7-25-2013

# Jackknife Empirical Likelihood for the Variance in the Linear Regression Model

Hui-Ling Lin

# JACKKNIFE EMPIRICAL LIKELIHOOD FOR THE VARIANCE IN THE LINEAR REGRESSION MODEL

by

HUI-LING LIN

Under the Direction of Dr. Yichuan Zhao

## ABSTRACT

The variance of a random variable is $\sigma^2$. It is the measure of spread from the center. Therefore, how to accurately estimate variance ($\sigma^2$) has always been an important topic in recent years. In this paper, we consider a linear regression model which is the most popular model in practice. We use jackknife empirical likelihood (JEL) method to obtain the interval estimate of $\sigma^2$ in the regression model. The proposed JEL ratio converges to the standard chi-squared distribution. The simulation study is carried out to compare the JEL, extended JEL, adjusted JEL methods and standard method in terms of coverage probability and interval length for the confidence intervals of $\sigma^2$ from linear regression models. The proposed JEL, extended JEL and adjusted JEL has better performance than the standard method. We also illustrate the proposed methods using two real data sets.

INDEX WORDS: Variance of error, Empirical likelihood, Jackknife empirical likelihood, Adjusted jackknife empirical likelihood, Extended jackknife empirical likelihood, Coverage probability, Interval length

# JACKKNIFE EMPIRICAL LIKELIHOOD FOR THE VARIANCE IN THE LINEAR REGRESSION MODEL

by

HUI-LING LIN

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

2013

# JACKKNIFE EMPIRICAL LIKELIHOOD FOR THE VARIANCE IN THE LINEAR REGRESSION MODEL

by

HUI-LING LIN

Committee Chair: Dr. Yichuan Zhao

Committee: Dr. Xin Qi

Dr. Remus Osan

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2013

# ACKNOWLEDGEMENTS

This thesis work would not have been possible without the support of many people. First, I want to express my sincere gratitude to my adviser, Dr. Yichuan Zhao, it is only under his guidance and supports that I can complete this thesis successfully. I truly appreciate the opportunity he has given me to explore this field of statistics, which I have never dealt with before. I believe this puts a perfect ending to my life as a graduate student. I have encountered many difficulties along the way, and Dr. Zhao has always assisted me in everyday possible to help me to overcome these difficulties.

Then I would like to thank my committee members. Thanks for you taking your time to read my thesis and your valuable advice.

Finally, I would like to take this chance to thank my family and friends. For my parents, the dream of getting a degree in statistics would not have been possible without your support. For my friends Yan Zhang, and Xiaoxi Wei, thank you for all your encouraging words and support along the way.

# TABLE OF CONTENTS

# LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

Variance, according to Moore et al. (2009) is the square of standard deviation and is used to measure the variation from the mean. In statistics, variance is usually denoted as $\sigma^2$, and it has always been an important topic in the field. In the regression analysis, $\sigma^2$ measures the dispersion of $\varepsilon$, also known as the error component. It assists us in determining the variability of response value, denoted as $y$, at a specific value of $x$. (see Montgomery et al. (2006)) Over the years, statisticians have proposed many different ways to calculate $\sigma^2$ when facing different circumstances; however, it goes without saying that the process of estimating $\sigma^2$ becomes extremely difficult in a few specific situations especially when dealing with a small sample size.

Thomas and Grunkemeier (1975) was the first paper to use empirical likelihood while constructing confidence intervals for censored survival time data, and Owen (1988, 1990) looked into the relationship between empirical likelihood and nonparametric statistics and used the empirical likelihood ratio function to construct nonparametric confidence regions. According to Jing et al. (2009), "On the computational side, the empirical likelihood involves maximizing nonparametric likelihood supported on the data subject to some constraints. And if those constraints are linear, then the maximization problem becomes easy with the use of Lagrange multipliers." With the help of the empirical likelihood, constructing confidence region when the constraints are linear is no longer a difficult job. In this paper, instead of focusing solely on empirical likelihood, we shift our attention to a modified method known as the jackknife empirical likelihood (JEL) to estimate $\sigma^2$.

The main reason for choosing JEL instead of empirical likelihood is that the JEL is extremely simple to use in practice. In particular, the JEL is proven to be very effective when dealing with complicated U-statistics (Jing et al. (2009)). In this paper, we would like

to see if the JEL is a good method for calculating coverage probability and average length for the confidence interval of $\sigma^2$.

In Chapter 2, we review a few important concepts that are needed in this paper, including the concept of simple and multiple linear regression models, empirical distribution, empirical likelihood and propose jackknife empirical likelihood, adjusted jackknife empirical likelihood and extension of the jackknife empirical likelihood by expanding it domain for the $\sigma^2$ in linear regression. All the formulas entertained in the simulation study of this paper are provided as well. In Chapter 3, we focus on the simulation study, we calculate the coverage probability and average length for all confidence intervals at specific $\alpha$ levels using all three different types of JEL and compare them to the standard method. In Chapter 4, we shift our attention to two real data sets and repeat the process performed in Chapter 3. Finally in Chapter 5, we discuss the results we obtained in Chapters 3 and 4, weakness of the study, and some possible future work.

# CHAPTER 2

# METHODOLOGY

## 2.1 Linear Regression Models

Linear regression is an approach used to model the relationship between a scalar of dependent variable $\mathbf{y}$ and one or more explanatory variables denoted as $\mathbf{X}$ in matrix form, when only one explanatory variable is included then it is called a simple linear regression; other than that, it is known as a multiple linear regression (Montgomery et al. (2006)). Where $\mathbf{y}$ is a $n \times 1$ vector of $n$ scalar response variables, $\mathbf{X}$ is a design matrix of dimension $n \times (p+1)$ and $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of regression coefficients. Linear regression is known as an easy way of analyzing data and is extensively used in many practical applications. We follow the same notations as those in Montgomery et al. (2006), the general form of linear regression models can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.1}$$

Where $\boldsymbol{\varepsilon}$ is a $n \times 1$ vector of random errors. We make the assumptions that $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I})$ and $\boldsymbol{\varepsilon} \sim \mathbf{exp}\,(\mathbf{1}) - \mathbf{1}$ in our simulation study. The least-squares estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is shown below:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\,\mathbf{X}'\mathbf{y}. \tag{2.2}$$

In addition to estimating $\widehat{\boldsymbol{\beta}}$, $\sigma^2$ is another parameter that is usually estimated in order to perform hypothesis testing and construct confidence interval in the regression model (Montgomery et al. (2006)). In this paper, we focus on estimating $\sigma^2$. Denote the residual sum of squares as,

$$SS_{Res} = \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}. \tag{2.3}$$

We plug in $\widehat{\boldsymbol{\beta}}$ into the equation (2.3):

$$SS_{Res} = \mathbf{y}' \left[ \mathbf{I} - \mathbf{X} \left( \mathbf{X}'\mathbf{X} \right)^{-\mathbf{1}} \mathbf{X}' \right] \mathbf{y}. \tag{2.4}$$

As we know the residual sum of squares has $n - p$ degrees of freedom, where $p$ is $k + 1$, $k$ is the number of parameters in the model. The expected value of $SS_{Res}$ is $(n - p)\,\sigma^2$, so an unbiased estimator of $\sigma^2$ and the mean of the residual sum of square is written as

$$MS_{Res} = \frac{SS_{Res}}{n - p} = \widehat{\sigma}^2. \tag{2.5}$$

We now know $\widehat{\sigma}^2$ is calculated using the residual sum of squares. From Montgomery et al. (2006) we know that any violation of assumptions or misspecification of the error components will result an uselessness of $\widehat{\sigma}^2$. Under general assumptions, the error components are normally and independently distributed (i.i.d):

$$\frac{SS_{Res}}{\sigma^2} = \frac{(n-p)MS_{Res}}{\sigma^2} = \frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p} \tag{2.6}$$

In addition, the confidence interval estimation of $\sigma^2$ in the regression model is another important concept in our study. According to Montgomery et al. (2006) "The width of these confidence intervals is a measure of the overall quality of the regression line". Therefore, a $100\,(1 - \alpha)$ percent confidence interval on $\sigma^2$ is written as:

$$\frac{(n-p)\,MS_{Res}}{\chi^2_{\frac{\alpha}{2},n-p}} \leq \sigma^2 \leq \frac{(n-p)\,MS_{Res}}{\chi^2_{1-\frac{\alpha}{2},n-p}}, \tag{2.7}$$

where $\chi^2_{\frac{\alpha}{2},n-p}$ and $\chi^2_{1-\frac{\alpha}{2},n-p}$ are upper $\frac{\alpha}{2}$ quantile of $\chi^1_2$. The length of confidence interval is a useful tool which can help us in determining the accuracy of the regression model.

## 2.2 Empirical Likelihood

"In statistics, the empirical function is the cumulative distribution function (CDF) associated with the empirical measure of the sample" (seeOwen (2001)). Let $\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_n}$ be independent random vectors in $\mathbb{R}^p$ and for $p \geq 1$ with common distribution function $F_0$. $\delta_x$ denotes a point mass at $x$, for a more detailed description please see Owen (1990). The empirical distribution is given by

$$F_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}. \tag{2.8}$$

$F_n$ is known to be the nonparametric maximum likelihood estimate of $F_0$ based on $\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_n}$ (Owen (1990)).

As we mentioned in the previous chapter, empirical likelihood was first introduced by Thomas and Grunkemeier (1975) as a way to construct confidence intervals for survival functions with censored data. Empirical likelihood has sampling properties that are similar to the bootstrap, but it performs its goal through profiling a multinomial with one parameter for each data point to replace through resampling (Owen (1991)). The properties of empirical likelihood in i.i.d. settings are described in Owen (1988, 1990). Furthermore, Owen (1988, 1990, 1991) kept building on the idea proposed by Thomas and Grunkemeier (1975) and came up with an empirical likelihood ratio for nonparametric statistics. Owen (1990) mentioned that "Many techniques for improving the accuracy of parametric likelihood intervals apply also to empirical likelihood." For example, DiCiccio et al. (1991) show that a Bartlett correction reduces the central coverage errors to $O\left(n^{-2}\right)$. In the past, even though staatisticians could use parametric likelihood ratio functions to construct confidence intervals in some conditions, sometimes the distributions of parameters could be unknown to statisticians so it is hard to use the likelihood ratio (Owen (1988)). Therefore, Wilks (1938) proposed that under general conditions $-2logR_0$ approaches to $\chi^2_p$ distribution, where $R_0$ is the likelihood function. From Owen (1988, 1990), we know that no distribution needs to be specified when trying to draw inferences using the empirical likelihood. Based on the fact that it is not necessary to specify a distribution for the data, empirical likelihood has many

advantages over other parametric methods. The likelihood function that $F_n$ maximizes is

$$L(F) = \prod_{i=1}^{n} F\{x_i\}, \tag{2.9}$$

$F\{x_i\}$ is the probability of $\{x_i\}$ under $F$. As we mentioned in the empirical distribution, $x_i$ is the value that is observed from the $\mathbf{X_i}$ and $F$ is the probability measure on $\mathbb{R}^p$. Then from Owen (2001), $\sum_{i=1}^{n} p_i = 1$ and $L(F) = \prod_{i=1}^{n} p_i$, the empirical likelihood ratio function is given by

$$R(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} np_i. \tag{2.10}$$

We know $g_p(x) = \sum_{i=1}^{n} p_i I\{\mathbf{X_i} \leq x\}$, and the empirical liklihood evaluated at $\theta$ is defined by

$$L(\theta_p) = \max\left\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \theta(g_p) = \theta_p\right\},$$

where $\theta_p = Eg(\mathbf{X_1}, ..., \mathbf{X_n})$ is a parameter of interest. The profile empirical likelihood ratio function for $\theta$ can be rewritten as

$$R(\theta_p) = \frac{L(\theta_p)}{n^{-n}} = \max\left\{\prod_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i = 1, \theta(g_p) = \theta_p\right\}.$$

Using Lagrange multipliers, we can write

$$p_i = \frac{1}{n}\frac{1}{1 + \lambda(\mathbf{X_i} - \theta_p)}.$$

where $\lambda$ satisfies

$$f(\lambda) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{X_i} - \theta_p}{1 + \lambda(\mathbf{X_i} - \theta_p)} = 0.$$

Although empirical likelihood has many advantages for constructing confidence regions, there are still other difficulties for the empirical likelihood interval estimation of variance $\sigma^2$ in the linear regression. The reason is the estimate $\widehat{\sigma}^2$ of $sigma^2$ is a nonlinear functional.

### 2.3 Jackknife Empirical Likelihood

"The jackknife empirical likelihood (JEL) is the combined version of jackknife and empirical likelihood method. The key idea of the JEL method is to turn the statistic of interest into a sample mean based on the jackknife pseudo-values" (Jing et al. (2009)). The jackknife was invented by Quenouille (1956) as a resampling method and it is proved to be useful when the sample size $n$ is small. Miller (1974) applied the jackknife to the linear model andn Hinkley (1977) proposed an idea for the unbalanced nature of regression data points by modifying the jackknife. According to Jing et al. (2009), the simplicity is the major advantage of the JEL method and it is an easy application of empirical likelihood to the sample mean of jackknife pseudo-values. We extend JEL from the U-statistics to general case in the regression model. We let $Z_i = (x_i, y_i), i = 1, ..., n$ in the general regression model and we also let

$$T_n = T\left(Z_1, ..., Z_n\right) = \widehat{\sigma}^2 \tag{2.11}$$

be the estimation of the parameter $\sigma^2$. The jackknife pseudo-values is defined as:

$$\widehat{V}_i = nT_n - (n-1)T_{n-1}^{(-i)},\ i = 1, ..., n. \tag{2.12}$$

$T_{n-1}^{(-i)} := T(Z_1, ..., Z_{i-1}, Z_{i+1}, ..., Z_n)$. So $T_{n-1}^{(-i)}$ is computed on the sample $n-1$ variables formed from the original data set by deleting the $i$th observation. Thus the jackknife estimator of $\sigma^2$ is the average of all the pseduo-values which is

$$\widehat{T}_{n,jack} := \frac{1}{n}\sum_{i=1}^{n}\widehat{V}_i. \tag{2.13}$$

We know the empirical likelihood is an easy tool to use. We decide to apply empirical likelihood to the jackknife pseudo-values. The empirical likelihood at the $\sigma^2$ is given by

$$L(\sigma^2) = \max\left\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i\widehat{V}_i = \sigma^2, \sum_{i=1}^{n} p_i = 1\right\}. \tag{2.14}$$

Therefore, the jackknife empirical likelihood ratio at $\sigma^2$ is defined by

$$R\left(\sigma^2\right) = \frac{L(\sigma^2)}{n^{-n}} = \max\left\{\prod_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i\widehat{V}_i = \sigma^2, \sum_{i=1}^{n} p_i = 1, p_i \geq 0\right\}. \tag{2.15}$$

Using Lagrange multipliers we get

$$p_i = \frac{1}{n}\frac{1}{1+\lambda(\widehat{V}_i - \sigma^2)}, \tag{2.16}$$

where $\lambda$ satisfies

$$f(\lambda) \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{V}_i - \sigma^2}{1+\lambda(\widehat{V}_i - \sigma^2)} = 0. \tag{2.17}$$

We plug in $p_i$ into equation (2.15), then we get the jackknife empirical log-likelihood ratio

$$logR(\sigma^2) = -2logR(\sigma^2) = -2\sum_{i=1}^{n} log\left\{1+\lambda(\widehat{V}_i - \sigma^2)\right\}. \tag{2.18}$$

Using the similar argument of Jing et al. (2009), the following theorem explains how Wilk's theorem works and states how it can be used to construct a confidence region for $\sigma^2$.

**Theorem 1** Under the above conditions, $-2logR(\sigma^2)$ converges to $\chi_1^2$ in distribution.

Using **Theorem 1**, the JEL confidence interval for $\sigma^2$ is constructed as follows,

$$R_1 = \left\{\sigma^2 : -2logR\left(\sigma^2\right) \leq \chi_1^2(\alpha)\right\},$$

where $\chi_1^2(\alpha)$ is the upper $\alpha$-quantile of $\chi_1^2$.

### 2.4 Adjusted Jackknife Empirical Likelihood

In order to improve the performance of the JEL method, we applied another method known as the adjusted JEL. The idea of adjusted empirical likelihood is from Chen et al. (2008). According to Zheng and Yu (2012) the adjustment of empirical likelihood is better than the original method because it can reduce the amount of deviation. Therefore, we

applied this idea to the JEL to see if its performance is better than the JEL. For $i = 1, 2, .., n$ and $g_i^{ad}(\sigma^2) = g_i(\sigma^2) = \widehat{V}_i - \sigma^2$. The empirical likelihood at $\sigma^2$ is given by

$$L(\sigma^2) = \max\left\{\prod_{i=1}^{n+1} p_i : \sum_{i=1}^{n+1} = 1, \sum_{i=1}^{n+1} p_i g_i^{ad}(\sigma^2) = 0\right\}, \tag{2.19}$$

$g_{n+1}^{ad}(\sigma^2) = -a_n \bar{g}_n(\sigma^2)$, and $a_n$ is a constant number depending on $n$ which is $a_n = \max(1, log(n)/2)$ (see Chen et al. (2008)).

$$\bar{g}_n(\sigma^2) = \frac{1}{n}\sum_{i=1}^{n} g_i(\sigma^2) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{V}_i - \sigma^2). \tag{2.20}$$

Therefore, we define the adjusted jackknife empirical likelihood ratio at $\sigma^2$ by

$$R^{ad}(\sigma^2) = \prod_{i=1}^{n+1}\left\{(n+1)p_i^{ad}(\sigma^2)\right\}, \tag{2.21}$$

where

$$p_i^{ad}(\sigma^2) = \frac{1}{n+1}\frac{1}{1+\lambda(g_i^{ad}(\sigma^2))},\ i = 1, ..., n+1 \tag{2.22}$$

and $\lambda$ satisfies

$$f(\lambda) \equiv \sum_{i=1}^{n+1}\frac{g_i^{ad}(\sigma^2)}{1+\lambda(g_i^{ad}(\sigma^2))} = 0. \tag{2.23}$$

After we plug in the $p_i^{ad}$ into equation (2.21) and we can get the adjusted jackknife empirical log-likelihood ratio

$$logR^{ad}(\sigma^2) = -2logR^{ad}(\sigma^2) = -2\sum_{i=1}^{n+1} log\left\{1+\lambda(g_i^{ad}(\sigma^2))\right\}.$$

Combining Chen et al. (2008) and Jing et al. (2009), we know the Wilks' theorem holds as $n \to \infty$. The following theorem explains how Wilk's theorem works and states how it can be used to construct a confidence region for $\sigma^2$.

**Theorem 2** Under the above conditions, $-2logR^{ad}(\sigma^2)$ converges to $\chi_1^2$ in distribution.

Using **Theorem 2**, the adjusted JEL confidence interval for $\sigma^2$ is constructed as follows,

$$R_2 = \left\{\sigma^2 : -2logR^{ad}\left(\sigma^2\right) \leq \chi_1^2\left(\alpha\right)\right\},$$

where $\chi_1^2\left(\alpha\right)$ is the upper $\alpha$-quantile of $\chi_1^2$.

### 2.5 Extended Jackknife Empirical Likelihood

According to Tsao (2013), "the extended empirical likelihood can escape the convex hull constraint on the empirical likelihood and improve the coverage accuracy of the empirical likelihood ratio confidence region to $O(n^{-2})$ at the same time." The difference between the JEL and adjusted JEL is that we used $h_n^C(\sigma^2)$ to replace the true value, $\sigma^2$. According to Tsao (2013) we get

$$h_n^C(\sigma^2) = \widehat{T}_{n,jack} + \gamma(n, l(\sigma^2))(\sigma^2 - \widehat{T}_{n,jack}), \tag{2.24}$$

where $\gamma(n, l(\sigma^2))$ is the expansion factor given by, (see Tsao and Wu (2012)).

$$\gamma\left(n, l\left(\sigma^2\right)\right) = 1 + \frac{l\left(\sigma^2\right)}{2n}. \tag{2.25}$$

The extended jackknife empirical likelihood ratio for $\sigma^2$ in the domain is defined by

$$R(\sigma^2) = \sup\left\{\sum_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i(\widehat{V}_i - \sigma^2) = 0, \sum_{i=1}^{n} p_i = 1, p_i \geq 0\right\}. \tag{2.26}$$

we have

$$p_i = \frac{1}{n}\frac{1}{1 + \lambda\left[\widehat{V}_i - h_n^C(\sigma^2)\right]}, \tag{2.27}$$

when $\lambda$ satisfies

$$f(\lambda) \equiv \sum_{i=1}^{n} \frac{\widehat{V}_i - h_n^C(\sigma^2)}{1 + \lambda\left[\widehat{V}_i - h_n^C(\sigma^2)\right]} = 0. \tag{2.28}$$

We plug in the $p_i$ back into equation (2.27) and get extended jackknife empirical log-likelihood ratio

$$logR\left(\sigma^2\right) = -2logR(\sigma^2) = -2\sum_{i=1}^{n} log\left\{1 + \lambda\left[\widehat{V}_i - h_n^C(\sigma^2)\right]\right\}.$$

As $n \to \infty$, the Wilks' theorem holds as Tsao and Wu (2012) did. And the following theorem explains how Wilk's theorem works and states how it can be used to construct a confidence region for $\sigma^2$.

**Theorem 3** Under the above conditions, $-2logR(\sigma^2)$ converges to $\chi_1^2$ in distribution.

Using **Theorem 3**, the extended JEL confidence interval for $\sigma^2$ is constructed as follows,

$$R_3 = \left\{\sigma^2 : -2logR\left(\sigma^2\right) \leq \chi_1^2(\alpha)\right\},$$

where $\chi_1^2(\alpha)$ is the upper $\alpha$-quantile of $\chi_1^2$.

In the next chapter, we are going to apply these methods to conduct an extensive simulation study.

# CHAPTER 3

# SIMULATION STUDY

There are three sections in this chapter. In the first section, we generated the data from normal and uniform distributions. We estimated $\sigma^2$ for both simple and multiple linear regressions with six different sample sizes: 60, 90, 120, 150, 200 and 400. The number of repetitions we performed on each sample size was 5000. We then get an interval estimate of $\sigma^2$ using the JEL method. In addition, we estimated the coverage probability by checking whether $-2logR(\sigma^2)$ is less or equal to $\chi_1^2(\alpha)$ or not. We used three different significance level in this study, $\alpha = 0.01, 0.05$ and 0.1. When $\alpha = 0.05$ and $-2logR(\sigma^2) \leq 1.96^2$, the result will be counted as a 1, otherwise it will be counted as 0. Then we add up all the 1s we have, then divide it by the number of repetitions to get the coverage probability for $\alpha = 0.05$. Besides, we also estimated $\widehat{\sigma}^2$ using the standard method (2.7) and we check whether $\widehat{\sigma}^2$ is in the confidence region. If it is, then the result will be counted as 1, otherwise it will be counted as 0. The coverage probability is then calculated under the same fashion. We used both methods to investigate whether the JEL methods outperforms the standard method. Finally, we calculated the interval lengths for each $\sigma^2$ for all situations, as well as their averages. Investigating the length of each confidence interval is a very important part for this study because the method which results in shorter lengths has a higher accuracy in capturing the true $\sigma^2$.

In the second section, one of parameters and the error term were generated from an exponential distribution. The processes are very similar with previous section, the only difference is that we only considered multiple linear regression and calculated their coverage probabilities in this section. The last section is the conclusion, we compared the performance of the JEL methods and to the standard method.

### 3.1 Simple Linear Regression

For simple linear regression, we used $y = 1+2X+\varepsilon$ as our true model, where $X \sim U[0,2]$ and $\varepsilon \sim N[0,1]$.

From Table A.1, we could see that when $\alpha = 0.01$, the coverage probability of JEL methods is close to 99%. If we compared it to the standard method then we could see that both coverage probabilities are very close to each other especially when the sample size becomes large. For the average lengths, when the sample size increases, average lengths of JEL methods are shorter than ones of standard methods. The performance of the coverage probability of the JEL methods when $\alpha = 0.05$ is good even though most of coverage probabilities do not reach to our expectation. And the average lengths of the JEL methods are also shorter than ones of normal methods. Coverage probability for the both methods are close to 90% when $\alpha = 0.1$. The average lengths of JEL methods are shorter than those of standard methods too. Under general conditions, those JEL methods have similar coverage probability comparing to the standard method but shorter average lengths. Therefore we conclude that the JEL methods can result in a pretty accurate interval estimate of $\sigma^2$. Results are shown on the Appendix A.

### 3.2 Multiple Linear Regression

For the multiple case, we used $y = 1 + 3X_1 + 4X_2 + \varepsilon$ as the true model, where $X_1 \sim U[0,3]$, $X_2 \sim N[0,4]$ and $\varepsilon \sim N[0,1]$.

As we can see in Table A.2, both the coverage probability of the JEL methods and the standard method are close to 99% when $\alpha = 0.01$, the difference becomes less obvious when sample size increases. In addition, most of the average interval lengths of JEL methods are shorter than those standard methods except when sample size equals to 400. Although the standard method has shorter average interval lengths than those of JEL methods but the differences between them are less 0.01. For $\alpha = 0.05$, we compared both methods and

found out that the coverage probability of JEL methods gets closer to 95% when the sample size increases. And average interval lengths of JEL methods are shorter than the standard methods except when sample size equals to 400. In $\alpha = 0.1$, the coverage probability of both methods is close to 90% for all sample sizes. Most of the average interval lengths of JEL methods are shorter than those of standard methods when $\alpha = 0.1$. Although in some conditions, the JEL method does not have a good coverage but it is able to give shorter interval lengths than the standard method. Therefore, we should consider the JEL method as a useful way to give an interval estimate of $\sigma^2$ for multiple linear regression. Results are shown on the Appendix A.

### 3.3 Exponential Distribution

The simulation study for this section is different from the last section. In this section, we generated $X_1$ and error term from a exponential distribution. Our purpose is to investigate the difference in coverage probability between the JEL and standard methods. We then compare it to the results obtained previously using a normal distribution.

We used the multiple case for this section, we used $y = 1 + 3X_1 + 4X_2 + \varepsilon$ as the true model. Let $X_1 \sim exp\,(1)$, $X_2 \sim U\,[0, 1]$ and $\varepsilon \sim exp\,(1) - 1$.

From Table A.3, we can see that the coverage probability of the standard method did not perform well. The coverage probability decreases as the sample size increases. For example, when sample size is 400 and dealing with a 99% confidence interval, the coverage probability for standard method is 88.3%. On the other hand, when sample size is 400, the coverage probability for the JEL is 97.67%, the adjusted JEL is 97.77% and the extended JEL is 97.63%. Therefore, those JEL methods did not seem to suffer from the same issue, even though its performance did not reach our expectation. In conclusion, the JEL, adjusted JEL and extended JEL methods are quite robust even when the error distribution is very skewed. In contrast, the standard method is not robust when the error distribution is exponential distribution. Results are shown on the Appendix A.

### 3.4 Conclusion

From the results of our simulation study we know that even though the JEL methods did not outperform the standard method in some situations, they are better in most of the cases. It turns out that the adjusted JEL method has the best performance in calculating coverage probability. In the interval length, the JEL methods usually results in shorter intervals comparing to the standard method. The only case where the standard method results in shorter intervals comparing to the JEL and extended JEL is when sample size equals to 400. However, the differences are very small.

When an exponential distribution is entertained, the coverage probabilities of standard method did not perform well. The coverage probability decreases as the sample size increases. On the other hand, the JEL methods did not seem to suffer from the same issue, even though the results did not reach our expectation. From the result, we know the proposed JEL, adjusted JEL and extended JEL methods are robust when the distribution of the error is very skewed. This is the advantages of the JEL methods.

# CHAPTER 4

# REAL DATA ANALYSIS

## 4.1 Real Data

In this chapter, we applied the JEL and standard methods to two real data sets. We calculated the interval length and used three different significance level, $\alpha$ =0.01, 0.05 and 0.1. Both data set came from Montgomery et al. (2006).

The sample size for the first data set is 38, and each observation represents one type of wine. From the data description, we can know that the quality of Pinot Noir is thought to be related to the properties of clarity, aroma, flavor, oakiness and region.

For the second data set, the sample size is 62. And response variable is the dimensionless factor for the pressure drop through a bubble cap. From the data description, we know there are four variables that are likely to affect the response which are superficial fluid velocity of the gas, kinematic viscosity, mesh opening and dimensionless number relating the superficial fluid velocity of the gas to the superficial fluid velocity of the liquid.

## 4.2 Conclusion

As we mentioned before, an interval which is shorter is higher in accuracy. From Table B.1, the adjusted and extended JEL methods result in shorter interval lengths than standard methods when confidence level is 99%. Although the JEL methods did not perform well when the confidence level is 95% and 90%, the results are very close to the standard method.

According to Table B.2, the JEL methods have better performance than the standard method because we can find all of the interval lengths for the JEL methods are shorter than standard method among all three different significance levels. Therefore, we conclude that the JEL methods are useful ways to give an interval estimate of $\sigma^2$.

# CHAPTER 5

# SUMMARY AND FUTURE WORK

## 5.1 Summary

In this thesis, we develop the interval estimate of $\sigma^2$ by using JEL, adjusted JEL and extended JEL methods. And according to the simulation study, there are two major reasons that we come to conclude that the JEL method is a useful method. The first reason is that the JEL methods can estimate a coverage probability better than the standard method in most of the time when the sample size is small. In addition, all three JEL methods resulted in shorter average lengths than the standard method. For the real data examples, although the JEL methods did not perform well while calculating interval lengths, the differences are very small comparing to the standard method. And we think the reason for not being to perform well might be because sample size is too small. By applying the JEL methods to the second dataset, we obtained shorter interval lengths comparing to the standard method in all three significance levels. Therefore, we conclude that the JEL methods can provide an interval estimate of $\sigma^2$ comparing to the standard method. By using the JEL methods, estimator with high accuracy can be obtained and a large sample size is not required. Researcher can save a lot of budget and time when collecting data necessary for the study.

## 5.2 Future Work

From the simulation and real data study we know that the JEL, adjusted JEL and extended JEL methods are really useful methods because they can provide an interval estimate of $\sigma^2$. However, the extended JEL fail to meet our expectation in this study. According to Tsao (2013), the extended JEL is believed to have better performance than other methods. Therefore, for our future work, we should investigate this problem closely.

## REFERENCES

Chen, J., Variyath, A., and Abraham, B. (2008). Adjusted empirical likelihood and its properties. *J Comput Graph Stat*, 17:426–443.

DiCiccio, T. S., Hall, P., and Romano, J. (1991). Empirical likelihood is bartlett correctable. *The Annals of Statistics*, 19:1053–1061.

Hinkley, D. (1977). Jackknifing in unbalanced situations. *Technometrics*, 19:285–292.

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

Miller, R. G. (1974). An unbalanced jackknife. *The Annals of Statistics*, 2:880–891.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2006). *Introduction to Linear Regression Analysis*. John Wiley and Sons Inc.

Moore, D. S., McCabe, G. P., and Craig, B. A. (2009). *Introduction to the Practice of Statistics*. W. H. Freeman and Company.

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

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

Owen, A. B. (1991). Empirical likelihood for linear models. *Journal of the American Statistical Association*, 19(4):1725–1747.

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

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

Thomas, D. R. and Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, 70(352):865–871.

Tsao, M. (2013). Extending the empirical likelihood by domain expansion. *The Canadian Journal of Statistics*, 40(4):1–18.

Tsao, M. and Wu, F. (2012). Extended empirical likelihood for estimating equations. *Biometrika*, 99(1):1–14.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9:60–62.

Zheng, M. and Yu, W. (2012). Empirical likelihood method for multivariate cox regression. *Comput Stat*.

# Appendix A

## TABLES FOR SIMULATION STUDY

Table A.1 Coverage probability and average length of confidence intervals of $\sigma^2$ for simple linear regression model

| $1-\alpha$ | $n$ | | $JEL$ | $AJEL^a$ | $EJEL^b$ | $Standard$ |
|---|---|---|---|---|---|---|
| 0.99 | 60 | Coverage | 0.9790 | 0.9816 | 0.9806 | 0.9902 |
| | | Length | 0.9594 | 0.8962 | 0.9829 | 1.0582 |
| | 90 | Coverage | 0.9856 | 0.9866 | 0.9864 | 0.9908 |
| | | Length | 0.7690 | 0.6049 | 0.7935 | 0.8289 |
| | 120 | Coverage | 0.9862 | 0.9882 | 0.9870 | 0.9892 |
| | | Length | 0.6716 | 0.5844 | 0.6839 | 0.7051 |
| | 150 | Coverage | 0.9900 | 0.9914 | 0.9908 | 0.9906 |
| | | Length | 0.6097 | 0.5293 | 0.6088 | 0.6234 |
| | 200 | Coverage | 0.9880 | 0.9902 | 0.9898 | 0.9908 |
| | | Length | 0.5297 | 0.4649 | 0.5270 | 0.5330 |
| | 400 | Coverage | 0.9917 | 0.9918 | 0.9923 | 0.9920 |
| | | Length | 0.3746 | 0.3358 | 0.3775 | 0.3705 |
| 0.95 | 60 | Coverage | 0.9424 | 0.9420 | 0.9372 | 0.9474 |
| | | Length | 0.7262 | 0.6870 | 0.6516 | 0.7772 |
| | 90 | Coverage | 0.9426 | 0.9492 | 0.9448 | 0.9516 |
| | | Length | 0.5664 | 0.4504 | 0.4751 | 0.6162 |
| | 120 | Coverage | 0.9424 | 0.9446 | 0.9416 | 0.9462 |
| | | Length | 0.5181 | 0.4384 | 0.4420 | 0.5274 |
| | 150 | Coverage | 0.9492 | 0.9544 | 0.9516 | 0.9542 |
| | | Length | 0.4632 | 0.3920 | 0.4203 | 0.4679 |
| | 200 | Coverage | 0.9530 | 0.9482 | 0.9462 | 0.9492 |
| | | Length | 0.4002 | 0.3105 | 0.3988 | 0.4014 |
| | 400 | Coverage | 0.9497 | 0.9507 | 0.9500 | 0.9513 |
| | | Length | 0.2823 | 0.2266 | 0.2785 | 0.2804 |

[a]Adjusted jackknife empirical likelihood
[b]Extended jackknife empirical likelihood

| $1-\alpha$ | $n$ | | $JEL$ | $AJEL^a$ | $EJEL^b$ | $Standard$ |
|---|---|---|---|---|---|---|
| 0.90 | 60 | Coverage | 0.8822 | 0.8924 | 0.8852 | 0.8956 |
| | | Length | 0.5628 | 0.5730 | 0.5079 | 0.6430 |
| | 90 | Coverage | 0.8998 | 0.8992 | 0.8938 | 0.9000 |
| | | Length | 0.4887 | 0.4014 | 0.3942 | 0.5123 |
| | 120 | Coverage | 0.8884 | 0.8930 | 0.8884 | 0.8938 |
| | | Length | 0.4327 | 0.3800 | 0.3664 | 0.4395 |
| | 150 | Coverage | 0.8928 | 0.9100 | 0.9052 | 0.9098 |
| | | Length | 0.3863 | 0.3464 | 0.3517 | 0.3905 |
| | 200 | Coverage | 0.9010 | 0.8966 | 0.8930 | 0.8944 |
| | | Length | 0.3335 | 0.3037 | 0.3333 | 0.3355 |
| | 400 | Coverage | 0.8915 | 0.8923 | 0.8900 | 0.8917 |
| | | Length | 0.2355 | 0.2213 | 0.2435 | 0.2349 |

[a]Adjusted jackknife empirical likelihood
[b]Extended jackknife empirical likelihood

Table A.2 Coverage probability and average length of confidence intervals of $\sigma^2$ for multiple linear regressions model

| $1-\alpha$ | $n$ | | $JEL$ | $AJEL^a$ | $EJEL^b$ | $Standard$ |
|---|---|---|---|---|---|---|
| 0.99 | 60 | Coverage | 0.9818 | 0.9856 | 0.9840 | 0.9906 |
| | | Length | 0.9739 | 0.9056 | 0.9995 | 1.0691 |
| | 90 | Coverage | 0.9864 | 0.9877 | 0.9876 | 0.9886 |
| | | Length | 0.7770 | 0.6042 | 0.8012 | 0.8356 |
| | 120 | Coverage | 0.9882 | 0.9900 | 0.9896 | 0.9914 |
| | | Length | 0.6785 | 0.5859 | 0.6896 | 0.7082 |
| | 150 | Coverage | 0.9875 | 0.9884 | 0.9880 | 0.9910 |
| | | Length | 0.6131 | 0.5299 | 0.6132 | 0.6249 |
| | 200 | Coverage | 0.9900 | 0.9908 | 0.9906 | 0.9910 |
| | | Length | 0.5336 | 0.4697 | 0.5292 | 0.5356 |
| | 400 | Coverage | 0.9907 | 0.9907 | 0.9907 | 0.9903 |
| | | Length | 0.3744 | 0.3369 | 0.3776 | 0.3706 |
| 0.95 | 60 | Coverage | 0.9416 | 0.9478 | 0.9448 | 0.9510 |
| | | Length | 0.7348 | 0.7043 | 0.6617 | 0.7848 |
| | 90 | Coverage | 0.9472 | 0.9530 | 0.9502 | 0.9554 |
| | | Length | 0.5766 | 0.4555 | 0.4758 | 0.6211 |
| | 120 | Coverage | 0.9456 | 0.9488 | 0.9474 | 0.9472 |
| | | Length | 0.5216 | 0.4393 | 0.4458 | 0.5296 |
| | 150 | Coverage | 0.9432 | 0.9476 | 0.9446 | 0.9494 |
| | | Length | 0.4654 | 0.3945 | 0.4199 | 0.4690 |
| | 200 | Coverage | 0.9468 | 0.9502 | 0.9484 | 0.9498 |
| | | Length | 0.4025 | 0.3127 | 0.3998 | 0.4034 |
| | 400 | Coverage | 0.9500 | 0.9513 | 0.9503 | 0.9523 |
| | | Length | 0.2816 | 0.2884 | 0.2786 | 0.2806 |
| 0.90 | 60 | Coverage | 0.8918 | 0.9042 | 0.8942 | 0.9056 |
| | | Length | 0.5743 | 0.5884 | 0.5190 | 0.6491 |
| | 90 | Coverage | 0.8992 | 0.9062 | 0.9014 | 0.9092 |
| | | Length | 0.4934 | 0.4048 | 0.3959 | 0.5163 |
| | 120 | Coverage | 0.8926 | 0.9010 | 0.8950 | 0.8982 |
| | | Length | 0.4358 | 0.3858 | 0.3696 | 0.4413 |
| | 150 | Coverage | 0.8980 | 0.9034 | 0.8998 | 0.9002 |
| | | Length | 0.3883 | 0.3507 | 0.3515 | 0.3914 |
| | 200 | Coverage | 0.8968 | 0.9026 | 0.8984 | 0.8996 |
| | | Length | 0.3358 | 0.3057 | 0.3380 | 0.3371 |
| | 400 | Coverage | 0.9030 | 0.9013 | 0.9040 | 0.8973 |
| | | Length | 0.2360 | 0.2192 | 0.2443 | 0.2350 |

[a]Adjusted jackknife empirical likelihood
[b]Extended jackknife empirical likelihood

Table A.3 Coverage probability of confidence intervals of $\sigma^2$ for multiple linear regressions model

| $1-\alpha$ | $n$ | | $JEL$ | $AJEL^a$ | $EJEL^b$ | $Standard$ |
|---|---|---|---|---|---|---|
| 0.99 | 60 | Coverage | 0.9370 | 0.9452 | 0.9342 | 0.8314 |
| | 90 | Coverage | 0.9504 | 0.9544 | 0.9564 | 0.8262 |
| | 120 | Coverage | 0.9606 | 0.9636 | 0.9616 | 0.8172 |
| | 150 | Coverage | 0.9640 | 0.9668 | 0.9630 | 0.8196 |
| | 200 | Coverage | 0.9750 | 0.9772 | 0.9746 | 0.8344 |
| | 400 | Coverage | 0.9767 | 0.9777 | 0.9800 | 0.8210 |
| 0.95 | 60 | Coverage | 0.8672 | 0.8750 | 0.8704 | 0.7082 |
| | 90 | Coverage | 0.8868 | 0.8974 | 0.8888 | 0.7002 |
| | 120 | Coverage | 0.8990 | 0.9072 | 0.8972 | 0.6918 |
| | 150 | Coverage | 0.9028 | 0.9064 | 0.9012 | 0.6858 |
| | 200 | Coverage | 0.9134 | 0.9178 | 0.9126 | 0.6982 |
| | 400 | Coverage | 0.9257 | 0.9263 | 0.9283 | 0.6780 |
| 0.90 | 60 | Coverage | 0.8044 | 0.8212 | 0.8086 | 0.6220 |
| | 90 | Coverage | 0.8238 | 0.8364 | 0.8252 | 0.6098 |
| | 120 | Coverage | 0.8332 | 0.8406 | 0.9310 | 0.6036 |
| | 150 | Coverage | 0.8468 | 0.8526 | 0.8456 | 0.6086 |
| | 200 | Coverage | 0.8558 | 0.8600 | 0.8618 | 0.6102 |
| | 400 | Coverage | 0.8790 | 0.8820 | 0.8827 | 0.5983 |

[a]Adjusted jackknife empirical likelihood
[b]Extended jackknife empirical likelihood

# Appendix B

## TABLES FOR REAL DATA

Table B.1 Interval length of confidence intervals of $\sigma^2$ for wine quality data

| $1-\alpha$ | | *JEL* | | *AJEL*[a] | | *EJEL*[b] | | *Standard* | |
|---|---|---|---|---|---|---|---|---|---|
| | | UB[c] | LB[d] | UB | LB | UB | LB | UB | LB |
| | | 2.9072 | 0.6868 | 2.8249 | 0.6378 | 2.8654 | 0.7077 | 2.9901 | 0.7860 |
| 0.99 | Length | 2.2204 | | 2.1871 | | 2.1577 | | 2.2041 | |
| | | 2.4534 | 0.8311 | 2.3791 | 0.7814 | 2.4243 | 0.8480 | 2.4648 | 0.8963 |
| 0.95 | Length | 1.6223 | | 1.5975 | | 1.5763 | | 1.5685 | |
| | | 2.2458 | 0.9113 | 2.1752 | 0.8609 | 2.2226 | 0.9259 | 2.2422 | 0.9610 |
| 0.90 | Length | 1.3345 | | 1.3143 | | 1.2968 | | 1.2812 | |

[a]Adjusted jackknife empirical likelihood
[b]Extended jackknife empirical likelihood
[c]Upper bound
[d]Lower bound

Table B.2 Interval length of confidence intervals of $\sigma^2$ for pressure drop data

| $1-\alpha$ | | $JEL$ | | $AJEL^a$ | | $EJEL^b$ | | $Standard$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | UB[c] | LB[d] | UB | LB | UB | LB | UB | LB |
| | | 39.7073 | 15.9994 | 38.4892 | 17.3848 | 39.1638 | 16.1938 | 43.0935 | 16.2381 |
| 0.99 | Length | 23.7079 | | 21.1044 | | 22.9700 | | 26.8554 | |
| | | 36.3988 | 18.0021 | 34.5643 | 16.3596 | 35.5302 | 18.1334 | 37.6784 | 17.9655 |
| 0.95 | Length | 18.3967 | | 18.2047 | | 17.3968 | | 19.7129 | |
| | | 34.1478 | 19.1010 | 32.7690 | 17.5662 | 33.7697 | 19.1965 | 35.2505 | 18.9463 |
| 0.90 | Length | 15.0468 | | 15.2028 | | 14.5732 | | 16.3042 | |

[a]Adjusted jackknife empirical likelihood
[b]Extended jackknife empirical likelihood
[c]Upper bound
[d]Lower bound