., n. \tag{3.1}$$

The jackknife estimator $\hat{\rho}_J(X,Y)$ is defined as

$$\hat{\rho}_J(X,Y) = n^{-1}\sum_{i=1}^{n}\hat{V}_i. \tag{3.2}$$

The jackknife empirical likelihood of $\rho(X,Y)$ can be then defined as

$$J(\rho(X,Y)) = \sup_{\mathbf{p}=(p_1,...,p_n)}\left(\prod_{i=1}^{n} np_i; p_i \geq 0; \sum_{i=1}^{n} p_i = 1; \sum_{i=1}^{n} p_i(\hat{V}_i - \rho(X,Y)) = 0\right). \tag{3.3}$$

The maximum of p_i occurs at

$$p_i = \frac{1}{n}\left(1 + \lambda(\hat{V}_i - \rho(X,Y))\right)^{-1},\ i = 1, ..., n,$$

where λ is the solution of the following equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{V}_i - \rho(X,Y)}{1 + \lambda(\hat{V}_i - \rho(X,Y))} = 0. \tag{3.4}$$

With λ, we can now calculate the -2log of the empirical likelihood ratio as

$$-2logJ(\rho(X,Y)) = 2\sum_{i=1}^{n} log\{1+\lambda((\hat{V}_i - \rho(X,Y))\} \tag{3.5}$$

Five conditions are needed to obtain asymptotic results.

1. The density function $f_U(u)$ of the random variable U is bounded away from 0 and satisfies the Lipschitz condition of order 1 on $\mathcal{U}$, which is a compact support set of U.
2. $\phi(\cdot)$, $\psi(\cdot)$ have three bounded and continuous derivatives. $E[\phi(U)] = 0$ and $E[\psi(U)] = 0$.
3. The kernel function $K(\cdot)$ is a univariate bounded, continuous and symmetric density function about zero.
4. $E[|X|^4] < \infty, E[|Y|^4] < \infty$.
5. As $n \to \infty$, $nh^4 \to 0$, $log^2 n/(nh^2) \to 0$.

We can derive the Wilks' theorem as follows:

Theorem 3.1. *Assume that conditions 1-5 hold. Let $\rho_0(X,Y)$ be the true value of $\rho(X,Y)$. When $n \to \infty$, we have*

$$-2\log J(\rho_0(X,Y)) \xrightarrow{\mathcal{D}} \chi_1^2.$$

Following the theorem, the JEL confidence interval for $\rho(X,Y)$ is obtained by

$$I^J_{\rho(X,Y)} = \{\rho_{(X,Y)} : -2\log R(\rho_{(X,Y)}) \le \chi^2_{1-\alpha}(1)\}, \tag{3.6}$$

where $\chi^2_{1-\alpha}(1)$ is the $1-\alpha$ quantile of $\chi^2(1)$.

CHAPTER 4

ADJUSTED AND MEAN JEL FOR CORRELATION COEFFICIENTS WITH ADDITIVE ERRORS

Simulation studies have shown that under-coverage issues still exist when the sample size is smaller than 25. Thus, we use adjusted jackknife empirical likelihood to improve the performance of JEL. In order to construct an adjusted jackknife empirical likelihood ratio for $\rho(X,Y)$, first define W_i as

$$W_i(\rho(X,Y)) = \hat{V}_i - \rho(X,Y), i = 1, ..., n, \tag{4.1}$$

and then add one more pseudo value W_{n+1} to W_i

$$W_{n+1}(\rho(X,Y)) = -\frac{a_n}{n}\sum_{i=1}^{n} W_i(\rho(X,Y)), \tag{4.2}$$

where $a_n = \max(1, \log(n)/2)$ according to Chen et al. (2008). Let $\widehat{W}_i$ denote the new vector obtained from W_i. AJEL is an adjustment to the JEL. Thus, we can calculate the AJEL estimator as follows by implementing the adjustment to the JEL estimator

$$\hat{\rho}_A(X,Y) = \hat{\rho}_J(X,Y) + \frac{1}{n+1}\sum_{i=1}^{n+1} W_i(\hat{\rho}_J(X,Y)). \tag{4.3}$$

The adjusted jackknife empirical likelihood ratio for $\rho(X,Y)$ is defined as

$$J_A(\rho(X,Y)) = \sup_{\mathbf{p}=(p_1,\ldots,p_{n+1})} \left(\prod_{i=1}^{n+1}(n+1)p_i : p_i \geq 0; \sum_{i=1}^{n+1} p_i = 1; \sum_{i=1}^{n+1} p_i \widehat{W}_i(\rho(X,Y)) = 0 \right). \tag{4.4}$$

Hence, the adjusted jackknife empirical log-likelihood ratio for $\rho(X,Y)$ is

$$l_A(\rho(X,Y)) = -\sum_{i=1}^{n+1} \log(1+\lambda_a \widehat{W}_i(\rho(X,Y))), \tag{4.5}$$

where λ_a is a solution to the following equation:

$$\frac{1}{n+1}\sum_{i=1}^{n+1} \frac{\widehat{W}_i(\rho(X,Y))}{1+\lambda_a \widehat{W}_i(\rho(X,Y))} = 0.$$

Once we calculate λ_a, the -2log of adjusted jackknife empirical likelihood ratio can be obtained by

$$-2logJ_A(\rho(X,Y)) = 2\sum_{i=1}^{n+1} log\{1+\lambda \hat{W}_i\}. \tag{4.6}$$

The Wilks' theorem also holds for the adjusted jackknife empirical likelihood and it states as follows:

Theorem 4.1. *Suppose that $\rho_0(X,Y)$ is the true value of $\rho(X,Y)$. Under the same assumptions in Theorem 3.1, when $n \to \infty$*

$$-2\log J_A(\rho_0(X,Y)) \xrightarrow{\mathcal{D}} \chi_1^2. \tag{4.7}$$

Thus, following the theorem, the $100(1-\alpha)\%$ AJEL confidence interval is defined as:

$$I^A_{\rho(X,Y)} = \{\rho_a(X,Y) : -2\log J_A(\rho_a(X,Y)) \leq \chi^2_{1-\alpha}(1)\}. \tag{4.8}$$

By using AJEL, the length of AJEL confidence interval is usually longer than JEL but, the coverage probability of AJEL is better in small sample cases. To combine the methods of mean and jackknife empirical likelihood, first we let M denote the pseudo vector calculated from $\hat{W}_i$, where

$$M = \Big\{\frac{\hat{W}_i + \hat{W}_j}{2} : 1 \leq i \leq j \leq n\Big\}. \tag{4.9}$$

Through the equation above, the original $\hat{W}_i$ is expanded into a vector of size $N =$

$n(n+1)/2$. Meanwhile, M maintains the same mean as $\hat{W}_i$. The expected value of the new M is close to 0. Similar to the adjusted jackknife estimator, the mean jackknife estimator can be defined as follows by adding an adjustment term to the jackknife estimator:

$$\hat{\rho}_M(X,Y) = \hat{\rho}_J(X,Y) + \frac{1}{N}\sum_{i=1}^{N} M_i(\hat{\rho}_J(X,Y)). \tag{4.10}$$

Now, we can construct the empirical likelihood based on the new vector M. The mean empirical likelihood ratio, denoted as $\hat{R}^M(\rho(X,Y))$, is defined as:

$$\hat{R}^M(\rho(X,Y)) = \max_{\mathbf{p}=(p_1,\dots,p_n)} \left(\prod_{i=1}^{N} Np_i; p_i \geq 0; \sum_{i=1}^{N} p_i = 1; \sum_{i=1}^{N} p_i M_i(\rho(X,Y)) = 0 \right). \tag{4.11}$$

By the properties of empirical likelihood, the log-likelihood $\hat{l}^M(\rho(X,Y))$ can then be calculated as:

$$\begin{aligned} \hat{l}^M(\rho(X,Y)) &= \frac{-2\log \hat{R}^M(\rho(X,Y))}{n+1} \\ &= \frac{2}{n+1}\sum_{i=1}^{N} \log(1+\lambda M_i(\rho(X,Y))), \end{aligned} \tag{4.12}$$

where λ is the solution of the following equation

$$N^{-1}\sum_{i=1}^{N} \frac{M_i(\rho(X,Y))}{1+\lambda M_i(\rho(X,Y))} = 0. \tag{4.13}$$

To construct the confidence interval of $\rho(X,Y)$, we obtain Wilk's theorem as follows:

Theorem 4.2. *Assuming the same condition as we did in Theorem 3.1. We have that*

$$\hat{l}^M(\rho_0(X,Y)) \xrightarrow{\mathcal{D}} \chi_1^2. \tag{4.14}$$

Then, the mean jackknife empirical likelihood confidence interval is defined as follows:

$$I^M_{\rho(X,Y)} = \{\rho(X,Y) : \hat{l}^M\,(\rho(X,Y)) \leq \chi^2_{1-\alpha}(1)\},$$

where $\chi^2_{1-\alpha}(1)$ is the $1-\alpha$ quantile of χ^2_1.

Like the jackknife empirical likelihood, there is no exact formula for the confidence interval of mean jackknife empirical likelihood. Thus, the calculation of CI for MJEL is the same as that for EL by fitting a vector of estimators into the non-parametric models and compare the $l^M(\rho(X,Y))$ to the desired quantile. In general cases, the length of confidence interval decreases with the larger sample size. Simulation studies have shown that the confidence interval of MJEL is longer than that of empirical likelihood. The results of simulation studies will be shown in Chapter 6.

CHAPTER 5

ADJUSTED MEAN AND MEAN ADJUSTED JEL FOR CORRELATION COEFFICIENTS WITH ADDITIVE ERRORS

To increase the performance in small sample situations, we combine the methods of MJEL and AJEL and propose the methods of adjusted mean jackknife empirical likelihood (AMJEL) and mean adjusted jackknife empirical likelihood (MAJEL). In AMJEL, we calculate the vector M from equation (4.8) and then add one more point to the vector. For MAJEL, we first obtain the vector $\hat{W}_i$ from equation (4.2) and then expand the vector using the equation similar to equation (4.8).

For AMJEL, M is obtained by using equation (4.8) and has $N = n(n+1)/2$ elements. We add one additional point to M using

$$M_{N+1} = -\frac{a_N}{N}\sum_{i=1}^{N} M_i, \tag{5.1}$$

where $a_N = max(1, log(N)/2)$. The adjusted mean jackknife estimator is then defined as

$$\hat{\rho}_{AM}(X,Y) = \hat{\rho}_M(X,Y) + \frac{1}{N+1}\sum_{i=1}^{N+1} M_i(\hat{\rho}_M(X,Y)). \tag{5.2}$$

The adjusted mean jackknife empirical likelihood ratio is then defined as:

$$J_{AM}(\rho(X,Y)) = sup\left(\prod_{i=1}^{N+1}(N+1)p_i; p_i \geq 0; \sum_{i=1}^{N+1} p_i = 1; \sum_{i=1}^{N+1} p_i M_i(\rho(X,Y)) = 0\right). \tag{5.3}$$

The log-likelihood of AMJEL can be then calculated by

$$l_{AM}(\rho(X,Y)) = \frac{2}{n+1}\sum_{i=1}^{N+1} log(1 + \lambda M_i(\rho(X,Y))), \tag{5.4}$$

where λ is a solution to the following equation:

$$\frac{1}{N+1}\sum_{i=1}^{N+1}\frac{M_i(\rho(X,Y))}{1+\lambda M_i(\rho(X,Y))}=0. \tag{5.5}$$

The Wilks' theorem holds for AMJEL as follows:

Theorem 5.1. *Under the same assumptions in Theorem 3.1, when $n \to \infty$,*

$$-2logJ_{AM}(\rho_0(X,Y)) \xrightarrow{\mathcal{D}} \chi_1^2. \tag{5.6}$$

Following the theorem, the $100(1-\alpha)\%$ AMJEL confidence interval is as follows

$$I_{\rho(X,Y)}^{AM} = \{\rho(X,Y) : -2logJ_{AM}(\rho(X,Y)) \leq \chi_{1-\alpha}^2(1)\}. \tag{5.7}$$

The confidence interval of AMJEL appears to be longer than the confidence interval of MJEL, which will be shown in the simulation study in Chapter 6.

For MAJEL, we use $\widehat{W}_i$ obtained from equations (4.1)-(4.2) and calculate M^A as follows:

$$M^a = \left\{\frac{\widehat{W}_i+\widehat{W}_j}{2} : 1 \leq i \leq j \leq n+1\right\}. \tag{5.8}$$

The expectation of M^a remains close to 0 and M^a has $N^a = \dfrac{(n+1)(n+2)}{2}$ values. The mean adjusted jackknife estimator is then defined as

$$\hat{\rho}_{MA}(X,Y) = \hat{\rho}_A(X,Y) + \frac{1}{N^a}\sum_{i=1}^{N^a} M^a(\hat{\rho}_A(X,Y)). \tag{5.9}$$

The mean adjusted jackknife empirical likelihood ratio is then defined as follows:

$$J_{MA}(\rho(X,Y)) = sup\left(\prod_{i=1}^{N^a} N^a p_i; p_i \geq 0; \sum_{i=1}^{N^a} p_i = 1; \sum_{i=1}^{N^a} p_i M_i^a = 0\right). \tag{5.10}$$

The log-likelihood of MAJEL can be calculated by the following equation:

$$l^{MA}(\rho(X,Y)) = \frac{2}{n+2}\sum_{i=1}^{N^a} log(1+\lambda M_i^a(\rho(X,Y))), \tag{5.11}$$

where λ is the solution of the following equation:

$$\frac{1}{N^a}\sum_{i=1}^{N^a}\frac{M_i^a(\rho(X,Y))}{1+\lambda M_i^a(\rho(X,Y))} = 0. \tag{5.12}$$

We can also obtain the Wilks' theorem for MAJEL.

Theorem 5.2. *Under the same assumption in Theorem 3.1, when $n \to \infty$,*

$$-2logJ_{MA}(\rho_0(X,Y)) \xrightarrow{\mathcal{D}} \chi_1^2. \tag{5.13}$$

From the theorem, we can construct the $100(1-\alpha)\%$ MAJEL confidence interval as follows:

$$I_{\rho(X,Y)}^{MA} = \{\rho(X,Y) : -2logJ_{MA}(\rho(X,Y)) \leq \chi_{1-\alpha}^2(1)\}. \tag{5.14}$$

MAJEL performs better than AMJEL, JEL, MJEL and AJEL when the sample size is small. The average length of the MAJEL confidence interval is longer than AJEL and MJEL. The simulation study in Chapter 6 shows the comparison of estimation, coverage probability and average length of confidence intervals among the methods.

CHAPTER 6

SIMULATION STUDIES

For the simulation study, (X, Y) is generated by multivariate normal distribution with $\mu = (2, 4)$ and $\rho(X, Y) = -0.9, -0.5, 0, 0.5, 0.9$. To ensure X and Y are generated with predefined correlation coefficient, we let the (1,1) and (2,2) of the covariance matrix to be 1 while the (1,2) and (2,1) elements equal to the predefined correlation coefficients. U is simulated with Normal, Beta and Weibull distributions. The Normal distribution of $\mu = 2$ and $\sigma = 1$ is used to generate U. We set $\psi(U) = U - 2$ and $\phi(U) = 2 - U$. In the Beta distribution, $Beta(\alpha, \beta)$, we let $\alpha = 2$ and $\beta = 8$ such that $\psi(U) = U - 0.2$ and $\phi(U) = 0.2 - U$. For the Weibull distribution, $W(\lambda, \kappa)$, we have $\lambda = 1.2$ and $\kappa = 1$. We also let $\psi(U) = U - 0.9407$ and $\phi(U) = 0.9407 - U$ to ensure $E[\psi(U)] = 0$ and $E[\phi(U)] = 0$. The observed values $(\tilde{X}, \tilde{Y})$ are set up as:

$$\begin{cases} \tilde{X} = X + \psi(U), \\ \tilde{Y} = Y + \phi(U). \end{cases}$$

Each simulation was repeated 2,000 times with the sample size $n = 25, 50, 75, 100$. For the kernel function, we choose to use the Epanechnikov kernel functions, $K(t) = 0.75(1 - t^2)^+$, as suggested by Zhang et al. (2017a). The bandwidth is chosen as suggested by Silverman (1986) such that $h = \hat{\sigma}_U n^{-1/3}$, where $\hat{\sigma}_U$ is the sample standard deviation of U. Six methods, EL, JEL, AJEL, MJEL, AMJEL and MAJEL are compared in terms of estimators, coverage probability and average lengths of 95% confidence intervals. The results are shown in the following tables.

Table (6.1) **Comparison of all methods under the Normal distribution**

		EL			JEL			AJEL			MJEL			AMJEL			MAJEL		
$\rho(X,Y)$	n	$\hat{\rho}(X,Y)$	AL	CP	$\hat{\rho}_J(X,Y)$	AL	CP	$\hat{\rho}_A(X,Y)$	AL	CP	$\hat{\rho}_M(X,Y)$	AL	CP	$\hat{\rho}_{AM}(X,Y)$	AL	CP	$\hat{\rho}_{MA}(X,Y)$	AL	CP
-0.9	25	-0.895	0.162	0.854	-0.903	0.255	0.893	-0.903	0.294	0.909	-0.903	0.292	0.913	-0.903	0.496	0.986	-0.903	0.314	**0.928**
	50	-0.898	0.112	0.910	-0.900	0.151	0.935	-0.900	0.163	0.943	-0.900	0.166	0.948	-0.900	0.168	**0.949**	-0.900	0.173	0.951
	75	-0.899	0.090	0.917	-0.900	0.113	0.938	-0.900	0.120	0.941	-0.900	0.122	0.944	-0.900	0.122	0.944	-0.900	0.125	**0.948**
	100	-0.898	0.078	0.916	-0.899	0.095	0.942	-0.899	0.099	0.947	-0.899	0.100	**0.951**	-0.899	0.100	**0.951**	-0.899	0.102	0.956
-0.5	25	-0.485	0.551	0.835	-0.496	0.935	0.895	-0.496	1.076	0.911	-0.496	1.071	0.917	-0.496	1.817	0.992	-0.496	1.150	**0.929**
	50	-0.494	0.408	0.894	-0.499	0.576	0.926	-0.499	0.622	0.934	-0.499	0.634	0.938	-0.499	0.640	0.940	-0.499	0.661	**0.94**4
	75	-0.5	0.337	0.909	-0.504	0.439	0.942	-0.504	0.464	**0.950**	-0.504	0.471	0.955	-0.504	0.473	0.956	-0.504	0.485	0.965
	100	-0.496	0.295	0.918	-0.498	0.368	0.947	-0.498	0.384	**0.951**	-0.498	0.388	0.954	-0.498	0.389	0.954	-0.498	0.398	0.960
0	25	-0.005	0.695	0.841	0.002	1.215	0.904	0.002	1.398	0.926	0.002	1.392	0.930	0.002	2.361	0.995	0.002	1.493	**0.945**
	50	-0.002	0.529	0.886	0.001	0.758	0.928	0.001	0.819	0.937	0.001	0.834	0.942	0.001	0.843	0.946	0.001	0.870	**0.951**
	75	0.003	0.444	0.918	0.004	0.586	**0.953**	0.004	0.619	0.960	0.004	0.630	0.965	0.004	0.632	0.966	0.004	0.649	0.970
	100	0.003	0.388	0.930	0.004	0.488	**0.957**	0.004	0.510	0.958	0.004	0.515	0.959	0.004	0.516	0.959	0.004	0.528	0.963
0.5	25	0.481	0.551	0.864	0.500	0.928	0.906	0.500	1.068	0.927	0.500	1.061	0.928	0.500	1.807	0.995	0.500	1.139	**0.944**
	50	0.490	0.410	0.905	0.499	0.578	0.936	0.499	0.624	0.944	0.499	0.635	**0.950**	0.499	0.641	0.953	0.499	0.661	0.959
	75	0.490	0.341	0.915	0.497	0.443	0.941	0.497	0.468	0.947	0.497	0.475	**0.952**	0.497	0.477	0.954	0.497	0.490	0.960
	100	0.496	0.295	0.924	0.500	0.367	0.946	0.500	0.383	**0.952**	0.500	0.387	0.953	0.500	0.388	0.953	0.500	0.396	0.957
0.9	25	0.884	0.177	0.856	0.904	0.280	0.894	0.904	0.322	0.912	0.904	0.320	0.911	0.904	0.546	0.986	0.904	0.344	**0.927**
	50	0.894	0.114	0.886	0.901	0.154	0.915	0.901	0.167	0.926	0.901	0.169	0.930	0.901	0.171	0.931	0.901	0.177	**0.939**
	75	0.895	0.093	0.906	0.899	0.118	0.936	0.899	0.124	0.941	0.899	0.126	0.942	0.899	0.126	0.944	0.899	0.130	**0.950**
	100	0.899	0.078	0.929	0.901	0.094	0.945	0.901	0.098	**0.950**	0.901	0.099	0.951	0.901	0.099	0.953	0.901	0.102	0.954

Table (6.2) **Comparison of all methods under the Beta distribution**

		EL			JEL			AJEL			MJEL			AMJEL			MAJEL		
$\rho(X,Y)$	n	$\hat{\rho}(X,Y)$	AL	CP	$\hat{\rho}_J(X,Y)$	AL	CP	$\hat{\rho}_A(X,Y)$	AL	CP	$\hat{\rho}_M(X,Y)$	AL	CP	$\hat{\rho}_{AM}(X,Y)$	AL	CP	$\hat{\rho}_{MA}(X,Y)$	AL	CP
-0.9	25	-0.894	0.161	0.857	-0.902	0.248	0.891	-0.902	0.285	0.905	-0.902	0.284	0.910	-0.902	0.482	0.983	-0.902	0.304	**0.924**
	50	-0.898	0.110	0.912	-0.902	0.148	0.924	-0.902	0.160	0.936	-0.902	0.163	0.941	-0.902	0.164	0.943	-0.902	0.169	**0.948**
	75	-0.899	0.089	0.925	-0.901	0.112	0.945	-0.901	0.118	0.947	-0.901	0.120	**0.950**	-0.901	0.120	**0.950**	-0.901	0.124	0.956
	100	-0.899	0.077	0.916	-0.900	0.093	0.940	-0.900	0.097	0.946	-0.900	0.098	**0.949**	-0.900	0.099	**0.949**	-0.900	0.101	0.951
-0.5	25	-0.494	0.544	0.863	-0.511	0.908	0.907	-0.511	1.044	0.923	-0.511	1.039	0.926	-0.511	1.764	0.993	-0.511	1.114	**0.939**
	50	-0.499	0.407	0.904	-0.505	0.566	0.943	-0.505	0.612	**0.950**	-0.505	0.623	0.954	-0.505	0.629	0.954	-0.505	0.649	0.960
	75	-0.498	0.338	0.913	-0.503	0.435	0.937	-0.503	0.459	0.945	-0.503	0.466	**0.950**	-0.503	0.468	0.951	-0.503	0.481	0.955
	100	-0.499	0.293	0.909	-0.501	0.360	0.940	-0.501	0.376	0.943	-0.501	0.380	0.946	-0.501	0.380	0.946	-0.501	0.389	**0.948**
0	25	-0.001	0.685	0.834	0.003	1.172	0.891	0.003	1.349	0.910	0.003	1.341	0.915	0.003	2.279	0.988	0.003	1.439	**0.931**
	50	0.002	0.528	0.888	0.000	0.745	0.929	0.000	0.805	0.938	0.000	0.819	0.946	0.000	0.828	**0.948**	0.000	0.854	0.952
	75	0.002	0.442	0.915	0.002	0.577	0.940	0.002	0.609	**0.948**	0.002	0.619	0.952	0.002	0.622	0.952	0.002	0.638	0.957
	100	-0.003	0.386	0.918	-0.003	0.480	0.946	-0.003	0.502	**0.950**	-0.003	0.507	0.953	-0.003	0.508	0.953	-0.003	0.519	0.956
0.5	25	0.493	0.544	0.857	0.507	0.908	0.900	0.507	1.044	0.919	0.507	1.039	0.925	0.507	1.764	0.990	0.507	1.115	**0.941**
	50	0.497	0.404	0.898	0.503	0.563	0.924	0.503	0.608	0.930	0.503	0.619	0.936	0.503	0.625	0.940	0.503	0.645	**0.944**
	75	0.497	0.337	0.912	0.501	0.434	0.945	0.501	0.459	**0.951**	0.501	0.466	0.955	0.501	0.468	0.957	0.501	0.480	0.961
	100	0.497	0.294	0.915	0.500	0.363	0.940	0.500	0.379	0.943	0.500	0.383	0.945	0.500	0.384	0.945	0.500	0.392	**0.949**
0.9	25	0.895	0.159	0.848	0.901	0.246	0.877	0.901	0.283	0.896	0.901	0.282	0.902	0.901	0.480	0.984	0.901	0.302	**0.923**
	50	0.899	0.110	0.885	0.902	0.146	0.915	0.902	0.158	0.922	0.902	0.161	0.928	0.902	0.162	0.928	0.902	0.167	**0.932**
	75	0.900	0.089	0.929	0.901	0.111	0.941	0.901	0.118	0.946	0.901	0.119	0.947	0.901	0.120	**0.949**	0.901	0.123	0.954
	100	0.899	0.077	0.916	0.900	0.093	0.944	0.900	0.098	0.946	0.900	0.099	**0.948**	0.900	0.099	**0.948**	0.900	0.101	0.953

Table (6.3) **Comparison of all methods under the Weibull distribution**

		EL			JEL			AJEL			MJEL			AMJEL			MAJEL		
$\rho(X,Y)$	n	$\hat{\rho}(X,Y)$	AL	CP	$\hat{\rho}_J(X,Y)$	AL	CP	$\hat{\rho}_A(X,Y)$	AL	CP	$\hat{\rho}_M(X,Y)$	AL	CP	$\hat{\rho}_{AM}(X,Y)$	AL	CP	$\hat{\rho}_{MA}(X,Y)$	AL	CP
-0.9	25	-0.895	0.160	0.865	-0.902	0.233	0.882	-0.902	0.268	0.903	-0.902	0.266	0.909	-0.902	0.453	0.980	-0.902	0.286	**0.923**
	50	-0.898	0.112	0.914	-0.900	0.142	0.923	-0.900	0.154	0.930	-0.900	0.157	0.936	-0.900	0.158	0.938	-0.900	0.163	**0.943**
	75	-0.899	0.091	0.923	-0.900	0.110	0.937	-0.900	0.116	0.943	-0.900	0.118	0.946	-0.900	0.119	**0.947**	-0.900	0.122	0.953
	100	-0.900	0.077	0.927	-0.901	0.091	0.938	-0.901	0.095	0.941	-0.901	0.096	0.944	-0.901	0.096	0.946	-0.901	0.098	**0.951**
0.5	25	-0.493	0.550	0.870	-0.503	0.862	0.906	-0.503	0.991	0.923	-0.503	0.987	0.927	-0.503	1.673	0.992	-0.503	1.059	**0.943**
	50	-0.497	0.410	0.912	-0.500	0.545	0.933	-0.500	0.588	0.942	-0.500	0.599	0.947	-0.500	0.605	**0.948**	-0.500	0.625	0.953
	75	-0.499	0.339	0.926	-0.501	0.421	0.937	-0.501	0.445	0.944	-0.501	0.451	0.949	-0.501	0.453	**0.950**	-0.501	0.465	0.956
	100	-0.497	0.297	0.915	-0.500	0.356	0.944	-0.500	0.371	0.946	-0.500	0.375	**0.949**	-0.500	0.376	**0.949**	-0.500	0.384	0.953
0	25	-0.006	0.691	0.868	0.000	1.110	0.908	0.000	1.277	0.925	0.000	1.270	0.931	0.000	2.159	0.995	0.000	1.362	**0.943**
	50	-0.001	0.532	0.918	0.003	0.718	0.943	0.003	0.775	**0.956**	0.003	0.789	0.963	0.003	0.797	0.963	0.003	0.823	0.969
	75	0.001	0.444	0.926	0.004	0.559	0.945	0.004	0.590	**0.951**	0.004	0.599	0.955	0.004	0.602	0.956	0.004	0.618	0.962
	100	-0.001	0.389	0.928	0.001	0.469	**0.949**	0.001	0.490	0.953	0.001	0.495	0.955	0.001	0.496	0.955	0.001	0.507	0.959
0.5	25	0.483	0.556	0.869	0.510	0.878	0.903	0.510	1.011	0.919	0.510	1.007	0.923	0.510	1.705	0.991	0.510	1.080	**0.939**
	50	0.490	0.412	0.913	0.500	0.547	0.941	0.500	0.591	0.946	0.500	0.601	**0.949**	0.500	0.607	0.951	0.500	0.626	0.956
	75	0.496	0.340	0.917	0.502	0.422	0.933	0.502	0.446	0.938	0.502	0.452	0.944	0.502	0.454	0.945	0.502	0.466	**0.949**
	100	0.497	0.296	0.935	0.502	0.354	0.944	0.502	0.370	**0.950**	0.502	0.373	0.952	0.502	0.374	0.953	0.502	0.382	0.958
0.9	25	0.884	0.176	0.878	0.906	0.259	0.891	0.906	0.298	0.908	0.906	0.296	0.912	0.906	0.505	0.989	0.906	0.318	**0.925**
	50	0.892	0.117	0.923	0.900	0.148	0.939	0.900	0.160	0.943	0.900	0.163	0.948	0.900	0.165	**0.950**	0.900	0.170	0.954
	75	0.895	0.093	0.919	0.900	0.112	0.947	0.900	0.119	**0.950**	0.900	0.120	0.954	0.900	0.121	0.954	0.900	0.124	0.960
	100	0.897	0.079	0.936	0.901	0.093	0.945	0.901	0.097	0.947	0.901	0.098	**0.951**	0.901	0.098	**0.951**	0.901	0.100	0.953

Figure 6.1 shows trends of average length and coverage probability from the simulation under Weibull distribution. The trends are similar across all scenarios with different $\rho(X, Y)$, but actual values vary. If $\rho(X, Y)$ is closer to 0, average length tends to be longer.

Conclusions from the simulations are as follows:

1) The average lengths of all five jackknife methods are longer than those of EL. The average length of the AJEL confidence interval is longer than that of JEL but the length of MJEL and AJEL are close. The AL of the MJEL confidence interval is shorter than that of AJEL when the sample size is as small as 25. When the sample size is greater than 50, MJEL has longer length than AJEL.

2) All new methods give better performance with an increase in the sample size.

3) MJEL and AJEL have similar performances. AMJEL and MAJEL show overall improvement from MJEL and AJEL.

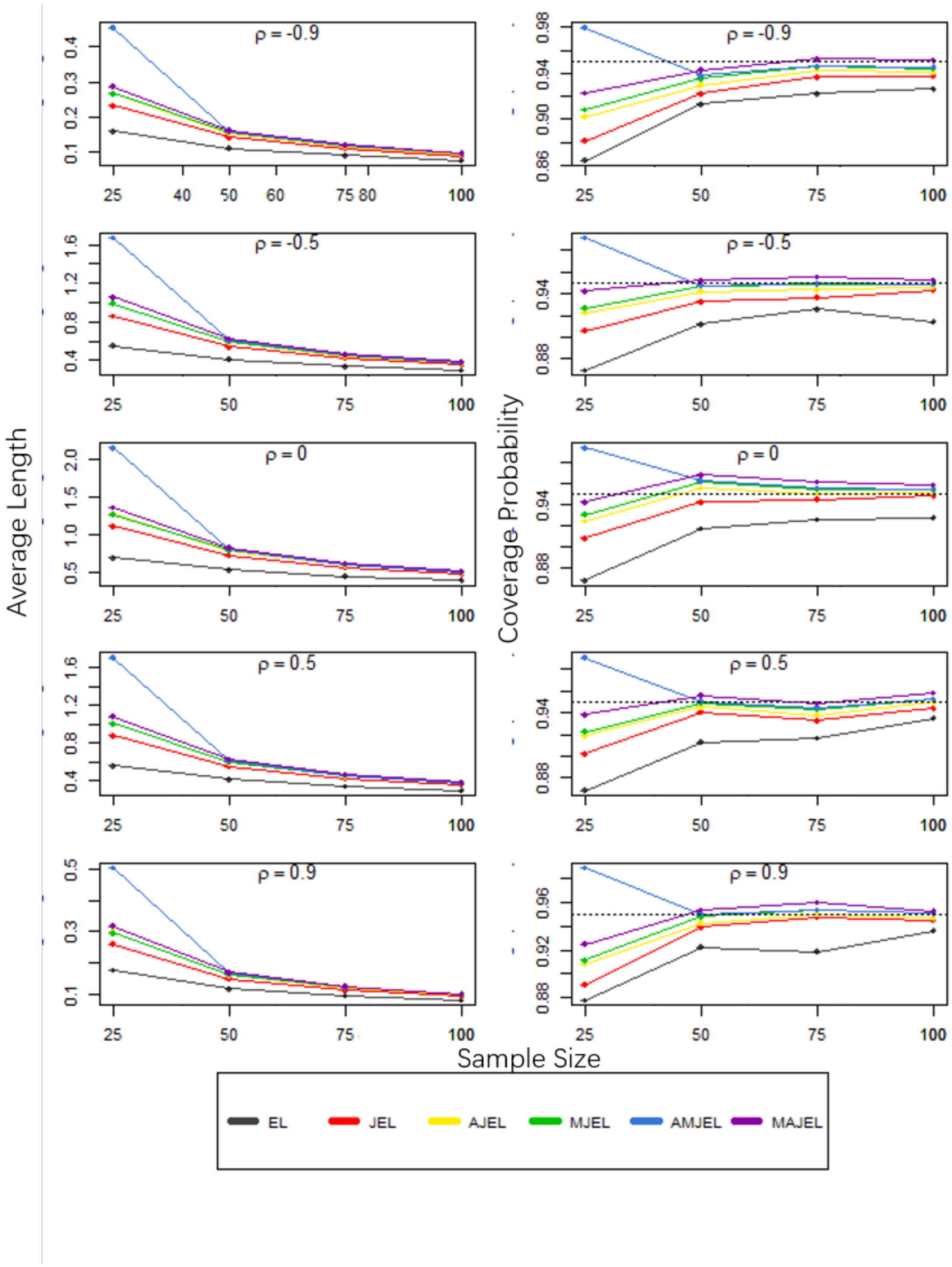
4) The performance of estimators varies under the same sample size. However, with larger sample sizes, JEL gives better estimators than the EL method.

5) The estimator, coverage probability and average length are consistent under the normal, Beta and Weibull distributions. The methods are consistent regardless of symmetric or asymmetric distributions.

6) The AMJEL method has over-coverage when sample size is 25. However, when the sample size increases, the coverage probability of the AMJEL confidence interval is close to the nominal level 0.95.

7) MAJEL outperforms all the other methods in the small sample size ($n = 25$) situations. When the sample sizes are larger than 50, MAJEL has over-coverage issues, which is similar to what we have observed in the MJEL method.

Figure (6.1) Average length and coverage probability trend plot.



CHAPTER 7

REAL DATA ANALYSIS

To compare the new methods to the original EL method proposed by Zhang et al. (2017a), we use the 1993 new car data and Boston house price data to conduct the real data analysis. The 1993 new car data is collected by Lock (1993). The data has 93 observations with 27 variables. We choose the horsepower as $\tilde{X}$ and the highway MPG as $\tilde{Y}$. Weights of cars are considered as the confounding distortion errors U. The Boston house data is retrieved from Harrison and Rubinfield (1978). The data contains 506 observations with 14 variables. We study the correlation coefficient between the house prices ($medv$) and distance to employment centers (dis), where dis is considered as $\tilde{X}$ and $medv$ is considered as $\tilde{Y}$. The lower status of the population ($lstat$) is considered as the confounding variable U. In the real data analysis, to calculate the bandwidth h for the kernel function $K_h(\cdot) = h^{-1}K(\cdot/h)$, we let $h = \hat{\sigma}_U n^{-1/3}$ where $\hat{\sigma}_U$ is the sample standard deviation of confounding variable. For the first part of the real data analysis, we compare the EL, JEL, AJEL, MJEL, AMJEL and MAJEL methods using the whole dataset.

Table (7.1) **1993 new car data analysis**

Method	Estimator	Lower	Upper	Length
EL	-0.1083	-0.3451	-0.0599	0.4050
JEL	-0.1220	-0.4518	0.1450	0.5968
AJEL	-0.1220	-0.4671	0.1574	0.6245
MJEL	-0.1220	-0.5145	0.1803	0.6948
AMJEL	-0.1220	-0.5154	0.1810	0.6964
MAJEL	-0.1220	-0.5320	0.1481	0.6801

Figure 7.1 concludes the analysis result for the new car dataset. The vertical lines show the confidence intervals for all methods including EL. The blue line shows the change of confidence interval length. The results are consistent with simulation studies that the 95%

Table (7.2) **Comparison of Boston house price analysis**

Method	Estimator	Lower	Upper	Length
EL	-0.2522	-0.3420	-0.1549	0.1701
JEL	-0.2478	-0.3524	-0.1495	0.2029
AJEL	-0.2478	-0.3536	-0.1484	0.2052
MJEL	-0.2478	-0.3531	-0.1492	0.2039
AMJEL	-0.2478	-0.3532	-0.1492	0.2040
MAJEL	-0.2478	-0.3543	-0.1481	0.2062

confidence intervals of new methods are longer than those of EL. The AMJEL confidence interval is longer than MJEL confidence interval and MAJEL confidence interval is longer than AJEL confidence interval. The naive correlation coefficient between horsepower and highway MPG from the 1993 cars data, $\rho(\tilde{X}, \tilde{Y})$, is -0.8107, which indicates that there exists a strong negative correlation between MPG and horsepower. After taking the confounding variable of weight into consideration, the 95% confidence intervals of all proposed methods contain zero, meaning the horsepower and MPG are uncorrelated. However, the EL confidence interval proposed by Zhang et al. (2017a) does not include zero.

For the Boston housing price analysis, the naive correlation coefficient, $\rho(\tilde{X}, \tilde{Y})$, between the distance and median price is 0.2499. However, the new methods indicate a negative correlation between these two variables. Also, all confidence intervals are exclusively less than zero, meaning the distance to employment centers and house prices are negatively correlated. The next part of the real data analysis focuses on the fact that new methods outperform EL with the small sample size. Thus, the Boston data set is partitioned into five sets depending on the lower status of the population. The partition and the results are as shown in Table 7.3. A plot is drawn in Figure 7.2 to demonstrate how all the methods perform in the partitioned Boston house price analysis.

The results show that the correlation coefficient between the distance and house price increases with the increase of the lower status of population (lstat). House price is positively

Figure (7.1) Confidence intervals for the new car dataset.

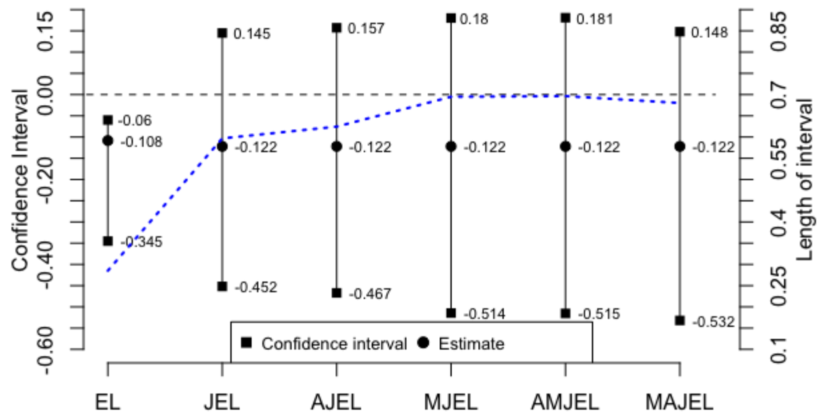
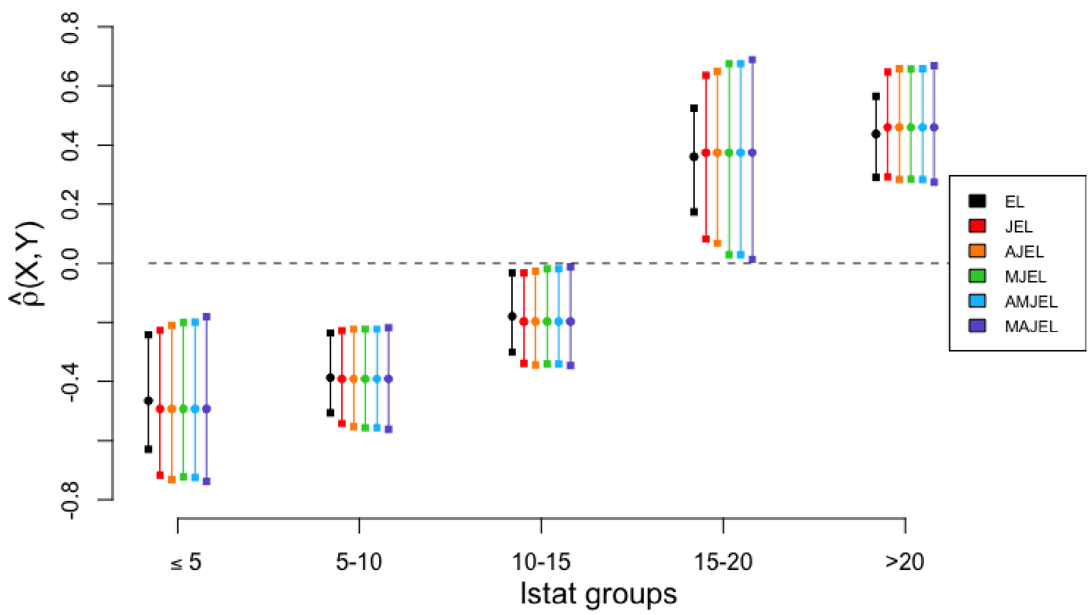


Table (7.3) **Partitioned Boston house price analysis**

lstat	(0,5]	(5,10]	(10,15]	(15,20]	(20,100]
n	62	157	125	88	74
$\hat{\rho}$(LB, UB)	-0.465 (-0.630, -0.242)	-0.387 (-0.506, -0.236)	-0.180 (-0.301, -0.033)	0.360 (0.174, 0.524)	0.437 (0.290, 0.565)
$\hat{\rho}_J$(LB, UB)	-0.493 (-0.717, -0.227)	-0.391 (-0.542, -0.228)	-0.197 (-0.339, -0.033)	0.374 (0.082, 0.636)	0.460 (0.292, 0.647)
$\hat{\rho}_A$(LB, UB)	-0.493 (-0.732, -0.210)	-0.391 (-0.553, -0.223)	-0.197 (-0.344, -0.027)	0.374 (0.068, 0.649)	0.460 (0.283, 0.658)
$\hat{\rho}_M$(LB, UB)	-0.493 (-0.723, -0.200)	-0.391 (-0.557, -0.223)	-0.197 (-0.341, -0.019)	0.374 (0.029, 0.675)	0.460 (0.284, 0.657)
$\hat{\rho}_{AM}$(LB, UB)	-0.493 (-0.725, -0.199)	-0.391 (-0.557, -0.223)	-0.197 (-0.341, -0.019)	0.374 (0.029, 0.675)	0.460 (0.283, 0.658)
$\hat{\rho}_{MA}$(LB, UB)	-0.493 (-0.738, -0.181)	-0.391 (-0.562, -0.218)	-0.197 (-0.346, -0.013)	0.374 (0.013, 0.689)	0.460 (0.274, 0.668)

correlated with the distance to employment centers when lstat is greater than 15 and the correlation is moderate. Also, when the lower status of population (lstat) is less than 5, house prices can be moderately correlated to the distance in a negative fashion.

Figure (7.2) Confidence intervals for the partitioned house price dataset.



CHAPTER 8

CONCLUSIONS

In this paper, we proposed the JEL, AJEL and MJEL for a correlation analysis when the response variable is influenced by a confounding variable and the error terms are assumed to be additions to the unobserved true values of interest. By the nature of JEL, MJEL and AJEL, all confidence intervals are larger than those of empirical likelihood by Zhang et al. (2017a). AJEL and MJEL have longer confidence intervals than JEL. AJEL provides longer confidence intervals compared to MJEL, when sample sizes are as small as 25. When the sample size, n, is between 50 and 100, the length of MJEL is larger than that of AJEL. Both MJEL and AJEL have larger length than JEL. When the true value of $\rho(X, Y) = 0$, all proposed methods generate longer confidence intervals compared with other situations. In the cases with $25 \leq n \leq 100$, all new methods provide better coverage probability compared to the conventional empirical likelihood. But the performance varies case by case. MJEL and AJEL could have over-coverage when the sample size is greater than or equal to 75. Generally, MAJEL and AMJEL show better performances when $n = 50$. The coverage probability of AMJEL could be close to 0.99 when the sample size is 25 and drops when the sample sizes become 50. The reason of unexpected overcoverage for AMJEL with the small sample sizes needs further investigation. MAJEL shows the best performance when sample sizes are 25. When applied to real data sets, the new methods make it more convenient to partition a data set into smaller subgroups without losing efficacy such that the analysis of trend is possible even when we are dealing with small data sets. For future research, the tool of jackknife empirical likelihood can be further applied to scenarios where the measurement error acts as a factor to the unobserved variables of interest.

REFERENCES

Alemdjrodo, K., Zhao, Y., Reduce the computation in jackknife empirical likelihood for comparing two correlated Gini indices, *Journal of Nonparametric Statistics*, 31(4), 849-866, 2019.

Chen, Y., Ning, W., Adjusted jackknife empirical likelihood, *arXiv:1603.04093*, 2016.

Chen, J., Variyath, A., Abraham, Bovas., Adjusted empirical likelihood and its properties, *Journal of Computational and Graphical Statistics*, 17(2), 426-443, 2008.

Cheng, G., Zhao, Y., Li, B., Empirical likelihood inferences for the semiparametric additive isotonic regression, *Journal of Multivariate Analysis*, 112, 172-182, 2012.

Cheng, Y., Zhao, Y., Bayesian jackknife empirical likelihood, *Biometrika*, 106(4), 981-988, 2019.

Feng, Z., Zhang, J., Chen, Q., Statistical inference for linear regression models with additive distortion measurement errors, *Statistical Papers*, 1-27, 2018.

Harrison, D.J., Rubinfeld, D.L., Hedonic housing prices and demand for clean air, *J Environ Econ Manage*, 5, 81-108, 1978.

Huang, H., Zhao, Y., Empirical likelihood for the bivariate survival function under univariate censoring, *Journal of Statistical Planning and Inference*, 194, 32-46, 2018.

Liang, W., Dai, H., He, S., Mean empirical likelihood, *Computational Statistics and Data Analysis*, 138(C), 155-169, 2019.

Lin, H., Li, Z., Wang, D., Zhao, Y., Jackknife empirical likelihood for the error variance in linear models, *Journal of Nonparametric Statistics*, 29, 151-166, 2017.

Lock, R., New car data, *Journal of Statistical Education*, 1(1), 1993.

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

Owen, A.B., Empirical likelihood confidence regions, *Annuals of Statistics*, 18, 90-120, 1990.

Owen, A.B., *Empirical Likelihood*, Chapman and Hall, London, 2001.

Sang, Y., Dang, X., Zhao, Y., Jackknife empirical likelihood methods for Gini correlations and their equality testing, *Journal of Statistical Planning and Inference*, 199, 45-59, 2019.

Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.

Silverman, B.W., *Monographs on Statistics and Applied Probability*, Chapman and Hall, London, 1986.

Senturk, D., Muller, H., Covariate adjusted correlation analysis via varying coefficient models, *Journal of Statistics*, 32(3), 365-383, 2005.

Senturk, D., Muller, H., Generalized varying coefficient models for longitudinal data, *Biometrika*, 95(3), 653-666, 2008.

Wang, D., Wu, T., Zhao, Y., Penalized empirical likelihood for the sparse Cox regression model, *Journal of Statistics Planning and Inference*, 201, 71-85, 2019.

Wang, D., Zhao, Y., Gilmore, D., Jackknife empirical likelihood confidence interval for the Gini index, *Statistics and Probability Letters*, 110, 289-295, 2015.

Yang, H., Zhao, Y., Smoothed jackknife empirical likelihood for the one-sample difference of quantiles, *Computational Statistics and Data Analysis*, 120, 58-69, 2017.

Yu, X., Zhao, Y., Empirical likelihood inference for semi-parametric transformation models with length-biased sampling, *Computational Statistics and Data Analysis*, 132(C), 115-125, 2019.

Yu, X., Zhao, Y., Jackknife empirical likelihood inference for the accelerated failure time model, *Test*, 28, 269–288, 2019.

Zhang, J., Chen, Q., Zhou, N., Correlation analysis with additive distortion measurement errors, *Journal of Statistical Computation and Simulation*, 87(4), 664-688, 2017a.

Zhang, J., Zhou, Y., Lin, B., Yu, Y., Estimation and hypothesis test on partial linear models with additive distortion measurement errors, *Computational Statistics and Data Analysis*, 112, 114-128, 2017b.

Zhang, J., Zhu, J., Zhou, Y., Cui, X., Lu, T., Multiplicative regression models with distortion measurement errors, *Statistical Papers*, 1-27, 2019.

Zhao, Y., Meng, X., Yang, H., Jackknife empirical likelihood inference for the mean absolute deviation, *Computational Statistics and Data Analysis*, 91, 92-101, 2015.