International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Investigating the Significance of a Correlation Coefficient using Jackknife Estimates

Anthony Akpanta[a], Idika Okorie[b]

[a,b]*Department of Statistics, Abia State University Uturu,Nigeria*

[a]Email: *ac_akpa@yahoo.com*

[b]*Email:iokorie@yahoo.com*

Abstract

Often in Applied statistics, population parameters are not known and could be inferred using the available sample data and this is the underpinning of statistical inference. Resampling technique such as jackknife offers effective estimates of parameters and its asymptotic distribution. In this paper, we present the jackknife estimate of the parameters of a simple linear regression model with particular interest on the correlation coefficient. This procedure provides an effective alternative test statistic for testing the null hypothesis of no association between the explanatory variables and a response variable.

Keywords: Jackknife; simple linear regression; correlation coefficient; ols estimates; bias.

1. Introduction

After estimation of parameters in applied statistics it is always crucial to assess the accuracy of the estimator by its standard error and construction of confidence intervals for the parameter [1].

Quenouille in 1956 developed a cross validation procedure known as jackknife (leave-one-out procedure) for estimating the bias of an estimator [2]. Two years later this method was further extended by John Tukey to estimate the variance of an estimator and the name Jackknife was coined for this cross validation method [3].

---

* Corresponding author.

E-mail address: ac_akpa@yahoo.com

The jackknife algorithm is an iterative procedure. The initial step is to estimate the parameter(s) from the entire sample. Then the ith element (datum) is sequentially dropped from the sample and the model parameters estimated from the reduced sample data. The resultant estimates are called the partial estimate (pseudo estimates) [4]. The mean of the pseudo estimates is referred to as the jackknife estimate used in place of the main parameter value [5]. Also, from the pseudo estimates the standard errors of the parameters could be estimated using the standard deviation in order to enable a statistically significant test of the parameters and the construction of the confidence interval [6].

Regression analysis has been widely used to explain the relationship between the explanatory variables and a response variable. However, jackknife was found viable in estimating the sampling distribution of the regression coefficients in the work of Efron [7], and further extended by Freedman [8] and Wu [9].

With a special case of the simple linear regression model, this article is aimed at illustrating an alternative to the classic test statistic for assessing the significance of the correlation coefficient using jackknife estimates.

2. Methods

The linear regression model could be given in matrix form

$$\mathbf{Y} = \mathbf{X\theta} + \boldsymbol{\varepsilon} \tag{1}$$

Where

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n\times p},$$

is the $n \times p$ design matrix (matrix of the explanatory variables) and the remaining quantities are vectors corresponding to $p \times 1$ regression parameters, $n \times 1$ response variable and $n \times 1$ normally distributed error term with zero mean and constant variance, defined by

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}_{p\times 1}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}_{n\times 1} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}.$$

The simple linear regression model with one explanatory variable ($x'_i, i = 1,2,3,\cdots,n$) and two parameters $\theta_{(0)}$ and $\theta_{(1)}$ corresponding to the intercept and slope parameter is a special case of (1). Hence, the ordinary least square (ols) estimator of this model is

$$\begin{pmatrix} \hat{\theta}^{ols}_{(0)} \\ \hat{\theta}^{ols}_{(1)} \end{pmatrix} = (\mathbf{X'X})^{-1}\mathbf{X'Y} \tag{2}$$

Where

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{n\times 2}$$

With variance covariance matrix of $\hat{\theta}_{(0)}^{ols}$ and $\hat{\theta}_{(1)}^{ols}$ given by

$$\text{var/cov}\begin{pmatrix} \hat{\theta}_{(0)}^{ols} \\ \hat{\theta}_{(1)}^{ols} \end{pmatrix} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}{}_{2\times 2} \tag{3}$$

Where the diagonal elements of (3) are the variances of $\hat{\theta}_{(0)}^{ols}$ and $\hat{\theta}_{(1)}^{ols}$ respectively, and the off-diagonals are their co-variances.

Also, the least squares estimate of the correlation coefficient which measures the strength of a linear relationship is given by the Pearson product moment estimate

$$\hat{\rho}_{x,y} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(n\sum_{i=1}^{n} x_i^2 - [\sum_{i=1}^{n} x_i]^2)(n\sum_{i=1}^{n} y_i^2 - [\sum_{i=1}^{n} y_i]^2)}}. \tag{4}$$

This measure of strength lies within $-1 \leq \hat{\rho}_{x,y} \leq 1$ where the closer it is to 1 the stronger the positive, if closer to -1 then the stronger the negative relationship, and the closer it is to 0, the weaker the relationship. Interestingly, -1, 0 and 1 estimates of this measure imply perfect negative, no and perfect positive relationships, respectively. Also, it is often necessary to test the significance of this parameter with the following hypothesis and test statistic

Hypothesis:

H_0: $\hat{\rho}_{x,y} = 0$

H_1: $\hat{\rho}_{x,y} \neq 0$

Test Statistic

$\frac{\hat{\rho}_{x,y}\sqrt{n-2}}{\sqrt{1-\hat{\rho}_{x,y}{}^2}} \sim t_{\alpha,(n-2)}$.

However, the test statistic above is classical, and in this article we propose a jackknife based statistic $\frac{\hat{\rho}_{x,y(J)}}{\sqrt{\text{var}(\hat{\rho}_{x,y(J)})}} \sim t_{\alpha,(n-2)}$ for testing the above hypothesis.

The jackknife estimates of (2) and (4) is obtained by leaving-out the ith observation of the pair $y_i, x'_i; i =$

$1,2,3,\cdots,n$ and evaluating $\tilde{\theta}^{ols}(J)$ and $\tilde{\rho}_{x,y}(J)$ the least squares estimates based on the remaining observations [10]. The estimates of $\hat{\theta}_J$ and $\hat{\rho}_J$, bias and variance using the pseudo values $(\tilde{\theta}_{Ji}$ and $\tilde{\rho}_{x,y(Ji)})$ are

$$\hat{\theta}_J = \frac{\sum_{i=1}^{n} \tilde{\theta}_{Ji}}{n} \tag{5}$$

With bias

$$\text{bias} = \frac{\sum_{i=1}^{n} (\hat{\theta}^{ols} - \tilde{\theta}_{Ji})}{n} \tag{6}$$

Or more succinctly

$$\text{bias} = \hat{\theta}^{ols} - \hat{\theta}_J \tag{7}$$

And the variance

$$\text{var}(\hat{\theta}_J) = \frac{\sum_{i=1}^{n} (\tilde{\theta}_{Ji} - \hat{\theta}_J)^2}{n(n-1)} \tag{8}$$

Also,

$$\hat{\rho}_{x,y(J)} = \frac{\sum_{i=1}^{n} \tilde{\rho}_{x,y(Ji)}}{n} \tag{9}$$

With bias

$$\text{bias} = \frac{\sum_{i=1}^{n} (\hat{\rho}_{x,y} - \tilde{\rho}_{x,y(Ji)})}{n} \tag{10}$$

Or

$$\text{bias} = \hat{\rho}_{x,y} - \hat{\rho}_{x,y(J)} \tag{11}$$

And variance

$$\text{var}(\hat{\rho}_{x,y(J)}) = \frac{\sum_{i=1}^{n} (\tilde{\rho}_{x,y(Ji)} - \hat{\rho}_{x,y(J)})^2}{n(n-1)} \tag{12}$$

2.1 Algorithm for Jackknifing Simple Linear Regression Model

Steps:

- Using a pair of independent sample of size (n) of explanatory and response variables $(x_i, y_i)'$, $i = 1, 2, 3, \cdots, n$.
- Drop the first datum in both variable and estimate the ordinary least squares (ols) regression coefficients $(\tilde{\theta}_{(0)J1}$ and $\tilde{\theta}_{(1)J1})$ and the correlation coefficient $(\tilde{\rho}_{x,y(J1)})$ using $n - 1$ observations.
- Drop the second datum and replace the initially dropped datum in (ii) and compute the ordinary least squares (ols) regression coefficients $(\tilde{\theta}_{(0)J2}$ and $\tilde{\theta}_{(1)J2})$ and the correlation coefficient $(\tilde{\rho}_{x,y(J2)})$ using $n - 1$ observations.
- Repeat steps (ii) and (iii) by replacing the $(i - 1)$th previously dropped observation and dropping the ith observation and then computing the ordinary least squares (ols) regression coefficients $(\tilde{\theta}_{(0)Ji}$ and $\tilde{\theta}_{(1)Ji}, i = 3, 4, 5, \cdots, n)$ and the correlation coefficient $(\tilde{\rho}_{x,y(Ji)}, i = 3, 4, 5, \cdots, n)$ using $n - 1$ observations at each iteration until all the observations in the pair $(x_i, y_i)'$, $i = 1, 2, 3, \cdots, n$ has been sequentially dropped and replaced in turns. Steps (ii) to (iv) results to an n dimensional vectors of pseudo values corresponding to $\tilde{\theta}'_{(0)Ji}, \tilde{\theta}'_{(1)Ji}$ and $\tilde{\rho}'_{x,y(Ji)}$.
- Compute the jackknife regression parameters, correlation coefficients and their corresponding bias and standard errors using (5), (7), (8), (9), (11), and (12).

3. Data and Simulation

We have used the total demand and supply of FOREX (USD million) data from January 2008 to May 20014 (77 data points) available on the Central Bank of Nigeria official website [11]. All computations are done using R programs for windows.

3.1 Simulation Results

Using the data in 2.0 we fit a simple linear regression model and the result is shown in Table 1.

Table 1: Parameter Estimates for the Fitted Simple Linear Regression Model

	Parameters		
	$\theta_{(0)}$	$\theta_{(1)}$	$\rho_{x,y}$
Estimate	925.33554	0.49293	0.7232257
Standard Error	182.61583	0.05435	-

3.1.1 Jackknifing the Simple Linear Regression Model

Table 2 shows the ols estimates of the pseudo values, jackknife estimates and their corresponding standard errors obtained from the leave-one-out procedure.

Table 2: ols Estimates

S/N	$\widetilde{\theta}_{(0)Ji}$	$\widetilde{\theta}_{(1)Ji}$	$\widetilde{\rho}_{x,y(Ji)}$
1	952.7329	0.4867450	0.7188903
2	966.5801	0.4822654	0.7111098
3	972.6817	0.4806240	0.7101433
4	959.0865	0.4842215	0.7119457
5	945.4296	0.4886822	0.7203511
⋮	⋮	⋮	⋮
73	868.0236	0.5149849	0.7233487
74	935.5823	0.4876679	0.7174566
75	936.0865	0.4877716	0.7162645
76	926.6243	0.4899442	0.7226179
77	920.3058	0.4905668	0.7264761
	$\widehat{\theta}_{(0)J}$	$\widehat{\theta}_{(1)J}$	$\widehat{\rho}_{x,y(J)}$
	925.023	**0.4930578**	**0.7232903**
	SE $\widehat{\theta}_{(0)J}$	SE $\widehat{\theta}_{(1)J}$	SE $\widehat{\rho}_{x,y(J)}$
	35.01666	**0.01321761**	**0.0142381**

Table 3: Comparison between ols and Jackknife ols Estimates

Estimates	ols	Jackknife	Bias
$\theta_{(0)}$	925.33554	925.023	0.31254
$SE(\theta_{(0)})$	182.61583	35.01666	-
$\theta_{(1)}$	0.49293	0.4930578	-0.0001278
$SE(\theta_{(1)})$	0.05435	0.01321761	-
$\rho_{x,y}$	0.7232257	0.7232903	-0.0000646
$SE(\rho_{x,y})$	-	0.0142381	-

3.1.2. Testing the significance of the correlation coefficient

We shall proceed to test the significance of the correlation coefficient at 5% level of significance as follows:

H_0: $\hat{\rho}_{x,y} = 0$

H_1: $\hat{\rho}_{x,y} \neq 0$

$$\text{classic} = \frac{\hat{\rho}_{x,y}\sqrt{n-2}}{\sqrt{1-\hat{\rho}_{x,y}{}^{2}}} = \frac{(0.7232903)\sqrt{77-2}}{\sqrt{1-(0.7232903)^{2}}} = 9.07093$$

$$\text{jackknife} = \frac{\hat{\rho}_{x,y(J)}}{SE\left(\hat{\rho}_{x,y(J)}\right)} = \frac{0.7232903}{0.0142381} = 50.79964$$

with critical value $t_{\alpha,(n-2)} = t_{0.05,(77-2)} = 1.992102$.

Decision:

Since both test statistics are larger than the critical value, we therefore conclude that there is enough evidence against the null hypothesis; hence, the correlation coefficient is significantly different from 0 at 5% level of significance.

3.1.3. Discussions

The jackknife (leave-one-out) ols estimator provides better estimates of the regression parameters than the ols method. From Table 3 above it could be seen that the Jackknife estimates of both the regression coefficients $(\theta_{(0)}$ and $\theta_{(1)})$ and the correlation coefficient $(\rho_{x,y})$ are approximately the ols estimates with very small bias, it is interesting to observe that the Jackknife estimates has smaller standard errors (Efficiency property), a unique feature of a good estimator in comparison to their ols counterpart. The classic test statistic value for testing the significance of the correlation coefficient is smaller than the value obtained from the proposed jackknife test statistic; this is a consequence of a large variance of the ols estimates.

4. Conclusion

Jackknife results are misleading when the sample size is not large enough $(n < 50)$, [12]. Factually, the 77 observations used in this study reveals that the Jackknife estimators are more efficient than their ols counterpart in estimating the coefficients of a linear regression model and the correlation coefficient. It also provides the asymptotic distribution of the above mentioned parameters, e.g., Table 2. The classic test statistic for testing the significance of the correlation coefficient is under-estimated, an effect of large standard error of the ols estimators and consequently, could lead to erroneously accepting the null hypothesis (Type II error). Without loss of generality, the jackknife based test statistic is better than its classic counterpart.

References

[1] M. R. Chernick. *Bootstrap Methods a Guide for Practitioners and Researchers*. 2nd ed; John Wiley & Sons Inc., New Jersey, 2008.

[2] M. H. Quenouille. "Notes on Bias in Estimation", *Biometrika*, 61, pp. 1-17, 1956.

[3] J. W.Tukey."Bias, and Confidence in not Quite Large Samples (Abstract)" *Annals of Mathematical*

Statistics, 29, pp. 614, 1985.

[4] H. Friedl and E. Stampfer. "Jackknife Resampling", *Encyclopaedia of Econometrics*, 2, pp. 1089-1098, 2002.

[5] S. Sahinler and D.Topuz. "Bootstrap and Jackknife Resampling Algorithms for Estimation of Regression Parameters", *Journal of Applied Quantitative Methods,* Vol. 2. No. 2. pp. 188-199, 2007.

[6] H. Abdi and J. L. Williams. "Jackknife", In Neil Salkind (Ed.), *Encyclopaedia of Research Design*. Thousand Oaks, CA: Sage,2010.

[7] B. Efron. "Bootstrap Method; another Look at Jackknife". *Annals of Statistics,* Vol. 7, pp. 1-26, 1979.

[8] D.A. Freedman. "Bootstrapping Regression Models", *Annals of Statistics*. Vol.1, No. 6, pp. 1218-1228, 1981

[9] C. F. J. Wu. "Jackknife, Bootstrap and other Resampling Methods in Regression Analysis", *Annals of Statistics*, Vol. 14, No. 4, pp. 1261-1295,1986.

[10] J. Shao and D. Tu. *The Jackknife and Bootstrap*, Springer- Verlag, New York, 1995.

[11] http//www.cbn.gov.ng, date accessed 1\5\2015.

[12] Zakariya, Y. A. and Khairy, B. R., (2010), “Re-sampling in Linear Regression Model Using Jackknife and Bootstrap”, Iraqi Journal of Statistical Science. Vol. 18, pp. 59-73.