Jackknife After Bootstrap Procedure as a Remedy of Outliers in Regression Model

Zakariya Yahya Algamal[1], Intisar Ibrahim Allyas[2]
[1]*Department of Statistics and Informatics, University of Mosul, Mosul, Iraq*
[2]*College of Administration and Economics, Nawroz University, Kurdistan region, Iraq*

Abstract: *Robust regression is an alternative to the least squares method that can be appropriately used when there is evidence that the distribution of the error term is non-normal (heavy-tailed) and/or there are outliers that affect the regression equation.* ***.*** *The least squares method has been in use in regression analysis mainly because of tradition and ease of computation, but this method may suffer a huge setback in the presence of unusual observation such as outliers and high leverage point. In this paper our main objective was to use jackknife after bootstrap procedure in most of robust regression method like, M-estimator and MM-estimator. Analytical examples are presented to show the effective of the deleted observation on the coefficients, and the behavior of jackknife after bootstrap in robust regression.*

Keywords: *Robust Regression, Outlier, Leverage point, Bootstrap, Jackknife after Bootstrap.*

I. Introduction

In multiple regression, ordinary least squares (OLS) estimation is used if assumptions are met to obtain regression weights when analyzing data, the assumptions of OLS are that residual errors should be normally distributed, have equal variance at all levels of the explanatory variables, and be uncorrelated with both the independent variables and each other (Yan &Su ,2009). In practice, the assumption of that residual errors should be normally distributed may not hold because of possibility of skewness or presence of outliers in data. In theory, when this assumption does not meet, the OLS estimation for the regression coefficients β will be biased and / or non-efficient.

In regression analysis, three types of outliers influence the OLS estimator. Rousseeuwuw and Leroy (1987) define them as (1) vertical outliers, which are those observations that have outlying values for the corresponding error term (the y-dimension) but are not outlying in the space of explanatory variables (the x-dimension) and their presence affects the OLS estimator and in particular the estimated intercept, (2) good leverage points are observation that are outlying in the space of explanatory variables but that are located close to the regression line, their presence does not affect the OLS estimator but it affects statistical inference since they do inflate the estimated standard errors. Finally, (3) bad leverage points are observations that are both outlying in the space of explanatory variables and located for from the true regression line, their presence affects significantly the OLS estimator of both the intercept and the slop.

Robust regression is designed to reduce or bound the influence of outliers, it should perform better than OLS method when there are outliers or the residuals have a non-normal distribution with many extreme residuals in the tails (Huber & Ronchetti, 2009). Bootstrapping is a way of testing the reliability of the data set. It allows one to assess whether the distribution of characters has been influenced by stochastic effects (Bancayrin, 2009). In this paper we use Jackknife after Bootstrap to assess the bootstrap procedure in robust regression using M-estimator and MM-estimator. In section 2, we briefly discuss the robust regression. The Jackknife after Bootstrap procedure has been discussed in section 3. in section 4, we present two numerical examples to show the identification procedure. Finally, a discussion was presented in section 5.

II. Robust Regression

Robust regression is an alternative to OLS that can be appropriately used when there is evidence that the distribution of the error term is non-normal (heavy-tailed) and/or there are outliers that affect the regression equation (Ryan, 1993). A least squares method weights each observation equally in getting parameter estimates, whereas robust methods enable the observations to be weighted unequally (Draper & Smith, 1998). In matrix notation, the linear regression model is given by:

$$y = X\beta + e \qquad \text{...}(1)$$

Where, for a sample of size n, y is the ($n \times 1$) vector containing the values for the response variable, X is the ($n \times p$) matrix containing the values for the P explanatory variables, and e is the ($n \times 1$) vector containing the error terms. The ($p \times 1$) vector β contains the unknown regression parameters. The vector of parameters estimated by OLS is then:

$$\hat{\beta}_{OLS} = \arg\min_{\hat{\beta}} \sum_{i=1}^{n} e_i^2(\beta) \qquad \text{...}(2)$$ where

$e_i(\beta)$) is the residuals

$$e_i(\beta) = y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_p X_{ip} \quad \text{for} \quad i = 1,2,3,\ldots,n$$

Numerous robust regression methods have been developed. Among of them are:

(2-1) M-estimators

First proposed by Huber (1964), which are based on the idea of replacing the squared residuals $e_i^2(\beta)$, with another function of the residuals where this function allows to increase normal efficiency while keeping robustness with respect to vertical outliers (Huber & Ronchetti,2009). An M-estimator is defined as

$$\hat{\beta}_M = \arg\min_{\hat{\beta}} \sum_{i=1}^{n} \rho\left(\frac{e_i(\beta)}{\sigma_e}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

where ρ is a symmetric function. M-estimates are calculated using iteratively reweighted least squares (IRLS) method. Taking the derivation of equation (3) and solving produces the score function

$$\sum_{i=1}^{n} \psi\left(\frac{e_i(\beta)}{\hat{\sigma}_e}\right) X_i = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

Two well-known objective function ρ are most widely used, they are, the Huber function and bisquare function.

(2-2) MM-estimators

Yohai (1987) introduced MM-estimators that combine high-breakdown and a high efficiency. These defined estimators are redescending M-estimators as in (3), but where the scale is fixed at S-estimators so an MM-estimator is defined as:

$$\hat{\beta}_{MM} = \arg\min_{\hat{\beta}} \sum_{i=1}^{n} \rho\left(\frac{e_i(\beta)}{\hat{\sigma}_s}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

Where $\hat{\sigma}_s$ is the robust estimator of the scale (Rousseeuw & Leroy ,1987).

III. Jackknife After Bootstrap Procedure

The use of the bootstrap and jackknife resampling methods is gradually increasing nowadays, due to increasing computer power. The basic idea of bootstrapping is to generate a large number of samples by randomly drawing observations with replacement from the original dataset, and to recalculated the estimates for each of these bootstrap samples, whereas the jackknife is generated by sequentially deleting single datum from the original sample (Efron & Tibshirani,1993). Bootstrapping robust regression had been studied by Rarrera and et al.(2002), Willems and Aelst (2005) applied bootstrap on the LTS estimator, Barrera (2006) used bootstrap on the MM-estimator with fixed explanatory variables. (Uraibi & et al.,2009) applied model selection bossing on bootstrapping LTS estimator with fixed explanatory variables. Jackknife-after-Bootstrap (JAB) method was proposed by Efron (1993) to investigate the effect of a single observation in bootstrap. Suppose we have drawn B bootstrap samples and calculate the standard error of the regression parameter $Se_B(\hat{\beta})$, we would like to have a measure of the uncertainty in $Se_B(\hat{\beta})$. The JAB method provides a way of estimating the standard error of $Se_B(\hat{\beta})$, $Se_{JAB}(Se_B(\hat{\beta}))$, using only information in our B bootstrap samples. The jackknife estimate of standard error of $Se_B(\hat{\beta})$ involves two steps:

1-For i=1,2,….n , leave out data point I and recomputed $Se_B(\hat{\beta})$ and called the result $Se_{B(i)}(\hat{\beta})$.

2-Define:

$$Se_{Jak}(Se_B(\hat{\beta})) = \left[\frac{n-1}{n}\sum_{i=1}^{n}(Se_{B(i)}(\hat{\beta}) - Se_{B(.)}(\hat{\beta}))\right]^{\frac{1}{2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

Where $Se_{B(.)}(\hat{\beta}) = \dfrac{\sum_{i=1}^{n} Se_{B(i)}(\hat{\beta})}{n}$ (Efron & Tibshirani,1993).

For each i, there are some bootstrap samples in which that the data point say x_i, dose not appear, and we can use those samples to estimate $Se_{B(i)}(\hat{\beta})$. Let C_i denote the indices of the bootstrap samples that don't contain data point x_i, and there are B_i such samples, then

$$Se_{B(i)}(\hat{\beta})) = \left[\frac{\sum_{B \in C_i} (\hat{\beta}_B - \overline{\hat{\beta}}_B)^2}{B(i)} \right]^{\frac{1}{2}} \quad ....(7)$$ where

$\overline{\hat{\beta}}_B = \dfrac{\sum_{B \in C_i} \hat{\beta}_B}{B(i)}$. The JAB estimate of standard error of $Se_B(\hat{\beta})$ is

$$Se_{JAB}(Se_B(\hat{\beta})) = [\frac{n-1}{n} \sum_{i=1}^{n} (Se_{B(i)}(\hat{\beta}) - Se_{B(.)}(\hat{\beta}))]^{\frac{1}{2}} \quad ....(8)$$ Where

$Se_{B(.)}(\hat{\beta}) = \dfrac{\sum_{i=1}^{n} Se_{B(i)}(\hat{\beta})}{n}$ (Efron & Tibshirani, 1993).

IV. Analytical Examples

In this section, we consider two data sets for assessing the performance of the JAB in the M-estimator and MM-estimator. R program is used to obtain the results.

(4-1) First word–Gesell Adaptive score data

First word–Gesell adaptive score data (Rousseeuw & Leroy, 1987) consist of two variables. The explanatory variable is the age (in months) at which a child utters its first word, and the response variable is its Gesell adaptive score. These data originally for 21 children. But here we delete the 18th observation.
The M-estimator, bootstrap estimator (B=10000), bias and standard error appear in table 1.

Table 1: M-estimator, bootstrap estimator, bias and standard error for firs word data

	M-estimator	bootstrap estimator M-estimator	bias	Std M	Std(BM)
β_0	107.025	109.2352	-2.2102	6.93	8.88
age(β_1)	-0.9405	-1.1611	0.2211	0.5	0.699

The JAB procedure displays a diagnostic for using bootstrap. In R program we use the Jackknife after Bootstrap function which displays a plot. This plot shows the sensitivity of statistic (Here the statistic is $\hat{\beta}_j, \ j = 1,2,...,k$) and of the percentiles of its bootstrapped distribution to deletion of individual observations.

The horizontal axis of the graph, labeled "standardized Jackknife value" is a measure of the influence of each observation on the coefficient. The observation indices corresponding to the points in the graph are shown near the bottom of the plot. The horizontal broken on the plot are quantiles of the bootstrap distribution of each statistic, centered at the value of the coefficient for the original sample. The points connected by solid lines show the quartiles estimator only from bootstrap samples in which each observation in turn did not appear.

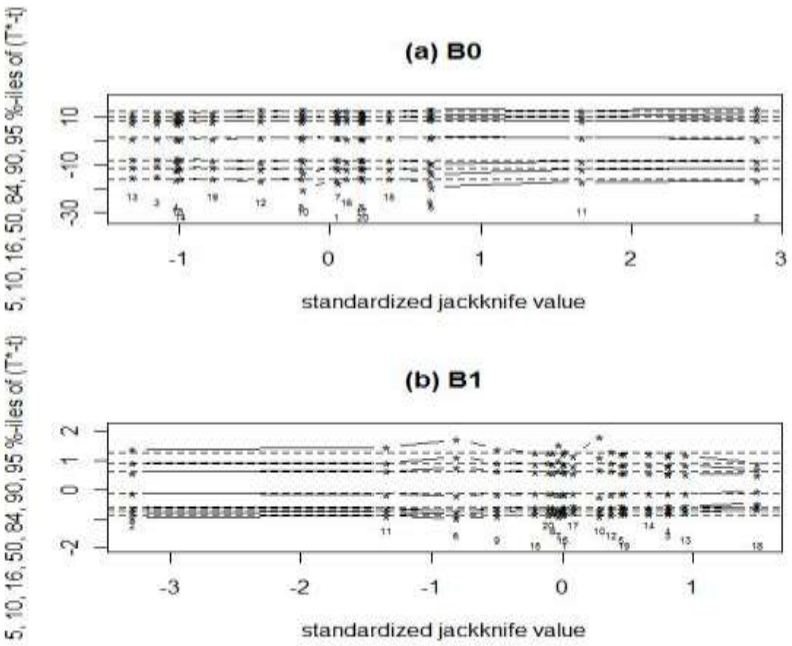

Figure 1: JAB plot for the First word-Gesell adaptive score data coefficients.

(4-2) Stack Loss Data

This data describe be the operation of a plant for the oxidation of ammonia to nitric acid and consist of 21 observations with three explanatory variables. The stack loss (y) has to be explained by the rate of operation (x_1), the cabling water inlet temperature (x_2) and the acid concentration (x_3) (Rousseeuwuw & Leroy,1987) the MM-estimator, bootstrap estimator (B=10.000), bias and standard error are showed in table 2.

Table 2: MM-estimator, bootstrap estimator, bias and standard error for stack loss data

	MM-estimator	Bootstrap estimator MM-estimator	bias	Std M	Std(BM)
β_0	-41.5231	-40.265	1.2586	9.307	8.79
Air (β_1)	0.938	0.889	-0.049	0.105	0.188
Temp .$(\beta 2)$	0.5795	0.6768	0.0973	0.287	0.526
Acid (β_3)	-0.1129	-0.1169	-0.004	0.122	0.122

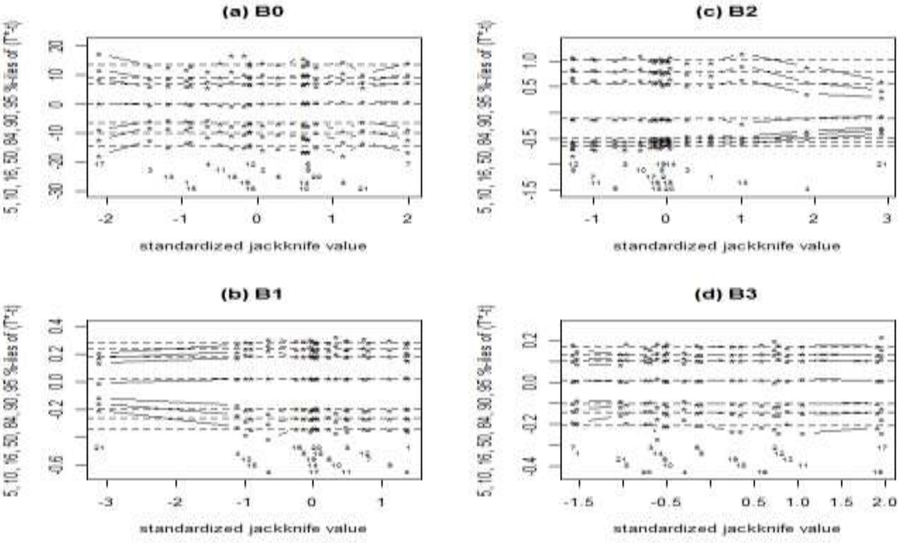

Figure 2: shows the JAB plot for the $\hat{\beta}_0$ and $\hat{\beta}_j, j = 1,2,3$ for the stack loss data.

V. Conclnsion

The analytical examples show that JAB procedure is proved to be very effective in the identification of the effective of the deleted observation. From Figure 1, we notice that no observation has decrease or increase the ($\hat{\beta}_0$) coefficient, whereas observation 6 and 10 serve to decrease the age coefficient. From Figure 2, observation 17 serve to increase the ($\hat{\beta}_0$) coefficient, while observations 8 and 7 have an increase effective on the ($\hat{\beta}_0$) coefficient. Observations 13, 4 and 21 serve to decrease the temperature ($\hat{\beta}_2$) coefficient. Observation 21 has decrease effective on the Air ($\hat{\beta}_1$) coefficient. No observation has an effective on the Acid ($\hat{\beta}_3$) coefficients.

References

[1] Bancayrin,C.,(2009)."**Performance of Median and Least squares Regression for slightly skewed Data**", world Academy of science, Engineering and Technology Vol.53, PP.226-230.

[2] Draper, N. R. and Smith, l.,(1998),"**Applied Regression Analysis**", 3rd ed., John Wiley & Sons, Inc., Canada.

[3] Efron, B. and Tibshirni, R.,(1993),"**An introduction to the bootstrap**", Chapman & Hall Inc. USA.

[4] Efron, B.,(1992),"**Jackknife-After-Bootstrap standard Errors and Influence Functions**", Journal of the Royal statistical. Series B, Vol. 54, No. 1, PP.83-127.

[5] Huber, P. J. and Ronchetti, E. M.,(2009),"**Robust statistics**" 2nd ,John Wiley & Sons, Inc., New Jersey.

[6] Huber, P. J.,(1964),"**Robust estimation of a location parameter**", the Annals of Mathematical statistics Vol. 36, PP.1753-1758

[7] Rousseeuw, A. J. & Leroy, A. M.,(1987),"**Robust regression and outlier detection** ", John Wiley & Sons USA.

[8] Salibian-Barrera, M. and Zamar, R., H.,(2002),"**Bootstrapping Robust Estimates of Regression**", the Annnals of statistics, Vol. 30. No. 2, PP.556-582.

[9] Salibian-Barrera, M.,(2006),"**Bootstrapping MM-estimators for linear regression with fixed designs**", statistics and probability Letters, Vol.76, PP.1287-1297.

[10] Vraibi,H.,S., Midi,H., Talib, B., A., and Yousif, J., M.,(2009),"**Linear Regression Model Selection Based on Robust Bootstrapping Technique**", American Journal of Applied sciences Vol.6, No.6, PP.1191-1198.

[11] Willems, G. and Aelst., S.,(2005),"**Fast and Roust Bootstrap for LTS**", computational statistics and data analysis, Vol. 48, No. 4, PP.703-715.

[12] Yan, X. and Su, X. G.,(2009),"**Linear regression analysis theory and computing**", world scientific publishing Co. , UK.

[13] Yohai, V.,(1987),"**High Breackdown-point and High Efficiency Estimates for Regression** ", the Annals of statistics Vol.15, PP.642-665.

[14] Al-Fakih, A. M., Algamal, Z. Y., Lee, M. H., Abdallah, H. H., Maarof, H., & Aziz, M. (2016). Quantitative structure-activity relationship model for prediction study of corrosion inhibition efficiency using two-stage sparse multiple linear regression. Journal of Chemometrics, 30(7), 361-368.

[15] Al-Fakih, A. M., Aziz, M., Abdallah, H. H., Algamal, Z. Y., Lee, M. H., & Maarof, H. (2015). High dimensional QSAR study of mild steel corrosion inhibition in acidic medium by furan derivatives. International Journal of Electrochemical Science, 10, 3568-3583.

[16] Algamal, Z. Y. (2008). Exponentiated exponential distribution as a failure time distribution. IRAQI Journal of Statistical science, 14, 63-75.

[17] Algamal, Z. Y. (2011). Paired Bootstrapping procedure in Gamma Regression Model using R. Journal of Basrah Researches, 37(4), 201-211.

[18] Algamal, Z. Y. (2012). Diagnostic in poisson regression models. Electronic Journal of Applied Statistical Analysis, 5(2), 178-186.

[19] Algamal, Z. Y. (2017). Using maximum likelihood ratio test to discriminate between the inverse Gaussian and gamma distributions. International Journal of Statistical Distributions, 1(1), 27-32.

[20] Algamal, Z. Y., & Ali, H. T. M. (2017). An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression. Electronic Journal of Applied Statistical Analysis, 10(1), 242-256.

[21] Algamal, Z. Y., & Ali, H. T. M. (2017). Bootstrapping pseudo - R2 measures for binary response variable model. Biomedical Statistics and Informatics, 2(3), 107-110.

[22] Algamal, Z. Y., & Lee, M. H. (2015). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. Expert Systems with Applications, 42(23), 9326-9332.

[23] Algamal, Z. Y., & Lee, M. H. (2015). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. Computers in Biology and Medicine, 67, 136-145.

[24] Algamal, Z. Y., & Lee, M. H. (2015). Penalized Poisson regression model using adaptive modified elastic net penalty. Electronic Journal of Applied Statistical Analysis, 8(2), 236-245.

[25] Algamal, Z. Y., & Lee, M. H. (2015). High dimensional logistic regression model using adjusted elastic net penalty. Pakistan Journal of Statistics and Operation Research, 11(4), 667-676.

[26] Algamal, Z. Y., & Lee, M. H. (2015). Adjusted adaptive lasso in high-dimensional Poisson regression model. Modern Applied Science, 9(4), 170-176.

[27] Algamal, Z. Y., & Lee, M. H. (2015). Applying penalized binary logistic regression with correlation based elastic net for variables selection. Journal of Modern Applied Statistical Methods, 14(1), 168-179.

[28] Algamal, Z. Y., & Lee, M. H. (2017). A new adaptive L1-norm for optimal descriptor selection of high-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives. SAR and QSAR in Environmental Research, 28(1), 75-90.

[29] Algamal, Z. Y., Lee, M. H., & Al-Fakih, A. M. (2016). High-dimensional quantitative structure-activity relationship modeling of influenza neuraminidase a/PR/8/34 (H1N1) inhibitors based on a two-stage adaptive penalized rank regression. Journal of Chemometrics, 30(2), 50-57.

[30] Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., & Aziz, M. (2015). High-dimensional QSAR prediction of anticancer potency of imidazo[4,5-b]pyridine derivatives using adjusted adaptive LASSO. Journal of Chemometrics, 29(10), 547-556.

[31] Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., & Aziz, M. (2016). High-dimensional QSAR modelling using penalized linear regression model with L1/2-norm. SAR and QSAR in Environmental Research, 27(9), 703-719.

[32] Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., & Aziz, M. (2017). High-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives based on the sparse logistic regression model with a bridge penalty. Journal of Chemometrics (doi:10.1002/cem.2889).

[33] Algamal, Z. Y., Qasim, M. K., & Ali, H. T. M. (2017). A QSAR classification model for neuraminidase inhibitors of influenza A viruses (H1N1) based on weighted penalized support vector machine. SAR and QSAR in Environmental Research, 1-12.

[34] Kahya, M. A., Al-Hayani, W., & Algamal, Z. Y. (2017). Classification of breast cancer histopathology images based on adaptive sparse support vector machine. Journal of Applied Mathematics & Bioinformatics, 7(1), 49-69.