OUTLIER EFFECT IN REGRESSION INFERENCE BASED ON T TEST STATISTIC USING STANDARD DEVIATION METHOD

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

[1]Department of Mathematical Sciences, Faculty of Science, Taraba State University Jalingo, Nigeria
[2-3]Department of Mathematic, Faculty of Science, Bauchi State University Gadau, Nigeria

ABSTRACT

One of the essential requirements for smooth regression analysis based on ordinary least squares (OLS) method is normality of the data. When the dataset pass the normality test, it is virtually free from unusual observations, (OLS) method can then be applied effectively to obtain the required estimate of the regression parameters and make inference. It is quite obvious that presence of outlier distorts regression inference when non robust methods are used. This article highlighted on the consequences of outlier in the analysis of regression models based on the t test statistics for testing significant of regression coefficient, since present of outlier could disturb significant test which in turn may lead to misinterpretation of the final result. Standard deviation (SD) method were employed in measuring the effect of outlier on t test statistics taking into account the sample sizes and the number of regressor (s) at different intensity of outlier both using real examples and simulation study. It was discovered that outlier affect the estimate of regression coefficient negatively which in return render the classical regression estimators inefficient. In addition it can alter the odds of making both type I and type II error as well as influence the estimate of regression that are of essential interest.

Keywords: *Outlier, Regression, Standard Deviation, t Statistic, Smoothing Regression Analysis and normality test*

INTRODUCTION

Consider the general linear regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon_i \qquad (1)$$

Which can also be written in matrix form as $Y = X\beta + \varepsilon$

where Y is an $n \times p + 1$ matrix of the predictor variables (the first column takes values 1), p is the number of regression parameters (including the intercept β_0) and n is the sample size, β is a column vector of the unknown regression parameters, ε is a column vector of unobservable random errors. The solution of equation (1) can be achieved by minimizing the difference between the estimated and observed response which can be obtained using the least square loss function defined by;

$$L(X, Y, \beta) = \sum_{i=1}^{n} (Y - X\beta)^2 \qquad (2)$$

Although several methods can be used to find the estimates of regression coefficient in (2), ordinary least square (OLS) method is the most commonly used in practice due to its simplicity and elegant properties. The (OLS) estimator is defined as β^= ("X" ^T "X")^(-1) "X" ^T "Y" , where X is a matrix of predictors and y is the observed response which often contain some outlier that have unusual small or large data point when compare to the remaining values in a dataset. When a dataset contain some outlier and other uncleanness, the OLS method produces corrupt estimates and may lead to wrong inference (Mielke2016). Outliers arise for several reasons, such as data from the heavy tail distribution functions or erroneous measurements (Kanhere and Khanuja (2014). andSuykens et al (2002)). Outlier can cause a lot of damages in statistical data analysis. According to Osborne and Overbay

(2004), outlier causes increase in variance and reduces the power of statistical test. In addition it can alter the odds of making both type I and type II as well as influence the estimate of regression that are of essential interest. Olewuezi (2011) compared some outlier labeling techniques which include Standard Deviation (SD), the MAD and the Median formulas. Hodge and Jim (2004) conducted a survey on the outlier detection methodologies and identify their advantages and disadvantages in data analysis. For detail on outlier detection method and effect in data analysis see (Hawkins et al., (2002), Dodge (1997), Rousseeuw and Leroy (2005) and Stevens (1984)). There are commonly 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 to test individual coefficients. It can also be used to test individual coefficients. This paper considered Standard Deviation (SD) method in measuring the effect of outlier on the t statistics taking into account the sample sizes and number of regressor at different intensity of outlier.

1.1 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 calculated statistic F_0 is such that $F_0 > F_{\alpha,k,n-(k+1)}$

1.2 Procedure for test of significance us the t test statistic

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

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

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 statistic $-t_{\alpha/2,n-2} < T_0 < t_{\alpha/2,n-2}$ lies within acceptance region

<

Numerical Examples and Simulation Study

In this section two real life example a simulation are discussed. The first dataset w used in Becker (1988) and the second exam dataset in Harrison and Rubinfeld (1978), Bel et al. (1980) and Tsung (2011) respectively study outlier related issues. The simulati procedure is based on 6 different size regress models with one, two, three, four and regressor. Observations of the explanat variables were generated from the unifo distribution. The errors ε_i were taking from standard normal distribution, with mean zero a variance one ($\varepsilon_i \sim N(0, 1)$). The parameters of model in each case were set as $\beta_0 = \beta_i =$ (hence each parameter is significant).

conducted *t*-test on the last parameter in each case. Two outlier levels (5% and 10%) at different sample sizes and different intensities of outliers were considered. Outlier intensity in this case, refers to degree or extent of deviation of the outlying observations from the normal mean. Six outlier intensities in succession at 5SD (5 standard deviation of the mean), 6SD, 7SD, 8SD, 9SD and 10 SD were considered. The results displayed in the table are for power of the *t* test (in percent) and each is based on 10000 replications using R. Power of a test refers to the ability of a test statistic to reject the null hypothesis (H_0) when it is false. In this case, H_0 is always false, the effect of outlier is apparent therefore with non-rejection of H_0 or less power. This means reduction in power is proportional to the outlier effect. One can deduce from the results obtained that outlier negative effect on the inference relies on the following factors:

I. Number of Outliers per Sample

The more the number of outliers relative to the sample size, the less the resistant of the test statistic to the effect of the outliers in the inference. For example, in the single regressor case, at n = 40, 5% outlier at 8SD, 9SD and 10SD, the powers are 86.19, 78.12 and 70.58 respectively; in the same case but at 10% outlier, the powers reduced to 59.1, 50.69 and 43.52 respectively.

II. Degree/ Intensity of the Outlier

The further the deviation of outlying observations from the normal observations, the greater the effect of the outlier, in distorting the inference.

III. Sample Size

The test statistic resists more, the negative effect of outliers in the inference in large samples than in small samples. For example, in two regressor case, at n = 60, 10% outlier at 5SD, 6SD and 7SD, the powers are 84.21, 72.31 and 59.9 respectively; under the same circumstance, but at n = 80, the powers improved to 93.18, 84.07 and 72.97 respectively.

IV. Number of Regressor

The test statistic resists more, the negative effect of outliers in the inference when there is less regressor than more. This means, smaller sized regression models resist outlier in the inference more than larger sized ones. For example, in two regressor case, at n = 60, 5% outlier at 7SD, 8SD and 9SD, the powers are 86.22, 77.82 and 69.64 respectively; under the same circumstance, but with three regressor, the powers reduced to 71.37, 62 and 52.68 respectively.

Two real data sets were considered. In each data, two levels of outliers at approximately 5% and 10% were used in turn. The outliers are inserted in place of original observations to contaminate the data. At each level, we considered 5SD - 10SD outlier intensities.

In the first data (Freeny Data), three of the parameters are significant at 95% (at least); but with both the 5% and 10% outliers, none is significant at each of the outlier intensity. The effect of the outlier is clear in this regard.

In the second real data set (Boston Housing Data), six of the parameters are significant at 95% at least. With 5% outliers in the data, the numbers of significant parameters at 90% level (at least) are 5, 5, 3, 2, 2 and 2 respectively for 5SD, 6SD, 7SD, 8SD, 9SD and 10SD. With 10% outliers in the data, the numbers of significant parameters at 90% level at least are 3, 1, 1, 1, 1 and 1 respectively for 5SD, 6SD, 7SD, 8SD, 9SD and 10SD.

Simulation results for power test of t statistic

Power of test

Number of regressor	Outlier intensity	N=20(5)	N= 20(10)	N=40(5)	n= 40(10)	N=60(5)	n= 60(10)	N=80(5)	n= 80(10)	N=100(5)	n=
Single regressor	5SD	82.87	61.32	99.78	90.74	100.00	98.50	100.00	99.83	100.00	
	6SD	72.29	49.29	97.42	80.34	99.92	94.17	100.00	98.21	100.00	
	7SD	63.59	40.16	92.79	68.65	99.03	86.47	99.91	94.55	100.00	
	8SD	56.04	33.74	86.19	59.10	96.88	77.11	99.41	88.07	99.95	
	9SD	49.30	28.69	78.12	50.69	93.06	67.54	97.90	81.16	99.44	
	10SD	45.50	25.03	70.58	43.52	87.25	58.63	95.26	73.07	98.35	
2regessor	5SD	62.74	39.79	90.36	66.75	98.33	84.21	99.80	93.18	99.97	
	6SD	52.64	31. 26	80.42	53.27	93.88	72.31	98.39	84.07	99.70	
	7SD	45.28	24.70	70.02	43.31	86.22	59.90	94.76	72.97	98.20	
	8SD	37.48	20.89	60.78	35.24	77.82	50.02	89.09	60.72	94.71	
	9SD	34. 11	17.44	51.91	29.55	69.64	41.64	82.28	52.90	89.14	
	10SD	28.28	15.11	44.98	25.35	62.85	35.48	73.21	45.67	82.91	
3regressor	5SD	50.02	28.65	78.20	52.02	92.62	69.31	98.07	80.78	99. 36	
	6SD	40.95	21.92	66.25	39.75	82.85	55.13	91.88	68.24	96.64	
	7SD	34.03	17.76	55.26	31.90	71.37	43.21	83.69	55.89	91.62	
	8SD	28.76	15.02	45.43	25.07	62.00	36.16	74.98	45.66	84.37	
	9SD	24.50	12.77	38.91	21.83	52.68	30.24	65.69	38.87	74.85	
	10SD	20.54	12.33	32.82	18.44	46.96	26.20	56.94	32.20	67.14	
4regressor	5SD	41.22	22.41	66.92	40.66	84.52	57.09	93.43	70.36	97.32	
	6SD	32.42	18.22	53.73	32.11	73.14	44.21	83.92	56.43	91.66	
	7SD	27.22	13.97	44.57	24.02	60.28	34.91	73.12	45.73	82.61	
	8SD	21. 94	12.23	37.41	19.99	50.28	28.37	62.49	36.21	72.94	
	9SD	18.25	10.16	31.37	17.12	43.04	23.89	53.53	30.75	63.28	
	10SD	15.34	9.23	26.14	15.45	36. 37	20.68	47.88	25.24	55.95	
10regressor	5SD	14.54	9.07	47.04	26.04	47.04	26.04	28.67	24.98	69.85	
	6SD	10.94	7.67	36.06	19.63	36.06	19.63	16.01	18.13	56.02	
	7SD	9.02	6.89	28.72	16.58	28.72	16.58	14.80	12.45	45.53	
	8SD	8.08	5.90	23.18	13. 81	23.18	13.81	13.09	11. 65	36.53	
	9SD	7.34	6.12	19.65	11.23	19.65	11.23	8.22	9.08	30.44	
	10SD	6.87	5.67	16.97	10.40	16.97	10.40	11.55	10.11	25.41	

Freeny dataset

	Without outliers	With outlier (5) 5SD	6SD	7SD	8SD	9SD	10SD	With outlier (10) 5SD	6SD	7SD	8SD	9SD
P values	0.091	0.735	0.756	0.773	0.786	0.796	0.805	0.641	0.661	0.675	0.686	0.695
	0. 390	0.796	0.801	0.805	0.808	0.810	0.812	0.901	0.912	0.920	0.927	0.932
	0.000	0.788	0.800	0.810	0.817	0.823	0.828	0.910	0.922	0.930	0.936	0.941
	0.000	0.862	0.846	0.834	0.825	0817	0.811	0.622	0.600	0.585	0.573	0.565
	0.013	0.673	0.673	0.691	0.705	0717	0.727	0.577	0.598	0.614	0.627	0.637

Boston Data set

	Without outliers	With outlier (5)						With outlier (10)					
		5SD	6SD	7SD	8SD	9SD	10SD	5SD	6SD	7SD	8SD	9SD	10SD
P values	0.020	0.333	0.428	0.505	0.577	0.640	0.694	0.129	0.164	0.196	0.2245	0.250	0.273
	0.017	0.945	0.890	0.764	0.668	0.594	0.535	0.717	0.835	0.929	0.996	0.936	0.888
	0.444	0.659	0.598	0.553	0.520	0.494	0.474	0.346	0.307	0.281	0.262	0.249	0.238
	0.526	0.518	0.458	0.416	0.384	0.361	0.343	0.297	0.261	0.238	0.220	0.208	0.199
	0.051	0.442	0.532	0.610	0.678	0.734	0.784	0.303	0.362	0.412	0.455	0.492	0.523
	0.483	0.436	0.451	0.465	0.478	0.489	0.499	0.944	0.981	0.991	0.969	0.951	0.936
	0.935	0.058	0.047	0.040	0.036	0.033	0.031	0.544	0.529	0.520	0.513	0.507	0.503
	0.000	0.046	0.077	0.112	0.149	0.187	0.224	0.083	0.124	0.167	0.209	0.247	0.283
	0.000	0.011	0.039	0.093	0.172	0.266	0.368	0.089	0.200	0.338	0.482	0.618	0.743
	0.464	0.418	0.435	0.452	0.465	0.478	0.489	0.135	0.138	0.141	0.145	0.148	0.151
	0.147	0.759	0.846	0.917	0.975	0.978	0.939	0.754	0.824	0.878	0.921	0.956	0.984
	0.041	0.068	0.084	0.101	0.117	0.131	0.145	0.120	0.147	0.170	0.191	0.209	0.226
	0.096	0.009	0.010	0.011	0.012	0.013	0.014	0.047	0.053	0.059	0.065	0.070	0.075
	0.001	0.974	0.797	0.630	0.509	0.420	0.355	0.862	0.961	0.828	0.726	0.649	0.588

CONCLUSION

In both the simulation and the real life examples different intensity of outlier using the standard deviation (SD) method were used at different sample sizes, number of regressor and percentage of outlier planted in each data set to study the consequences of presence of outlier in analysis of regression (t test statistics). Our result demonstrates categorically that outlier affects the value of regression estimate and power of the test for the hypothesis test as well as other inferences. It was also discovered that the power of the test of significant depends on the percentage of outlier planted, number of regressor, sample sizes and the intensity of outlier used. It's therefore important to check for the presence of outlier before carrying out any statistical data analysis to avoid committing type I and or type II error

REFERENCES

Mielke, A. (2016). Regression and Outliers.

Kanhere, P., and Khanuja, H. K. (2014).A Survey on Outlier Detection in Financial Transactions. *Int. J Computer Applications (0975–8887) Volume*. Suykens, J. A., De Brabanter, J., Lukas, L., &Vandewalle, J. (2002). Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, *48*(1), 85-105.

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

Olewuezi, N. P. (2011). Note on the comparison of some outlier labeling techniques. *Journal of Mathematics and Statistics*, *7*(4), 353-355.

Hodge, V., and Austin, J. (2004).A survey of outlier detection methodologies. *Artificial intelligence review*, *22*(2), 85-126.

Hawkins, S., He, H., Williams, G., and Baxter, R. (2002, September).Outlier detection using replicator neural networks.In *DaWaK* (Vol. 2454, pp. 170-180).

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

Rousseeuw, P. J., and Leroy, A. M. (2005). *Robust regression and outlier detection* (Vol. 589).John wiley& sons.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin*, *95*(2), 334.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.

Harrison, D., and Rubinfeld, D. L. (1978).Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, *5*(1), 81-102.

Belsey, D., Kuh, E., and Welch, R.E., R
Regression Diagnostics: Identif
Influential Data and Sources of Collinea
John Wiley, New York.

Cheng, T. C. (2012). On simultaneo
identifying outliers and heteroscedasti
without specific form. *Computati
Statistics & Data Analysis*, *56*(7), 2258-22