HYPOTHESIS TESTING FOR OUTLIER EFFECT ON THE REGRESSION COEFFICIENTS USING F TEST STATISTIC

*Baba, I. A.[1], Mohammad, A. A.[2], and Arzuka, I[3]

[1]Department of Mathematical Sciences, Faculty of Science, Taraba State University, Jalingo
[2-3]Department of Mathematic, Bauchi State University Gadau
*Email: ishaqbaba@yahoo.com

ABSTRACT

Outlying observation may lead to misleading inference plus biased estimate of the parameter and model misspecification among others. It is therefore important to highlight on it verse negative effect in the analysis of datasets. This paper, considered the effect of outliers on ordinary least squares (OLS) based on the test of hypothesis procedure, to test if there is any significant difference in testing the hypothesis of regression parameter for dataset with and without outliers using F test statistic. Four different dataset were considered to proof the set hypothesis. Our results showed that in the presence of outliers, the estimates of regression parameter sign may change which in turn may lead to wrong decision marking.

Keywords: *Outlier, Regression, Hypothesis and F test Statistics*

INTRODUCTION

Regression analysis modeled relationship between a response variable and a set of predictor variables, plus the random error term. The model in matrix notation assumes a functional relation of the form

$$Y = X\theta + \epsilon,$$

where Y represent the $n \times 1$ column vector of the observed response variable, X represent the data matrix of $n \times p'$ consisting of a column of ones and the p column vectors of the observations on the predictor variables, θ represent the $p' \times 1$ vector of parameters to be obtained by regressing y on x variable and ϵ represent the $n \times 1$ vector of random error which assumed to be normally distributed with mean 0 and variance σ^2. The regression model also assumed that there is no perfect Multicollinearity and all observations are equally reliable that is, they contain no contamination in the observed data point. Outlier may arise as a result of wrong entry, error in incorrect selection of model and/ or due to uncontrolled factors. For example in the study of HIV cases in the hospital, the practitioner found few individuals out of total population say 100, living with HIV but alive fine for certain years without taking any treatment, those individuals having this virus may be refer to as outlier, because they possess a contrary behaviors compared to the majority of the population under investigation. The presence of outlier in the data affects the scale estimates and deteriorates the effectiveness to identify both masking and swamping effect Cheng (2012). Generally speaking, there are two well-known techniques for estimation of regression parameter, the ordinary least square (OLS) and the Ordinary least squares is a method for estimating the unknown parameters in a linear regression model with the goal of minimizing the differences between the observed responses in some arbitrary data set and the responses produced by linear approximation of the data. This method obtains parameter estimates that minimize the sum of squared residuals, SSE. By using the formula

$$\min_{\theta} \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left(y_i - X\hat{\theta}\right)^2, \quad \text{where}$$

$$X\hat{\theta} = x_{i0}\hat{\theta}_o + \ldots + x_{ip}\hat{\theta}_p = \hat{y}$$

$$r_i = y_i - \hat{y}_i$$

denotes the difference between the observed response and the predicted values and n is total number of the observed data point. In particular, this approach is known to be non-resistance to outlier in the sense that it gives misleading estimates of regression parameter, which in turn affect all other inference of the regression model, among them are t values, p values, coefficient of determination and even the regression anova computations. Thus, outlier in regression analysis are categorized in three forms, outlier in y_i direction, outlier in x-axis which is also called leverage point and outlier in both y and x axis. Dodge (1997) deliberated on the effect of outlier both in the response and predictor variables and show that the least absolute deviation is more sophisticated in identifying outlying data point compared to OLS procedures. Osbone and Overbay (2004) argued on if outlier should be removed or not. They further explain the benefit of removing it and highlighted on it negative consequences in the estimation of parameter in parametric and non-parametric statistics. Recently, the outlier resistant test for heteroscedasticity in linear model based on the GoldfeldQuandt formula was proposed by Alih and Ong (2015), there proposed scheme uses two steps, the first step involved identification of outlier while in the second step, estimation of the proposed robust Goldfeld Quant procedure is performed and the method give a promising result compared to the existing approaches like S. Rana et al (2010) and Goldfeld and Quant (1965). But all these technique can only be applied to a single predictor problem. Usman and Oyejola (2013) studied the effect of outlier and excess zero on count data and established that presence of outliers in count data causes over dispersion. Cheng (2012) considered both identification outliers and heteroscedasticity without specific form in a dataset. Outlier has been deliberated by quite a lot of authors (Wainer (1976), Zimmerman(1994), Leroy and Rousseeuw (1987), Iglewicz and Hoagli (1993) and Drapper and John (1981)). Consequently, continuous research in this area has shown the need for researchers to explore more techniques of dealing with outliers in statistical study.

ORDINARY LEAST SQUARE ESTIMATOR

The commonly used method for solving equation is called the ordinary least squares (OLS), the standard formula for computing the estimates of regression coefficient is define by

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ M \\ y_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \Lambda & x_{1n} \\ 1 & x_{21} & x_{22} & \Lambda & x_{2n} \\ M & M & M & & M \\ 1 & x_{n1} & x_{n2} & \Lambda & x_{nn} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ M \\ \beta_n \end{bmatrix}$$

following the successful estimate of the regression coefficients in linear regression, test of hypothesis is usually carried out to the significance of each regression coefficient and or significance of the whole model. The test can be applied provided that it can be assumed that the random error terms are normally and identically distributed with mean zero and variance σ^2. There are popularly two types of hypothesis that can be used in multiple linear regression analysis (t and F test statistic). The t test checks the significance of the individual regression coefficients while F test is used to simultaneously check the significance of a number of regression coefficients. It can also be used test individual coefficients.

Procedure for test of Significance of Regression Coefficients using F test Statistic

The following steps are used for the implementation of F test statistic:

Step 1: Set the hypothesis that $H_0 : \beta_0 = \beta_1 = \beta_2 = \Lambda = \beta_p$ versus $H_1 : \beta_j \neq 0$ for at least one j

Step2: Compute the F test statistic as $F_0 = \frac{MS_R}{MS_E}$, where MS_R is the regression mean square and MS_E is the error mean square. If the null hypothesis, H_0 is true then the statistic F_0 follows the F distribution with k degree of freedom in numerator and $n-(k+1)$ degree of freedom in the denominator.

Step3: The hypothesis, H_0 is rejected if the calculated statistic, F_0 is such that $F_0 > F_{\alpha,k,n-(k+1)}$

Procedure for test of Significance Using the t test Statistic

The t statistic is used to check the significance of the coefficients in multiple linear regression models. It is a well-known fact that adding a significant variable to regression model makes the model more efficient and effective, whereas adding an insignificant variable makes the model worse. The hypothesis statement to test the significance of a particular regression coefficient β_j is employing using the following steps:

Step1: set the hypothesis $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$

Step2: Compute the t statistic $T_0 = \frac{\hat{\beta}_j}{Se(\hat{\beta}_j)}$, where $Se(\hat{\beta}_j)$ is the standard error

Step3: the null hypothesis, H_0 is rejected if the statistic $-t_{\alpha/2,n-2} < T_0 < t_{\alpha/2,n-2}$ lies within the acceptance region

PROPOSED ALGORITHM

Step1: $H_0^{'}$: Outliers does not affect the estimate of regression coefficients against $H_1^{'}$: Outliers affect the estimate of regression coefficients

Step2: Regress y against x using data without outliers by applying the OLS to obtain the estimate of regression coefficients.

Step2a: Set the hypothesis that $H_0 : \beta_0 = \beta_1 = \beta_2 = \Lambda = \beta_p = 0$ versus $H_1 : \beta_j \neq 0$ for at least one j, to test for the significance of the regression coefficient in the absence of the outliers

Step2b: Compute the F test statistic $F_0 = \frac{MS_R}{MS_E}$,

Step2c: The hypothesis, H_0 is rejected if the calculated statistic, F_0 is such that $F_0 > F_{\alpha,k,n-(k+1)}$

Step3: Regress y on x using the data with outliers by applying the OLS to obtain the estimates of regression coefficients.

Step4: Repeat step (2a) – (2c)

Step5: Compare the results of (2c) with step4, if the results of (2c) and step4 are equal, then $H_0^{'}$ should be accepted otherwise, $H_1^{'}$ should be accepted.

NUMERICAL EXAMPLES

In this section, four real life examples from the existing literature, which have been used by the researchers extensively to study the behavior of outlying observations, are presented. In the first, second and third examples, simple linear regression model are used to obtain estimates of the regression coefficients both without and with outliers. In the fourth example, we considered the used of multiple linear regression model in the estimation of the regression coefficients. Contamination is planted only in the predictor observations and the proposed algorithm is applied to all the said data set to verify the

effectiveness of the established hypothesis. The numerical computations are performed via the R programing language, which is free, open source statistical software for graphics and computing developed by Ihaka and Gentleman (1996).

Data for the Hertzsprung-Russell Diagram of the star cluster CYG OB1

These data set as deliberated in Leroy and Rousseeuw(1987), contains 47 observation which represent the stars in the direction of Czgnus, from C.Doom. The first variable is the logarithm of the effective temperature at the surface of the star (Te) and the second one is the logarithm of its light intensity (L/L0). Detection of outlier in both the horizontal (predictor) and vertical (response) direction were considered by Cook and Weisberg (1994) using this data set. In this example, we considered contaminating only the predictor observations. Applying the proposed algorithm to the data set under study before and after the outliers was inserted, using the lm function, which is used to fit a linear model in R software, we observed the result in Table 1.

Table 1: F test for Hertzsprung - Russell diagram data

Test procedure	F test without outliers	F test with outliers
F_{Cal}	2.0800^{r}	3.1242^{a}
F_{Tab}	2.8205^{a}	2.8205^{r}

The results given in Table 1 shows that the F test statistics fail to reject H_0 when the data were free from outliers. However, when applied to the dataset with outliers, F test statistic rejects the null hypothesis H_0 and accepts H_1, since the results of (2c) and Step 4 are not equal $H_0^{'}$ is rejected and concludes that outliers affect estimate of regression coefficients.

House Expenditure and Saving Data

These data as discussed in Pindyck and Rubinfeld (1998), Rana and Midi (2008), and Alin and Ong (2015), comprises of 20 observations which describe the relationship between the income as predictor variable and housing expenditure as the response variable. Inserting only two outliers in the dataset alter the F value which is the bench mark for making decision. As in the first example, implementation of the developed algorithm yields the result in the Table 2 and 3. Regarding the test power of the existing procedure and newly established algorithm, the new algorithm outperforms the old method in the presence of contamination as given in Table 2 and 3respectively.

Table 2: F test for House expenditure data

Test procedure	F test without outlier	F test with outliers
F_{Cal}	256.2700^{a}	0.1340^{r}
F_{Tab}	3.0070^{r}	3.0070^{a}

Table 3: Saving data

Test procedure	F test without outliers	F test with outliers
F_{Cal}	300.7320^{a}	0.1600^{r}
F_{Tab}	2.8870^{r}	2.8870^{a}

SIMULATION STUDY

In this section, we carry out simulation to investigate the performance of both existing and newly established F test procedure via the power test analysis. Both Simple and multiple linear regression models were considered where the data set were generated in this way: four different sample sizes were considered (n =15, n= 25, n=50 and n= 100) for both the two cases. We simulated the data from the normal distribution with different proportion of outlier (0, 5, 10, and 15) for each sample sizes. We vary the proportion of the outlier as to have prior information about contaminated and uncontaminated case. The response observations were generated using the next equation:

$$y_i = 0.5 + 2x_i + e_i, e_i \sim (n,25,2)$$

While the predictor observations were generated from $x_i \sim (n,90,5)$ for $i = 1,\mathrm{K}\ n$ the number of observations for each sample sizes with corresponding proportion of outlier as given in Table 4, for the multiple regression case we generated the response as:

$$y_i = 0.5 + 2x_i + 6x_i + e_i, e_i \sim (n,25,2)$$

And the corresponding predictors were generated as $x_{1i} \sim (n,90,5), x_{2i} \sim (n,75,10)\ i = 1,\mathrm{K}\ n$ for each set of data at different proportion of contamination. These experiments were repeated 1000 times in each case and the results are summaries in Table 4. The results of Table 4, show that the performance the newly developed algorithm is consistent with increase in the number of predictor since there is no significant difference between the power of the for simple and multiple linear regression model. Although when the sample size (n =15) at 10% and 15% the proposed algorithm gives less efficient results compare to when the sample sizes are large (n=25,50 and 50) at 10% and 15% contamination respectively. We can therefore conclude that the newly proposed algorithm for testing the significant of the regression is more powerful in the presence of outliers.

Table 4: Power test for 1000 iterations

n	Outlier proportion	$(1-\beta_{SLR})$	$(1-\beta_{MLR})$
15	0	0.000	0.000
	5	0.904	0.932
10		0.906	0.929
15		0.901	0.934
25	0	0.000	0.000
	5	0.916	0.921
	10	0.921	0.940
	15	0.922	0.945
50	0	0.000	0.000
	5	0.880	0.777
	10	0.903	0.932
	15	0.904	0.945
100	0	0.000	0.000
	5	0.783	0.063
	10	0.895	0.822
	15	0.912	0.911

CONCLUSION

A two-step F test statistic for testing the significant of regression coefficient were proposed and presented. The method is resistant to outlier for both simple and multiple linear regression cases when the sample sizes are fairly large. Our simulation study shows that the established algorithm is consistent and powerful in testing the significance of regression when the data is contaminated. Hence, it is therefore necessary for researchers especially the end user of statistical tools to check the presence of outlier before conducting the F statistic test of significant otherwise it may lead to wrong estimation and misleading conclusion, which may cause more harm than good to decision makers and practioners.

REFERENCES

Cheng, T. C. (2012). On simultaneously identifying outliers and heteroscedasticity without specific form. *Computational Statistics and Data Analysis*, 56(7), 2258-2272.

Dodge, Y. (1997). LAD regression for detecting outliers in response and explanatory variables. *Journal of multivariate analysis*, *61*(1), 144-158.

Osborne, J. W., and Overbay, A. (2004).The power of outliers (and why researchers should always check for them). *Practical assessment, research and evaluation*, *9*(6), 1-12.

Alih, E., and Ong, H. C. (2015).An outlier-resistant test for heteroscedasticity in linear models. *Journal of Applied Statistics*, *42*(8), 1617-1634.

Rana, M. S., Midi, H., and Imon, A. R. (2008).A robust modification of the goldfeld-quandt test for the detection of heteroscedasticity in the presence of outliers. *Journal of mathematics and Statistics*, *4*(4), 277.

Goldfeld, S. M., and Quandt, R. E. (1965). Some tests for homoscedasticity. *Journal of the American statistical Association*, *60*(310), 539-547.

Usman, M., and Oyejola, B. A. (2013).Models for Count Data in the Presence of Outliers and/or Excess Zero.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make nonevermind. *Psychological Bulletin*, *83*(2), 213.

Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric tests. *The journal of general psychology*, *121*(4), 391-401.

Leroy, A. M., and Rousseeuw, P. J. (1987). Robust regression and outlier detection. *Wiley Series in Probability and Mathematical Statistics. New York: Wiley*, *1987*.

Iglewicz, B., and Hoaglin, D. C. (1993). *How to detect and handle outliers* (Vol. 16). Asq Press.

Draper, N. R., and John, J. A. (1981).Influential observations and outliers in regression. *Technometrics*, *23*(1), 21-26.

Ihaka, R., and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, *5*(3), 299-314.

Leroy, A. M., and Rousseeuw, P. J. (1987).Robust regression and outlier detection. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley*, *1987*.

Cook, R. D., and Weisberg, S. (2009). *An introduction to regression graphics* (Vol. 405). John Wiley and Sons.

Pindyck, R. S., and Rubinfeld, D. L. (1998). *Econometric models and economic forecasts* (Vol. 4). Boston: Irwin/McGraw-Hill.

Rana, M. S., Midi, H., and Imon, A. R. (2008).A robust modification of the goldfeld-quandt test for the detection of heteroscedasticity in the presence of outliers. *Journal of mathematics and Statistics*, *4*(4), 277.