Iraqi Journal of Statistical Science (18) 2010
p.p. [59-73]

Re-sampling in Linear Regression Model Using Jackknife and Bootstrap

Zakariya Y. Algamal [*] **Khairy B. Rasheed** [**]

Abstract

Statistical inference is based generally on some estimates that are functions of the data. Resampling methods offer strategies to estimate or approximate the sampling distribution of a statistic. In this article, two resampling methods are studied, jackknife and bootstrap, where the main objective is to examine the accuracy of these methods in estimating the distribution of the regression parameters through different sample sizes and different bootstrap replications.

Keywords: Jackknife, Bootstrap, Multiple regression, Bias , Variance.

[*]Lecture/ Dept. of statistics/College of Computers Sciences and Mathematics/ University of Mosul. **zak_stat@yahoo.com**

[**]Lecture/ Dept. of statistics/College of Computers Sciences and Mathematics/ University of Mosul **khairy_line@yahoo.com**

Received:29/6 /2010 Accepted: 10 /10 / 2010

المعاينة بإرجاع نموذج الانحدار الخطي باستخدام Jackknife و Bootstrap

الملخص

إن الاستدلال الإحصائي يعتمد بصورة عامة على مقدرات إحـصائية والتـي تكون دوال لبيانات تلك المقدرات. وان طرائق إعادة المعاينة توفر إستراتيجيات لتقدير أو لتقريب توزيع المعاينة لهذه المقدرات الإحصائية. في بحثنـا هـذا تمـت دراسـة طريقتين من طرائق إعادة المعاينة وهما طريقـة Jackknife و Bootstrap. وقـد تركز هذا البحث على دراسة تأثير أحجام العينات المختلفة في نتائج مقدرات أسـاليب انحدار (Bootstrap) و (Jackknife) ، والتعرف على المـديات المثلـى لأسـلوب (Bootstrap) من خلال طرح العديد من قيم التكرارات . كما تضمن هـذا البحـث عرض أهم الأسباب التي تقودنا إلى استخدام هذه الأساليب عوضا عن طريقة المربعات الصغرى الاعتيادية في تقدير معلمات نموذج الانحدار الخطي .

1- Introduction

Two of the most important problems in applied statistics are the determination of an estimator for a particular parameter of interest and the evaluation of the accuracy of that estimator through estimates of the standard error of the estimator and the determination of confidence intervals for the parameter (Chernick, 2008). Jackknife and bootstrap resampling methods are designed to estimate standard errors, bias, confidence intervals, and prediction error. The jackknife preceded the bootstrap. The jackknife resampling is generated by sequentially deleting single datum from the original sample (Friedl and Stampfer, 2002). The bootstrap is a resampling method that draws a large collection of samples from the

original data. It is used to select the observation randomly with replacement from the original data sample (Efron and Tibshirani, 1993).

One of the most important and frequent types of statistical analysis is regression analysis, in which we study the effects of explanatory variables on a response variable. The use of the jackknife and bootstrap to estimate the sampling distribution of the parameter estimates in linear regression model was first proposed by Efron (1979) and further developed by Freedman(1981), Wu (1986).There has been considerable interest in recent years in the use of the jackknife and bootstrap in the regression context. In this study, we focus on the accuracy of the jackknife and bootstrap resampling methods in estimating the distribution of the regression parameters through different sample sizes and different bootstrap replications. The contents of this article may be divided into seven sections. In sections 2 and 3 we briefly review the jackknifing and bootstrapping regression model respectively. In section 4 we introduce our simulation design, whereas the simulation results, conclusion, and references are given in sections 5,6, and 7 respectively.

2- Jackknifing Linear Regression Model

For the linear regression model

$$Y = X\beta + e \qquad \ldots\ldots\ldots\ldots\ldots(1)$$

where Y denotes the $n\times 1$ vector of the response, $X = (x_1, x_2, ..., x_k)$ is the matrix of regressors with $n\times k$, and e is an $n\times 1$ vector of error which has normal distribution with zero mean

and variance σ_e^2 (Yan and Su, 2009). The least squares estimator is given by

$$\hat{\beta}^{ols} = (X'X)^{-1}X'Y \qquad \ldots\ldots\ldots\ldots(2)$$

The variance – covariance matrix of $\hat{\beta}_{ols}$ is

$$var-cov(\hat{\beta}_{ols}) = \hat{\sigma}^2 (X'X)^{-1} \qquad \ldots\ldots\ldots\ldots\ldots(3)$$

If β is estimated by $\hat{\beta}$, then θ is estimated by $\hat{\theta} = g(\hat{\theta})$, with respective jackknife values $\hat{\theta} = g(\hat{\beta}_i)$ .The jackknife estimation of the variance and bias of the $\hat{\theta}_{ols} = g(\hat{\beta}_{ols})$, delete the pair $(y_i, x_i'), (i = 1,2,..., n)$ and calculate $\hat{\theta}_{ols}(J)$, the least squares estimate of θ based on the rest of the data set (Shao and Tu,1995). The estimation of the $\hat{\beta}_J$, bias and variance with pseudo-values are

$$\hat{\beta}_J = \frac{1}{n}\sum_{i=1}^{n}\tilde{\beta}_{Ji} \qquad \ldots\ldots\ldots\ldots\ldots\ldots..(4)$$

$$bias(J) = (\frac{1}{n})\sum_{i=1}^{n}(\hat{\beta}_{ols} - \tilde{\beta}_{Ji}) \qquad \ldots\ldots\ldots\ldots\ldots(5)$$

$$V(\hat{\beta}_J) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\tilde{\beta}_{Ji} - \hat{\beta}_J)(\tilde{\beta}_{Ji} - \hat{\beta}_J)' \qquad \ldots\ldots\ldots\ldots(6)$$

respectively , where the $\tilde{\beta}_{Ji}$ is the pseudo-value and equals to

$$\tilde{\beta}_{Ji} = n\hat{\beta}_{ols} - (n-1)\hat{\beta}_{Ji} \qquad \ldots\ldots\ldots\ldots\ldots\ldots(7)$$

(Friedl and Stampfer, 2001).

The following are the steps of jackknifing linear regression model (Sahinler and Topuz, 2007):

1- Draw n sized sample from population randomly and label the elements of the vector $w_i = (y_i , x_{ji})'$.

2- Delete the first row of the vector $w_i = (y_i , x_{ji})'$ and label the remaining $n-1$ sized observation sets and estimate the ols regression coefficients $\hat{\beta}_{J1}$ from w_1. Then, omit second row of the vector $w_i = (y_i , x_{ji})'$ after that bring back the deleted first row, label remaining n-1 sized observation sets and estimate the ols regression coefficients $\hat{\beta}_{J2}$ from w_2. Similarly, omit each one of the n observation sets and estimate the regression coefficients as $\hat{\beta}_{Ji}$ alternately, where $\hat{\beta}_{Ji}$ is jackknife regression coefficient vector estimated after deleting of ith observation set from w_i.

3- Calculate the jackknife regression coefficient, bias, and standard error for each coefficient from equation (4),(5), and (6).

3- Bootstrapping Linear Regression Model

Efron(1979) has developed a new re-sampling procedure named as "Bootstrap" . Bootstrap is a resample consists of n elements that are drawn randomly from the n original data observations with replacement (Friedl & Stampfer, 2002). The all bootstrap samples are n^n, but we choose B bootstrap samples. Consider the linear regression model in equation (1), bootstrapping can be done by either re-sampling the residuals,in which the regressors $(x_1, x_2,..., x_k)$ are assumed to be fixed, or resampling the

y_i values and their associated x_i values, in which the re-gressors are assumed to be random. In our study, we deal with the residuals resampling. So, the bootstrapping residuals steps are:

1- On the original data set estimate the regression coefficients and compute the residuals e_i .

2- For $r = 1,2,...,B$ draw a n sized bootstrap sample with replacement $(e_{b1}, e_{b2},..., e_{bB})$ from the residuals e_i ,and compute the bootstrap y values

$$y_b = X\hat{\beta}_{ols} + e_b \qquad \text{……(8)}$$

3- Estimate the regression coefficients based on (8), using

$$\hat{\beta}_{br} = \hat{\beta}_{ols} + (X'X)^{-1} X' e_{br} \qquad \text{……(9)}$$

and repeat steps 2 and 3for r. Then the bootstrap estimate of the regression coefficient is:

$$\hat{\beta}_b = \frac{1}{B} \sum_{r=1}^{B} \hat{\beta}_{br} \qquad \text{……..(10)}$$

The bootstrap bias and the variance are given below(Shao and Tu,1995)

$$\text{bias}(b) = (\hat{\beta}_b - \hat{\beta}_{ols}) \qquad \text{…….(11)}$$

$$V(\hat{\beta}_b) = \frac{1}{B-1} \sum_{r=1}^{B} (\hat{\beta}_{br} - \hat{\beta}_b)(\hat{\beta}_{br} - \hat{\beta}_b)' \qquad \text{……(12)}$$

4- Simulation Study

In this section, we describe the design of our study. We consider our population size that is 1000, and we have three explanatory variables $(k = 3)$, each one has a uniform distribution

with $(0,1)$. The error distribution is assumed normal with mean 0 and variance 4.For multiple linear regression we consider $\beta = (1797.93,85.02,78.89,45.12)$. We draw four samples with sizes $10,30,50,\text{and}100$ respectively. Finally, we took three values of bootstrap samples $(B) = 100,1000,\text{and}10000)$ for each sample size. All computations are done by using R programs for windows.

5- Simulation Results

For each sample size we fit the ordinary least squares linear regression model and jackknifing and for B we fit the bootstrapping regression model. The results are shown in tables (1) and (2), which shows that both ols and jackknife have small difference between the MSE values when the sample size are 30 and 50, also the jackknife's MSE value when the sample size is 10 is greater than the ols since the jackknife samples have size n-1. In general, the MSE values for the bootstrap resampling with varying n and B are less than the ols and jackknife values. Comparing the estimated bootstrap and jackknife coefficients from equations (4) and (10) with the coefficients that are estimated by ols, show that there are a little bias in the jackknife and bootstrap coefficients, and the bias decreases when the sample size and B increase. The jackknifed standard error $S.E(\hat{\beta}_J)$ for the coefficients is greater than the $S.E(\hat{\beta}_{ols})$ and $S.E(\hat{\beta}_b)$ when $n = 10,30$, but when $n = 50,\text{and}100$ the $S.E(\hat{\beta}_J)$ become converge as compared with the $S.E(\hat{\beta}_{ols})$ and $S.E(\hat{\beta}_b)$.The bootstrapped standard error $S.E(\hat{\beta}_b)$ of the coefficients become smaller than the $S.E(\hat{\beta}_{ols})$ when B and n increase. The distributions of the bootstrapped and jackknifed regression coefficients for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}2,\text{and}\hat{\beta}_3$ are graphed in figures (1),(2),(3),

and (4). The histograms of the bootstrap estimates conform quite well to the normal distribution for all parameters when n equals to $10,30,50,\text{and}100$ and B are equal to $100,1000,\text{and}10000$. The jackknife's histograms of the estimated parameters also conform the normal distribution especially when n is $50,\text{and}100$.

Table(1): The least squares method and jackknifing results of the regression parameters

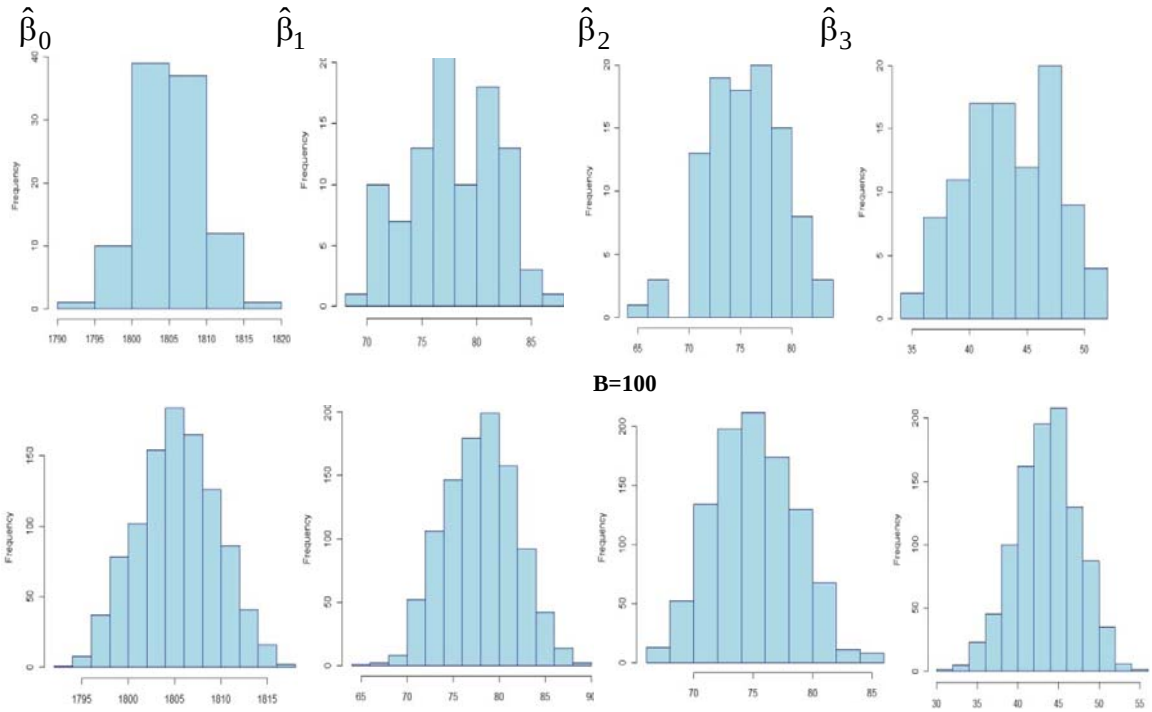
	Regression		**Jackknife**		
	$\hat{\beta}_{ols}$	$S.E(\hat{\beta}_{ols})$	$\hat{\beta}_J$	bias(J)	$S.E(\hat{\beta}_J)$
n=10	**1805.277**	**5.399**	**1806.391**	**-1.151**	**7.013**
	78.104	**4.937**	**76.974**	**1.152**	**8.061**
	75.181	**4.353**	**72.789**	**2.401**	**5.305**
	43.437	**4.747**	**44.749**	**-1.251**	**4.634**
	MSE=13.880		**MSE=14.165**		

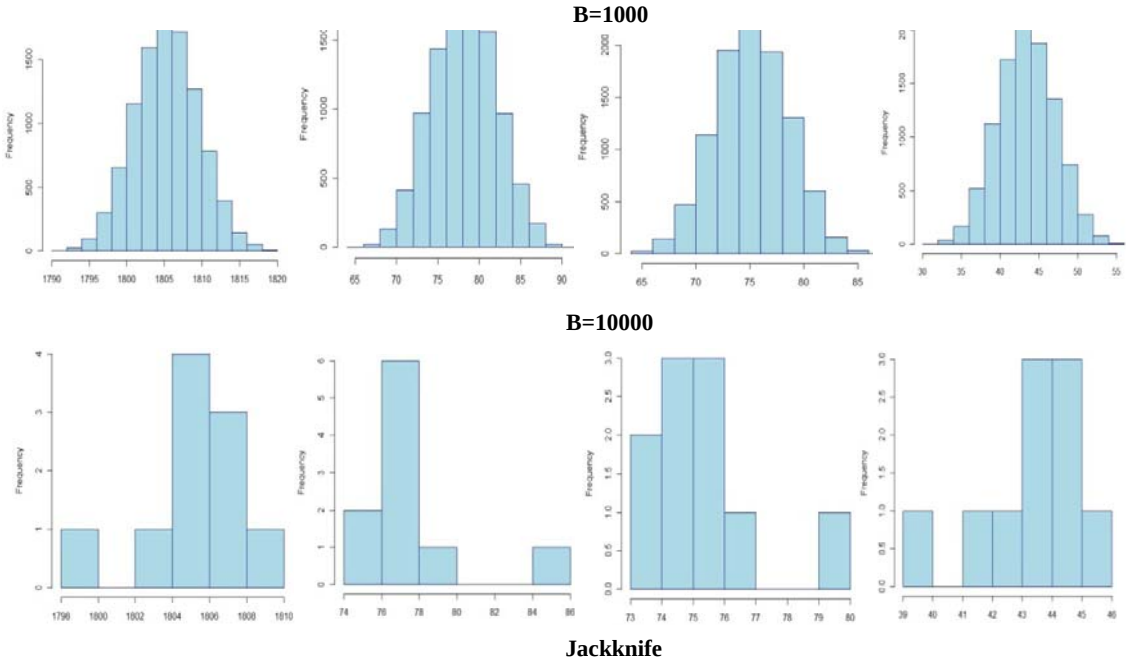
n=30	**1799.365**	**2.689**	**1799.236**	**0.134**	**3.476**
	84.679	**3.993**	**85.543**	**-0.869**	**5.368**
	80.065	**3.141**	**79.538**	**0.528**	**4.217**
	41.807	**3.287**	**42.178**	**-0.393**	**4.149**
	MSE=22.809		**MSE=22.783**		
n=50	**1798.134**	**2.123**	**1798.24**	**-0.098**	**2.323**
	84.986	**2.363**	**84.928**	**0.0549**	**2.464**
	81.013	**2.342**	**80.916**	**0.1036**	**2.18**
	43.827	**2.28**	**43.836**	**-0.034**	**2.086**
	MSE=20.532		**MSE=20.501**		
n=100	**1797.095**	**1.391**	**1797.097**	**0.0173**	**1.363**
	85.449	**1.533**	**85.428**	**0.0062**	**1.585**
	76.923	**1.405**	**76.941**	**-0.026**	**1.448**
	48.545	**1.48**	**48.537**	**0.0003**	**1.419**
	MSE=17.802		**MSE=17.765**		

Table(2): The bootstrap results of the regression parameters

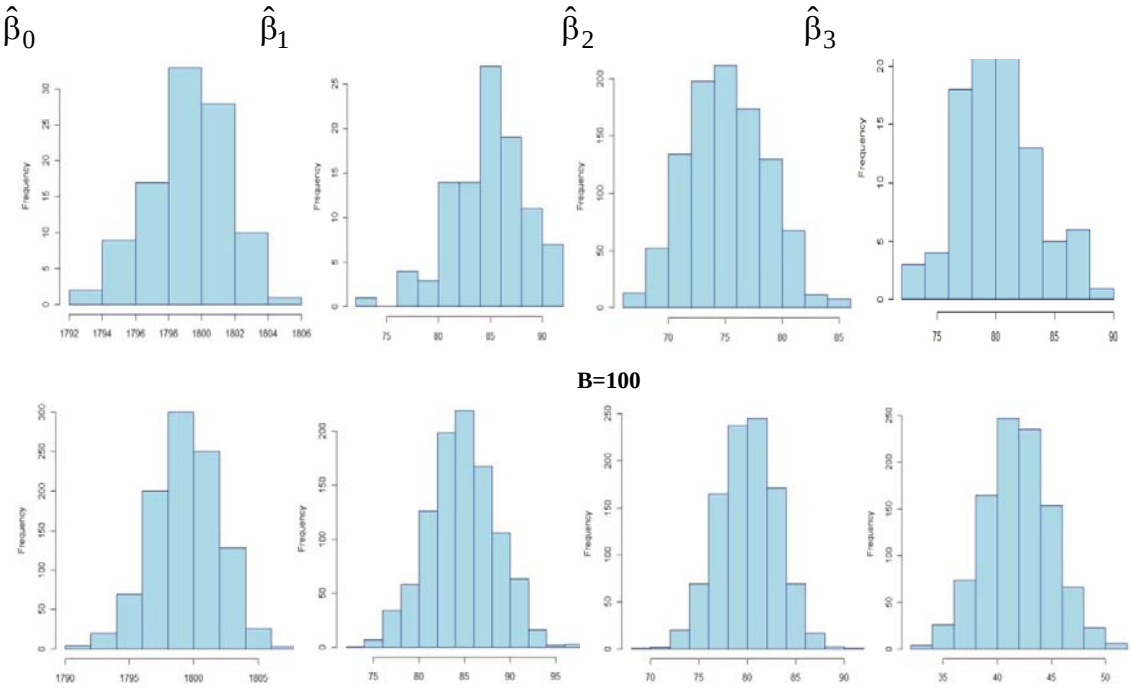
	B=100			**B=1000**			**B=10,000**		
	$\hat{\beta}_b$	bias(b)	S.E($\hat{\beta}_b$)	$\hat{\beta}_b$	bias(b)	S.E($\hat{\beta}_b$)	$\hat{\beta}_b$	bias(b)	S.E($\hat{\beta}_b$)
n=10	**1804.65**	**-0.621**	**4.139**	**1805.51**	**0.234**	**4.211**	**1805.27**	**-0.004**	**4.226**
	78.004	**-0.099**	**3.956**	**77.939**	**-0.164**	**3.877**	**78.136**	**0.031**	**3.894**
	75.674	**0.492**	**3.292**	**75.013**	**-0.168**	**3.411**	**75.136**	**-0.045**	**3.404**
	44.425	**0.987**	**3.754**	**43.262**	**-0.174**	**3.727**	**43.46**	**0.023**	**3.708**
	MSE=8.776			**MSE=8.306**			**MSE=8.447**		
n=30	**1799.65**	**0.289**	**2.303**	**1799.53**	**0.167**	**2.464**	**1799.35**	**-0.009**	**2.488**
	84.739	**0.059**	**3.956**	**84.554**	**-0.125**	**3.671**	**84.729**	**0.049**	**3.71**
	80.136	**0.071**	**2.544**	**79.984**	**-0.081**	**2.86**	**80.047**	**-0.017**	**2.935**

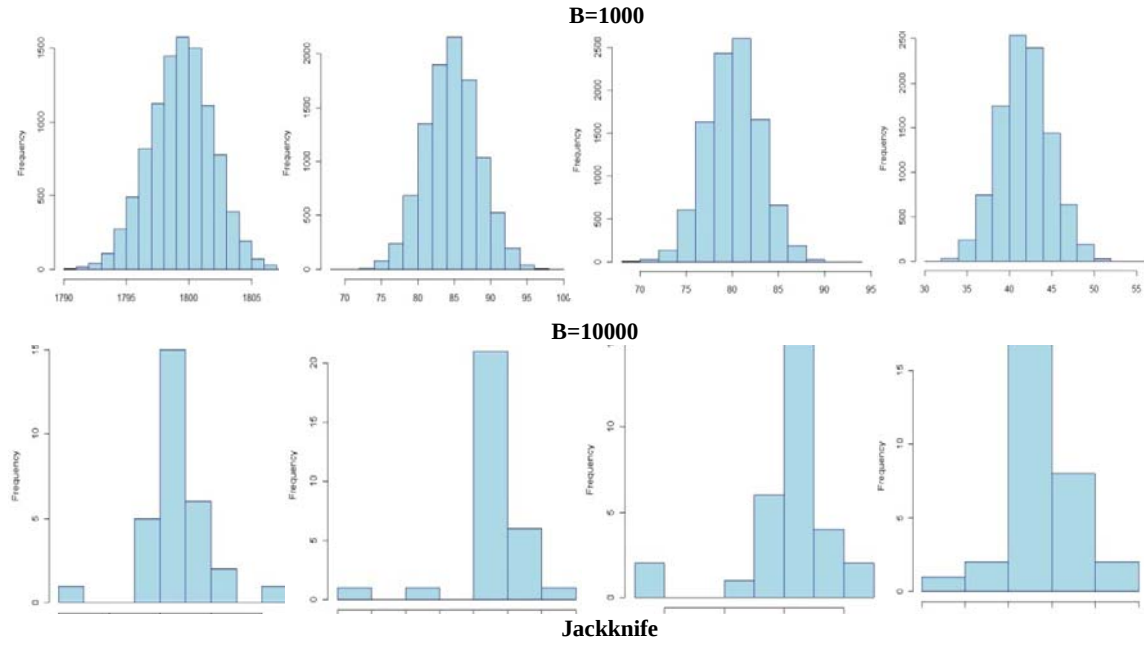
	41.019	**-0.788**	**3.011**	**41.669**	**-0.138**	**2.97**	**41.807**	**0**	**3.065**
	MSE=19.1022			**MSE=19.675**			**MSE=19.656**		
n=50	**1798.0**	**-0.035**	**2.112**	**1798.10**	**-0.29**	**2.045**	**1798.15**	**0.016**	**2.042**
	85.064	**0.077**	**2.008**	**85.013**	**0.027**	**2.292**	**84.946**	**-0.04**	**2.251**
	81.054	**0.04**	**2.109**	**81.056**	**0.042**	**2.154**	**81.022**	**-0.009**	**2.246**
	43.777	**-0.049**	**2.306**	**43.802**	**-0.024**	**2.158**	**43.822**	**-0.005**	**2.19**
	MSE=18.918			**MSE= 19.037**			**MSE= 18.852**		
n=100	**1797.2**	**0.105**	**1.535**	**1797.0**	**-0.021**	**1.379**	**1797.1**	**0.012**	**1.358**
	85.54	**0.091**	**1.544**	**85.428**	**-0.021**	**1.5**	**85.427**	**-0.021**	**1.505**
	76.649	**-0.273**	**1.367**	**77**	**0.076**	**1.381**	**76.921**	**-0.001**	**1.372**
	48.609	**0.063**	**1.413**	**48.527**	**-0.018**	**1.477**	**48.551**	**0.005**	**1.456**
	MSE=16.981			**MSE= 16.877**			**MSE= 17.026**		



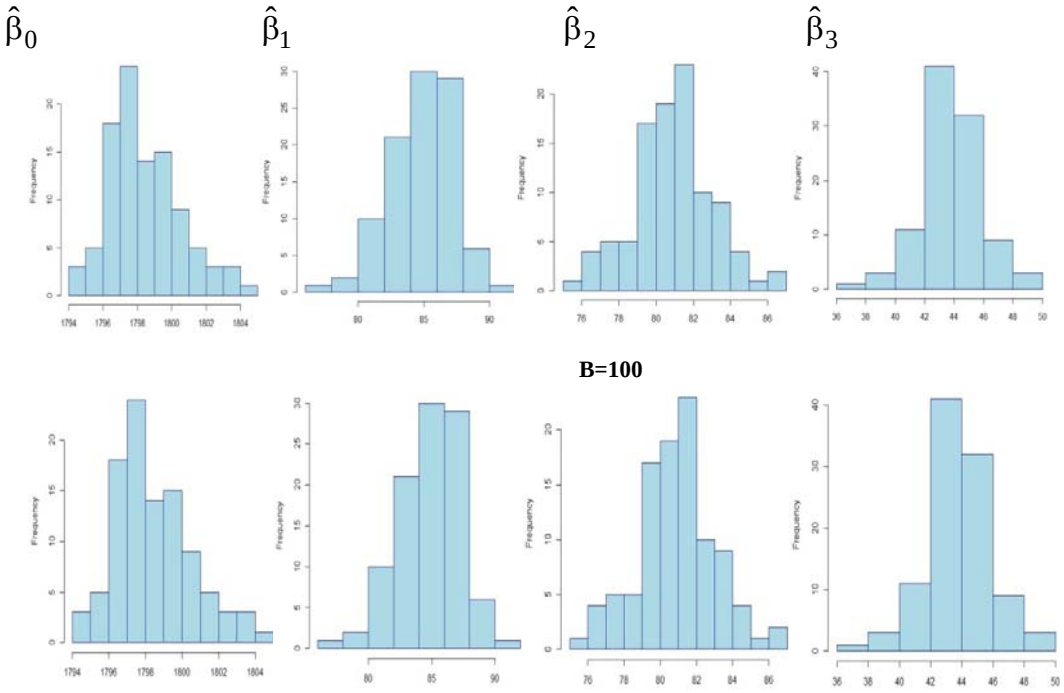


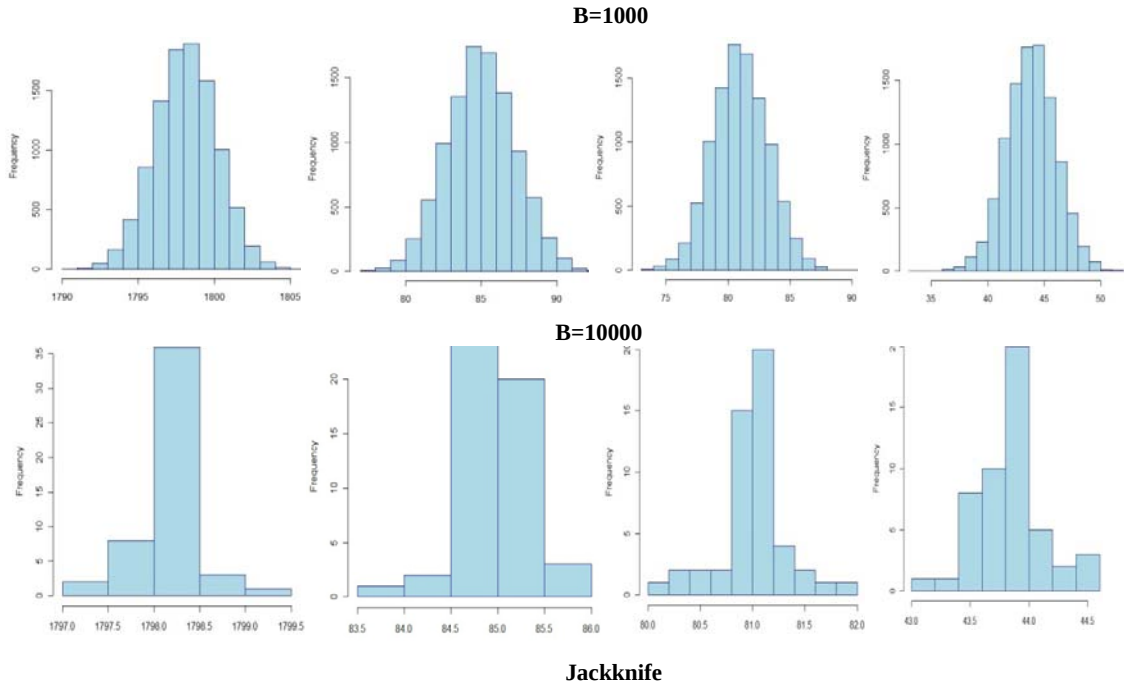
Figure(1): The histogram of the bootstrap and jackknife for n=10.



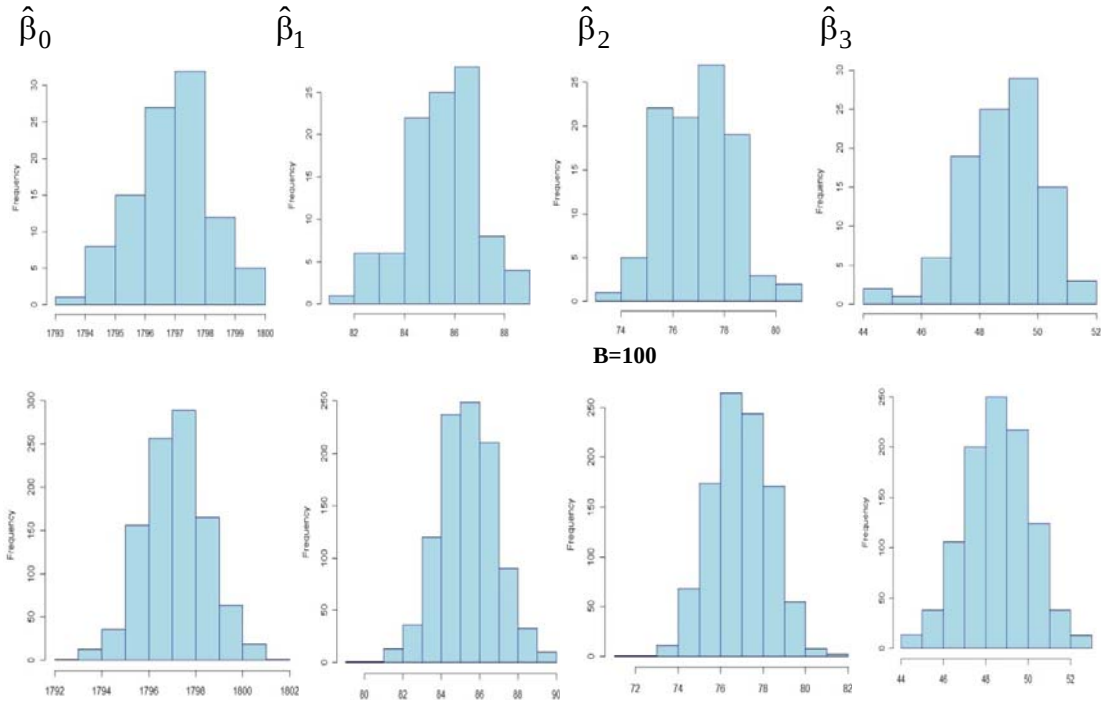


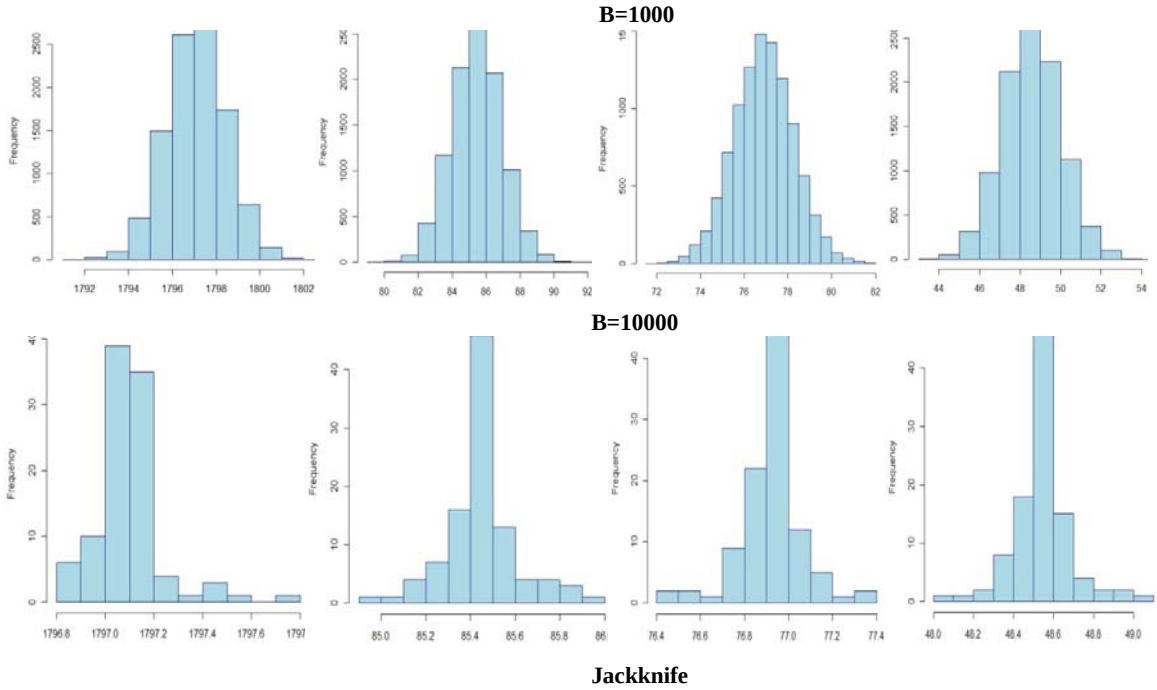
Figure(2): The histogram of the bootstrap and jackknife for n=30.





Figure(3): The histogram of the bootstrap and jackknife for n=50.





6-Conclusion

Bootstrap and jackknife methods are sample reuse techniques designed to estimate standard errors and confidence intervals. As a conclusion, we can rely on the jackknife results when the sample size is large enough $(n \geq 50)$. When B is increased we can get best results and less bias in bootstrap resampling . The histograms conform well to the normal distribution when the number of bootstrap replications B is enough large i.e. $B = 10000$ and the sample size being large too. The jackknife resampling results close

to the results of the bootstrap resampling when n is enough sufficient and B is large too.

7-Refrences

1- Chernick, M., R., (2008), " Bootstrap Methods, A Guide for Practitioners and Researchers" , 2nd ed., John Wiley & Sons, Inc.,New Jersey.
2- Efron, B. (1979) "Bootstrap Methods: Another look at Jackknife", Annals of Statistics,Vol.7, pp.1-26.
3- Efron, B. and Tibshirani, R., (1993), "An introduction to the bootstrap", Chapman and Hall , New York.
4- Freedman, D.,A.,(1981) "Bootstrapping Regression Models", Annals of Statistics,Vol.9, No.6, pp.1218-1228.
5- Friedl, H. and Stampfer, E.,(2002), "Jackknife Resampling", Encyclopedia of Environmetrics, 2, pp.1089-1098.
6- Sahinler, S. and Topuz, D., (2007), "Bootstrap and Jackknife Resampling Algorithms for Estimation of Regression Parameters", Journal of Applied Quantitative methods,Vol.2,No.2,pp.188-199.
7- Shao, J. and Tu, D.,(1995), "The Jackknife and Bootstrap", Springer-Verlag, New York.
8- Wu,C.,F.,J., (1986), "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis", Annals of tatistics,Vol.14, No.4 pp.1261-1295.
9- Yan, X. and Su, X.,G.,(2009), "Regression Analysis, Theory and Computing", World Scientific publishing Co., Uk.