Jackknife Algorithm on Linear Regression Estimation

Esemokumo Perewarebo Akpos
E-mail Address: contactperes4goo@gmail.com
Department of Statistics, School of Applied Science, Federal Polytechnic Ekewe
Yenagoa, Bayelsa State, Nigeria

Bekesuoyeibo Rebecca
E-mail Address: Fzimugha@gmail.com
Department of Statistics, School of Applied Science, Federal Polytechnic Ekewe
Yenagoa, Bayelsa State, Nigeria

Okenwe Idochi
E-mail Address: nwond@yahoo.com
Department of Statistics, School of Applied Sciences, Rivers State Polytechnic
PMB 20, Bori, Rivers State Nigeria

ABSTRACT
In this paper, interest was on the estimation of simple linear regression data using Jackknife algorithm. Thus, Jackknife delete-one algorithm was employed to provide estimates of simple linear regression coefficient. Observations on systolic blood pressure (SBP) and age for a sample of 30 randomly selected patients were collected from Federal Medical Centre Owerri Imo State Nigeria. It was discovered that all errors in the y-direction are normally distributed. The statistical software known as Stata version 9.1 was employed for the ease of the analysis. Pseudo-Values, Jackknife Estimates, and the Jackknife Standard Error were computed. From the analysis, it was revealed that the bias result of the correlation was positive. The result from the OLS shows that SBP on Age of patients is significant. The jackknife standard error and confidence intervals of the Age coefficient based on the distribution $F\left(\hat{\beta}^{(J)}\right)$ are substantially larger than the estimated OLS standard error due to the inadequacy of the jackknife in small samples. Comparing the jackknife coefficients averages $\overline{\hat{\beta}}_0^{(J)}$ and $\overline{\hat{\beta}}_1^{(J)}$ with the corresponding OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ shows that there is a little bias in the jackknife coefficients.

Keywords: Jackknife algorithm, simple regression, Pseudo-Values, Confidence interval, Bias, correlation coefficient

1. Introduction

According to Chernick (2008), there are two basic concerns in statistics, which 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. Friedl and Stampfer (2002) in their study worked on Jackknife re-sampling method which preceded the bootstrap to estimate standard errors, bias, confidence intervals, and prediction error. The jackknife re-sampling is generated by sequentially deleting single datum from the original sample. Specifically, the Jackknife is a re-sampling technique use for estimating the bias and standard error of an estimator and provides an approximate confidence interval for the parameter of interest. The principle behind jackknife method lies in systematically re-computing the statistic leaving out one or more observation(s) at a time from the sample set thereby generating n separate samples each of size n – 1 or n – d respectively. From this new set of replicates of the statistic, an estimate for bias and the variance of the statistic can be calculated

Efron and Tibshirani (1998). It is obvious that Efron (1979) first proposed the use of jackknife and bootstrap to estimate the sampling distribution of the parameter estimates in linear regression model, and it was further developed by Freedman (1981), Wu (1986).

According to Quenouille (1956), the jackknife or "leave one out" procedure is a cross-validation technique used to obtain the bias of an estimator. Again in 1958, Tukey then extended the use of the jackknife to incorporate variance estimation and tailored the name of jackknife due to like a jackknife-a pocket knife akin to a Swiss army knife and typically used by boy scouts-this technique can be employed as a "quick and dirty" substitute tool for a lot of more specific tools.

The jackknife can be used to estimate the actual predictive power of such models by predicting the dependent variable value of each observation as if this observation were a new observation. In order to do so, the predicted value(s) of each observation is (are) obtained from the model built on the sample of observations minus the observation to be predicted. The jackknife, in this context, is a procedure which is used to obtain an unbiased prediction (i.e. a random effect) and to minimize the risk of over-fitting. The aim of this study is to estimate the simple linear regression coefficient parameters using the jackknife delete-one algorithm.

2. Related work

Zakariya and Khairy (2010) worked on Re-sampling in Linear Regression Model using Jackknife and Bootstrap. Statistical inference was based generally on some estimates that are functions of the data. Re-sampling methods offer strategies to estimate or approximate the sampling distribution of a statistic. In their study, two re-sampling methods were studied, jackknife and bootstrap, where the main objective was to examine the accuracy of these methods in estimating the distribution of the regression parameters through different sample sizes and different bootstrap replications. Bootstrap and jackknife methods are sample reuse techniques designed to estimate standard errors and confidence intervals. As a conclusion, they relied on the jackknife results when the sample size is large enough ($n \geq 50$). When B is increased they can get best results and less bias in bootstrap re-sampling. 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 re-sampling results close to the results of the bootstrap re-sampling when n is enough sufficient and B is large too.

Iheagwara and Opara (2014) worked on minimization of error in exponential model estimation via jackknife algorithm. The data were on 25 samples of percentage sugar and percentage of Nitrogen in tobacco leaf for organic and inorganic chemical constituents. The study concerning the use of jackknife methods in estimating the parameters of non linear regression models were identified in the study. An algorithm for the estimation of nonlinear regression parameters was stated. For estimating these parameters, computer programs were written in Stata for the implementation of these algorithms. In the estimation of the nonlinear regression parameters, the results obtained from numerical problems using the Jackknife based algorithm developed yielded a reduced error sum of squares than the analytic result. As the number of d observations deleted in each re-sampling stage increases, so does the error sum of squares reduces minimally. This revealed the appropriateness of the algorithms for the estimation of nonlinear regression parameters and in the reduction of the error terms in nonlinear regression estimation.

Akpanta and Okorie (2015) worked on investigating the significance of a correlation coefficient using Jackknife estimates. In their study, they presented the jackknife estimate of the parameters of a simple linear regression model with particular interest on the correlation coefficient. The procedure provided an effective alternative test statistic for testing the null hypothesis of no association between the explanatory variables and a response variable.

Lu and Su (2015) carried a research on Jackknife model averaging for quantile regressions. In their study, they considered model averaging for Quantile Regressions (QR) when all models under investigation were potentially misspecified and the number of parameters was diverging with the sample size. To allow for the dependence between the error terms and regressors in the QR models, they proposed a Jackknife Model Averaging (JMA) estimator which selected the weights by minimizing a leave-one-out cross-validation criterion function and demonstrated its asymptotic optimality in terms of minimizing the out-of-sample final prediction error. They conducted simulations to demonstrate the finite-sample performance of the estimator and compared it with other model selection and averaging methods. They applied their JMA method to forecast quantiles of excess stock returns and wages.

3. Methodology

In jackknife re-sampling, either a single observation is deleted from the original sample which is called the delete-one jackknife or multiple observations are deleted from the original sample which is called the delete-d jackknife (Efron and Gong, 1983; Wu, 1986; Shao and Tu, 1995).

For instance, let the vector (p×1) $w_i = (y_i, x_{ji})'$, (i = 1,2,…,n) denote the values associated with ith observation, in this case, the set of observations are the vectors $(w_1, w_2, ..., w_n)$. In this study, the delete-one jackknife algorithms for the regression shall be used; thus, the algorithms are stated as:

Take n sized sample from a population randomly and label the elements vector $w_i = (Y_i, X_{ji})'$, as the vector $Y_i = (y_1, y_2, ..., y_n)'$ and the matrix $X_{ji} = (x_{j1}, x_{j2}, x_{j3}, ..., x_{jn})'$, where i=1,2,3,…,k, i=1,2,3,…,n.

i. Remove first row of the vector $w_i = (Y_i, X_{ji})'$, and name the remaining n-1 sized observation sets $Y_i^{(J)} = (Y_2^{(J)}, Y_3^{(J)}, ..., Y_n^{(J)})'$, and $X_{ji}^{(J)} = (x_{j2}^{(J)}, x_{j3}^{(J)}, ..., x_{jn}^{(J)})'$, as delete-one Jackknife sample $(w_1^{(J)})$ and estimate the OLS regression coefficient $\hat{\beta}^{(J_1)}$ from $(w_1^{(J)})$. Then, remove second row of the vector $w_i = (Y_i, X_{ji})'$, and name the remaining n-1 sized observation sets $Y_i^{(J)} = (y_1^{(J)}, y_3^{(J)}, ..., y_n^{(J)})'$, and $X_{ji}^{(J)} = (x_{j1}^{(J)}, x_{j3}^{(J)}, ..., x_{jn}^{(J)})'$ as $(w_2^{(J)})$ and estimate the OLS regression coefficients $\hat{\beta}^{(J_2)}$. Using the approach, expunge each one of the n observation sets and estimate the regression coefficients as $\hat{\beta}^{(J_i)}$ alternately, where $\hat{\beta}^{(J_i)}$ is Jackknife regression coefficient vector estimated after deleting of ith observation set from w_i.

ii. Determine the probability distribution $f(\hat{\beta}^{(J)})$ of Jackknife estimates $\hat{\beta}^{(J_1)}, \hat{\beta}^{(J_2)}, ..., \hat{\beta}^{(J_n)}$.

iii. Compute the jackknife regression coefficient estimate which is the average of the $f(\hat{\beta}^{(J)})$ distribution (Fox,1997) as;

$$\hat{\beta}^{(J)} = \frac{\sum_{i=1}^{n} \hat{\beta}^{(J_i)}}{n} = \bar{\hat{\beta}}^{(J_i)} \qquad (1)$$

iv. Then, the delete-one Jackknife regression equation becomes;

$$\hat{Y} = X\hat{\beta}^{(J)} + \varepsilon \qquad (2)$$

3.1 Jackknife Bias, Variance, and Confidence Interval

Obviously, the jackknife bias, variance and confidence intervals are estimated by using the following equations from $f(\hat{\beta}^{(J)})$distribution according to Miller (1974).

The jackknife bias is,

$$bi\hat{a}s_J(\hat{\beta}) = (n-1)(\hat{\beta}^{(J)} - \hat{\beta}) \quad (3)$$

The jackknife variance is given by;

$$\text{var}(\hat{\beta}^J) = \frac{(n-1)}{n}\sum_{i=1}^{n}(\hat{\beta}^{(Ji)} - \hat{\beta}^{(J)})(\hat{\beta}^{(Ji)} - \hat{\beta}^{(J)})' \quad (4)$$

where $\hat{\beta}^{(Ji)}$ is the estimate formed from the replicate with ith observation set deleted(Friedl and Stampfer, 2002).

Jackknife (1 – α) 100 % confidence interval according to Efron and Tibshirani (1993) equals;

$$\hat{\beta}^{(J)} - t_{n-p,\frac{\alpha}{2}} \times Se\left(\hat{\beta}^{([J])}\right) < \beta < \hat{\beta}^{(J)} + t_{n-p,\frac{\alpha}{2}} \times Se\left(\hat{\beta}^{([J])}\right) \quad (5)$$

where $t_{n-p,\frac{\alpha}{2}}$ is the critical value of **t** with probability α/2 the right for n-p degrees of freedom; and $S_e\left(\hat{\beta}^{([J])}\right)$ is the standard error of the $\hat{\beta}^{(J)}$.Hence, in this study, it shall be restricted to one explanatory variable.

3.2. Data Collection

Data used in this paper were collected from Federal Medical Center, Owerri Imo State Nigeria. Observations on systolic blood pressure (SBP) and age for a sample of 30 randomly selected patients were collected for the data analysis. The response variable (Y_n) is the SBP, while the predictor variable (X_n) is the age. Table 1 displayed the collected data for this study.

Table 1: Data on Systolic Blood Pressure and Age of Patients

S/N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
SBP	134	140	136	140	135	137	130	133	139	141	149	135	135	134	136
Age	25	33	26	29	27	25	23	26	37	31	38	32	31	28	26

S/N	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
SBP	132	139	136	140	135	137	130	133	139	143	139	151	154	154	141
Age	24	35	26	29	27	25	23	26	37	35	30	41	43	37	28

Data Analysis

First, the ordinary least squares regression model was fitted to data given in Table 1 and the results of the ordinary least squares regression was summarized in Table 2. The regression of SBP on Age is significant as result of variance analysis (p = 0.0000).

Table 2: The summary statistics of regression coefficients for OLS regression

SBP	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]
Age	.9644273	.1149095	8.39	0.000	.7290458 1.199809
cons	109.5374	3.514491	31.17	0.000	102.3383 116.7365
$R^2 = 0.7156, n = 30, SSE = 326.349685, F = 70.4412, MSE = 11.6553459, p - value = 0.0000$					

As a preliminary step, the data are analyzed by a standard regression analysis and we found that the regression equation is equal to:

$$\hat{Y}\,(\text{SBP}) = 109.5374 + 0.9644273\text{Age}$$

Let us now estimate the regression parameters and bias for the jackknife. So, for example, when we drop the first observation, we use the observation 2 through 30 to compute the regression equation with the partial estimates of the slope and intercept.

Table 3: The illustration of the jackknife (jackknife samples, each of size n-1=30-1=29) regression procedure from the data given in Table 1, calculating the jackknife estimates of the regression parameters for each sample for SBP model

	Variable	Observation sets						$\hat{\beta}_0^{(J)}$	$\hat{\beta}_1^{(J)}$
		1	2	3	4	…	30		
1	SBP (Y) Age (X)	Omitted	140 33	136 26	140 29	… …	141 28	109.4596	0.966597
2	SBP (Y) Age (X)	134 25	Omitted	136 26	140 29	… …	141 28	109.4440	0.969108
3	SBP (Y) Age (X)	134 25	140 33	Omitted	140 29	… …	141 28	109.2839	0.971228
4	SBP (Y) Age (X)	134 25	140 33	136 26	Omitted	… …	141 28	109.3544	0.967647
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
30	SBP (Y) Age (X)	134 25	140 33	136 26	140 29	… …	Omitted	109.0509	0.975458
$\hat{\beta}^{(J_0)} = \frac{\sum_{i=1}^{n} \hat{\beta}^{(J_i)}}{30}$								109.55172	0.963981

Table 4: Standard Errors of estimates, Correlation Coefficient and pseudo-values for the regression example of SBP & Age of Patients data

Obsns.	Standard Error estimates					Pseudo-Values	
	$(\hat{\beta}_0^{(J)})_{-n}$	$(\hat{\beta}_1^{(J)})_{-n}$	r_{-n}			$(\hat{\beta}_0^{(J)})_n$	$(\hat{\beta}_1^{(J)})_n$
1	3.756449	0.136177	0.842734			111.7941	0.9015114
2	3.710153	0.136609	0.846581			112.2457	0.8287003
3	3.700696	0.134643	0.845931			116.8893	0.7672025
4	3.676962	0.134579	0.848941			114.8458	0.8710477
5	3.694933	0.134657	0.844156			107.1033	1.025905
6	3.628458	0.132824	0.851704			131.0324	0.3651678
7	3.873474	0.140168	0.835703			94.43884	1.405334
8	3.723045	0.135561	0.842556			100.9932	1.193639
9	3.390208	0.124419	0.867468			149.4572	-0.5807106
. . .	. . .	. . .	. . .			. . .	. . .
29	3.871919	0.13903	0.855278			53.20645	3.144776
30	3.646144	0.133773	0.855687			123.6475	0.6445526
$\hat{\beta}^{(J_0)} = \frac{\sum_{i=1}^{n} \hat{\beta}^{(J_i)}}{30}$	3.766843	0.137599	0.84551			109.13015	0.978917067

The bias of the estimation of the coefficient of correlation is equal to:

$\beta_{jack}(r) = r - r^* = 0.84593 - 0.84551 = 0.00042$

The bias is positive and this implies (as required) that the coefficient of correlation over-estimates the magnitude of the population correlation.

Table 5: The Summary Statistics of the Regression Coefficients for Jackknife Regression

Variable	Observed	Average	S.E	Bias	[95% Conf. Interval]	
Constant	109.5374	109.55172	3.766843	0.415280	101.83723	117.26621
Age	0.9644273	0.963981	0.137599	-0.012943	0.68218	1.24578

4. Conclusion

The Jackknife delete-one algorithm has been used to estimate the regression parameters in this study, and the bias of the estimate, and it was observed that the bias result for the correlation coefficient was positive, which implies that the coefficient of correlation over-estimates the magnitude of the population correlation. The output from the traditional OLS method implies that SBP on Age of patients is significant. The jackknife standard errors and confidence intervals of the Age coefficient based on the distribution $F\left(\hat{\beta}^{(J)}\right)$ are substantially larger than the estimated OLS standard error due to the inadequacy of the jackknife in small samples. On the comparisons of the jackknife coefficients averages $\overline{\hat{\beta}}_0^{(J)}$ and $\overline{\hat{\beta}}_1^{(J)}$ with the corresponding OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ revealed that there is a little bias in the jackknife coefficients.

References

Akpanta, A. and Okorie, I. (2015). Investigating the Significance of a Correlation Coefficient using Jackknife Estimates**.** International Journal of Sciences: Basic and Applied Research (IJSBAR).Volume 22, No 2, pp 441-448.ISSN 2307-4531.

Chernick, M., R., (2008), " Bootstrap Methods, A Guide for Practitioners and Researchers" , 2nd ed., John Wiley & Sons, Inc.,New Jersey.

Efron, B. (1979) "Bootstrap Methods: Another look at Jackknife", Annals of Statistics,Vol.7, pp.1-26.

Efron, B., Gong, G.,(1983). A leisurely look at the bootstrap, the jackknife, and cross-validation, Amer. Statist., 37, pp. 36-48, 1983

Efron, B., Tibshirani, R.J.,(1993). An Introduction to the Bootstrap; Chapman & Hall, New York, 1993

Efron, B., Tibshirani, R.J.,(1993). An Introduction to the Bootstrap; Chapman &Hall, New York, 1993

Fox, J., (1997). Applied Regression Analysis, Linear Models and Related Methods; Sage, 1997

Freedman, D.,A.,(1981) "Bootstrapping Regression Models", Annals of Statistics,Vol.9, No.6, pp.1218-1228.

Friedl, H. and Stampfer, E.,(2002), "Jackknife Re-sampling", Encyclopedia of Environmetrics, 2, pp.1089-1098.

Friedl, H., Stampfer, E.(2002). Re-sampling Methods, Encyclopedia of Environmetrics, 3, Eds.: A. El-Shaarawi, W. Piegorsch, Wiley:Chichester, pp.1768-1770, 2002b

Hongchang, H., and Yuhe, X. (2013). Jackknifed Liu estimator in linear regression models. Wuhan University Journal of Natural Sciences. August 2013, Volume 18, Issue 4, pp 331-336.

Iheagwara, A.I. and Opara, J. (2014). Minimization of error in exponential model estimation via jackknife algorithm. International Journal of Research. Volume 02 Issue 02 February 2015.

Lu, X and Su, L.(2015). Jackknife model averaging for quantile regressions. Journal of EconometricsVolume 188, Issue 1, September 2015, Pages 40–58.

Miller, R.G. (1974). The jackknife: a review. Biometrika, **61**,117.

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

Shao, J., and Rao, J.N.K. (1993). Jackknife inference for heteroscedastic linear regression models. The Canadian Journal of Statistics. Vol.21, No. 4, 1993, pages 377-395.

Shao, J., Tu, D.,(1995). The Jackknife and Bootstrap; Springer, New York, 1995

Tukey, J.W. (1958). Bias and confidence in not quite large samples (abstract) Annals of Mathematical Statistic, 29, 614.

Wu, C.F.J. (1986). "Jackknife bootstrap and other re-sampling methods in regression analysis". *The Annals of Statistics* 1986, vol. 14, No. 4, pp 1261-1295.

Wu, C.F.J. (1986). Jackknife bootstrap and other re-sampling methods in regression analysis. The Annals of Statistics 1986, vol. 14, No. 4, 1261-1295

Zakariya, Y.A. and Khairy, B.R. (2010). Re-sampling in Linear Regression Model using Jackknife and Bootstrap. Iraqi Journal of Statistical Science (18) 2010