Methods of Detecting Outliers in A Regression Analysis Model.

Ogu, A. I. *, Inyama, S. C+, Achugamonu, P. C++

**Department of Statistics, Imo State University,Owerri*
+Department of Mathematics, Federal University of Technology, Owerri
++Department of Mathematics, Alvan Ikoku Federal College of Education, Owerri

Abstract

This study detects outliers in a univariate and bivariate data by using both Rosner's and Grubb's test in a regression analysis model. The study shows how an observation that causes the least square point estimate of a Regression model to be substantially different from what it would be if the observation were removed from the data set. A Boilers data with dependent variable Y (man-Hour) and four independent variables X_1 (Boiler Capacity), X_2 (Design Pressure), X_3 (Boiler Type), X_4 (Drum Type) were used. The analysis of the Boilers data reviewed an unexpected group of Outliers. The results from the findings showed that an observation can be outlying with respect to its Y (dependent) value or X (independent) value or both values and yet influential to the data set.

Key Words: Outliners, univariate, bivariate data, Regression Analysis,

---

1.0 Brief History and Background of Study

Outliers" are unusual data values that occur almost in all research projects involving data collection. This is especially true in observational studies where data naturally take on very unusual values, even if they come from reliable sources. Although definitions varies. An outlier is generally considered to be a data point that is far outside the norm for a variable or population Jarrell [4], Rasmussen [5]) and Steven [6].

1.1 Causes of Outliers

Outliers can arise from several different mechanisms or causes. Ascombe (1960) sorts into two major categories. Those arising from errors in the data and those arising from the inherent variability of the data.

1.2 Identification Of Outliers.

There is no such thing as a simple test. However, there are many ways to look at a distribution of numerical values, to see if certain points seem out of line with the majority of the data. This can be achieved by:

(I) By visual Aids
(II) By computation of IQR
(III) By plotting a scatter plot.

1.3 Dealing with Outliers

There is a great deal of debates as what to do with identified outliers. If

your data set contains an outlier two questions arises

(1) Are they merely fluke of some kind?

(2) How much have the coefficients error statistics and predictions been affected?

2.0 Data Presentation and Methodology

This study intends to examine the causes, problems, methods of detection and approaches to data analysis of outlier in a univariate and Bivariate data. In order to do this, a Broiler data were collected from Kelly Uscategui, university of Connecticut on Broilers.

2.1 Method of Data Analysis

2.1.1 ROSNER'S TEST (Rosner,1983)

The procedure entails removing from the data set the observation X that is fartherest from the mean. The test statistic R is calculated and compared with the critical value.

The Rosner's R test

$$R_{i+1} = \frac{\left| X^{(i)} - \overline{X}^{(i)} \right|}{S^{(i)}}$$

where

$\overline{X}^{(i)}$ is given by

$$\overline{X}^{(i)} = \frac{1}{n-1} \sum_{j=1}^{n-i} X_j \quad \ldots \quad (1)$$

and

$$S^{(i)} = \left[\frac{1}{n-1} \sum_{j=1}^{n-i} \left(X_j - \overline{X}^{(i)} \right)^2 \right]^{\frac{1}{2}}_{i,0\ldots k} \quad \ldots \quad (2)$$

$L_{i+1} \Rightarrow$ Tabled Critical Value for Comparison with R_{i+1}.

2.1.2 Test Criteria/Decision Rule Hypothesis

H_0: There is no outlier in the data set

H_{AK}: There is at least one outlier in the data set.

Decision Rule

Reject H_0 if $R_{i+1} > L_{i+1}$ at the stated level of significance otherwise do not reject H_0.

2.2 GRUBB'S TEST (Grubb, 1950)

Grubb's test detects one outlier at a time. This outlier is expunged from the data set and the test is iterated until no outlier is detected. A test statistic G is calculated and compared with the critical value. The Grubb's test statistic is given by

$$G = \frac{Max \left| Y_i - \overline{Y} \right|}{S}$$

Test Criteria/Decision Rule.

Hypothesis

H_O: There is no outlier in the given data set.

H_{ak}: There is outlier in the given data set.

Decision Rule:

Reject H_O if G

$$> \frac{\text{N-I}}{\text{N}}\sqrt{\frac{T^2\left(\frac{\alpha}{2N}\right),N-2}{N-2+T^2\left(\frac{\alpha}{2N}\right)N-2}}$$

at a given (α) level of significance, otherwise do not reject H_O.

Regression analysis is an estimating equation which expresses the functional relationship between two or more variables as well take care of the error term which is classified into; Simple linear regression and multiple linear regression.

2.3.1 Simple Linear Regression.

This is the type of linear regression that involves only two variables one independent and one dependent plus the random error term. The simple linear regression model assumes that there is a straight line (linear) relationship between the dependent variable Y and the independent variable X. This can be estimated by the least square estimate method expressed by.

2.3 Regression Analysis

$$\text{b i} = \frac{\text{n}\sum_{\text{i=I}}^{n} X_i Y_i - \left[\sum_{\text{i=I}}^{n} \text{X i}\right]\left[\sum_{\text{i=I}}^{n} \text{Y i}\right]}{n\sum_{i=I}^{n} X_i^2 - \left[\sum_{\text{i=I}}^{n} \text{X i}\right]^2} \quad \text{----------} \quad (4)$$

2.3.2 Multiple Linear Regressions

Multiple linear regression analyses three or more variables and the random error term.

This is expressed as follows.

$$Y_i = \beta_O + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + E \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (5)$$

$$\text{Where } Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1k} \\ 1 & X_{21} & X_{22} & \ldots & X_{2k} \\ \vdots & \vdots & & & \\ 1 & X_{1n} & X_{2n} & \ldots & X_{kn} \end{bmatrix}_{nxk}$$

$$\beta = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}_{kx1} \qquad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{nx1}$$

The matrix becomes.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1n} \\ 1 & X_{21} & X_{22} & \ldots & X_{2n} \\ \vdots & & & & \\ 1 & X_{1n} & X_{2n} & \ldots & X_{1n} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b\,k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

2.4 : Broilers Data As Used In The Study

Table 1

	Man- Hours	Boiler Capacity	Design Pressure	Boiler Type	Drum Type
S/N	Y	X_1	X_2	X_3	X_4
1	3137	120000	375	1	1
2	3590	65000	750	1	1
3	4526	150000	500	1	1
4	10825	1073877	2170	0	1
5	4023	150000	325	1	1
6	7606	610000	1500	0	1
7	3748	88200	399	1	1
8	2972	88200	399	1	1
9	3163	88200	399	1	1
10	4065	90000	1140	1	1
11	2048	30000	325	1	1
12	6500	441000	410	1	1
13	5651	441000	410	1	1
14	6565	441000	410	1	1
15	6387	441000	410	1	1
16	6454	627000	1525	0	1
17	6928	610000	1500	0	1
18	4268	150000	500	1	1
19	14791	1089490	2970	0	1
20	2680	125000	750	1	1
21	2974	120000	375	1	0

22	1965	65000	750	1	0
23	2566	150000	500	1	0
24	1515	150000	250	1	0
25	2000	150000	500	1	0
26	2735	150000	325	1	0
27	3698	610000	1500	0	0
28	2635	90000	1140	1	0
29	1206	30000	325	1	0
30	3775	441000	410	1	0
31	3120	441000	410	1	0
32	4206	441000	410	1	0
33	4006	441000	410	1	0
34	3728	627000	1525	0	0
35	3211	610000	1500	0	0
36	1200	30000	325	1	0

Note Y is the dependent variable which x_1, x_2,x_3 and x_4 are the independent variables.
For the purpose of this study, the following holds:
Y represents man hours
X_1 represent boiler capacity
X_2 represent design pressure
X_3 represent boiler type
X_4 represent drum type

3.0 Data Analysis
3.1 Dependent Variable Y Using Rosner's Test To Check For Outlier In The Dataset

1200	1206	1515	1965	2000	2048	2566	2635	2680
2735	2972	2974	3120	3137	3163	3211	3590	3698
3728	3738	3775	4006	4023	4065	4206	4268	4526
5651	6387	6454	6500	6565	6928	7606	10825	4791

I	**n-i**	$\bar{Y}_{0}^{i}$	**Sy$^{(i)}$**	**Y$^{(i)}$**	**R_{i+1}**	$\lambda_{i+1}\alpha(0.05)$
O	36	4290.7500	2702.7215	1200	1.144	2.99
1	35	4379.0571	2688.9615	14791	3.872	2.98
2	34	4072.8235	2016.9126	10825	3.348	2.97.

The decision rule is to reject H_O if $R_y.1 > \lambda_i +_1$ from our result above.
$R_y.1 < \lambda_i +_1$. That is 1.44 < 2.99 Accept H_O.
$R_y.2 > \lambda_i +_1$. That is 3.872 > 2.98 Reject H_O
$R_y. 3 > \lambda_i +_1$ That is 3.348 > 2.97 Reject H_O.
We therefore conclude that observation 14791 and 10825 are outliers.

3.2 3 Rosner's Test On The Independent Variable X_1 The Data Becomes.

30000	30000	30000	65000	65000	88200	90000
90000	120000	120000	125000	15000	150000	15000
50000	150000	150000	150000	441000	441000	441000
41000	441000	441000	441000	441000	610000	610000
61000	610000	627000	627000	1073877	1089490	

From the data above we have.

i	n-i	$\bar{X}_1(i)$	$S_{xi}^{(i)}$	$X_i^{(i)}$	R_i+_1	λ_{i+1} $(\alpha=0.05)$
0	36	318471.3056	281427.7874	3000	1.025	2.99
1	35	326713.3429	281093.5943	1089490	2.714	2.98
2	34	304278.7353	251511.9055	1073877	3.060	2.97

$R_{X1.1} > \lambda i +_1$ That is 1.025 < 2.99 accept H_O

$R_{X1.2} < \lambda i +_1$ That is 2.714 < 2.98 accept H_O

$R_{x1.3} > \lambda i +_1$ That is 3.060 > 2.97 Reject H_O

Therefore only observation 1073877 is an outlier in the data set of X_1 independent variable.

The null and alternative hypotheses are stated as fellows.

H_O: There are no outliers in the data set.

H_A: There is at least one outlier in the data set.

Critical region = $\frac{36-1}{\sqrt{36}} \sqrt{\frac{4.13}{36\text{-}2\text{+}4.13}}$

= 1.920

3.3 Grubb's Test

Grubb's Test on Y the Dependent Variable.

$\bar{Y}$ = 290.7500, S = 2702.7215.

i.	**G**			
1	0.427	(accept H_O)	< 1.920	Not an outlier
2	0.259	(accept H_O)	< 1.920	Not an outlier
3	0.087	(accept H_O)	< 1.920	Not an outlier
4	2.417	(reject H_O)	> 1.920	An outlier
5	0.099	(accept H_O)	< 1.920	Not an outlier
6	1.227	(accept H_O)	< 1.920	Not an outlier
7	0.201	(accept H_O)	< 1.920	Not an outlier
8	0.488	(accept H_O)	< 1.920	Not an outlier
9	0.417	(accept H_O)	< 1.920	Not an outlier
10	0.084	(accept H_O)	< 1.920	Not an outlier
11	0.830	(accept H_O)	< 1.920	Not an outlier
12	0.817	(accept H_O)	< 1.920	Not an outlier
13	0.503	(accept H_O)	< 1.920	Not an outlier

14	0.841	(accept H_O) < 1.920	Not an outlier
15	0.776	(accept H_O) < 1.920	Not an outlier
16	0.800	(accept H_O) < 1.920	Not an outlier
17	0.976	(accept H_O) < 1.920	Not an outlier
18	0.008	(accept H_O) < 1.920	Not outlier
19	3.885	(Reject H_O) > 1.920	An outlier
20	0.596	(accept H_O) < 1.920	Not an outlier
21	0.487	(accept H_O) < 1.920	Not an outlier
22	0.861	(accept H_O) > 1.920	Not an outlier
23	0.638	(accept H_O) < 1.920	Not an outlier
24	1.027	(accept H_O) < 1.920	Not an outlier
25	0.848	(accept H_O) < 1.920	Not an outlier
26	0.576	(accept H_O) < 1.920	Not an outlier
27	0.219	(accept H_O) < 1.920	Not an outlier
28	0.613	(accept H_O) < 1.920	Not an outlier
29	1.141	(accept H_O) < 1.920	Not an outlier
30	0.191	(accept H_O) < 1.920	Not an outlier
31	0.433	(accept H_O) < 1.920	Not an outlier
32	0.031	(accept H_O) < 1.920	Not an outlier
33	0.105	(accept H_O) < 1.920	Not an outlier
34	0.208	(accept H_O) < 1.920	Not an outlier
35	0.400	(accept H_O) < 1.920	Not an outlier
36	1.144	(accept H_O) < 1.920	Not an outlier

This shows that observation 4 and 19 are outliers on the dependent variable (Y) using Grubb's Test Method.

4.0 Conclusion

The above discussed statistical tests are used to determine if experimental observations are statistical outliers in the data set. Of course effective working with outliers in numerical data can be rather difficult and frustrating experience. Neither ignoring nor deleting them at all will be good solution if you do nothing, you will end up with a model that describes essentially none of the data neither the bulk of the data nor the outliers. Even though your numbers may be perfectly legitimate, if they lie outside the verge of most of the data, they can cause potential computational problem and thus influences problems.

4.1 Recommendation

Having carried out this study successfully the following recommendations were made.

(a) We recommend that experimenters should keep good record for each experiment.

All data should be recorded with any possible explanation or additional information.

(b) We recommend that analyst should employ robust statistical methods. These methods are minimally affected by outliers.

References

[1] Anscombe, F.J. (1960): *Rejection of Outliers Technometrics*, 2, 123-147.

]2] Grubbs,F.E (1950): *Sample Criteria for Testing Outlying observations: Annals of Mathematical sciences.*

[3] Jarrell M.G. (1994). A Comparison of two procedures, the Mabalanobis Distance and the Andrews – Pregibon statistics for identifying multivariate outliers. Researchers in the schools, 1:49-58.

[4] Rosner's Multiple Outlier Test Technometrics 25, No 2 May, (1983), 165-172.

[5] Rasmussen, J. L. (1988): Evaluating outlier identification tests: Mahalanobis D Squared and Comrey, 23 (2), 189-202.

[6] Steven, J.P. (1984). Outliers and Influential points in Regression Analysis. Psychological Bulletin, 95, 339-344..

Relative Efficiency of Split-plot Design (SPD) to Randomized Complete Block Design (RCBD)

Oladugba, A. V+, Onuoha, Desmond O*, Opara Pius N.++

+Department of Statistics, University of Nigeria, Nsukka,
*Dept of Maths/Statistics, Fed. Polytechnic Nekede, Owerri, 08035442403,
++Datafield Logistics Services, Port Harcourt, Rivers State.

Abstract

The relative efficiency of split-plot design (SPD) to randomized complete block design (RCBD) was computed using their error variance, sensitivity analysis and design planning. The result of this work showed that conducting an experiment using split-plot (SPD) without replication is more efficient to randomized complete block design (RCBD) based on comparison of their error variances, sensitivity analysis and design planning consideration.

Key words: Split-plot Design, Randomized Complete Block Design, Error variance, Sensitivity Analysis and Design planning.

Introduction

In experimental design, the Relative Efficiency (RE)of design say A to another design say B denoted as RE(A:B) is defined in terms of the number of replicates of design B required to achieve the same result as one replicate of design A. In view of this, the relative efficiency of split-plot design (SPD) to randomized complete block design (RCBD) denoted as RE (SPD:RCBD) is the number of replicates of RCBD required to achieve the same result as one replicate of SPD. Relative efficiency can be expressed in terms of percentage by multiplying it by 100. If RE (SPD:RCBD) > 100%, SPD is said to be more efficient to RCBD and if RE(SPD:RCBD) ≤ 100% SPD is said to be less efficient to RCBD.

The relative efficiency of two designs is mostly measured in terms of comparing their error variances and the design with the smallest variance is said to be more efficient than the other. This measure of relative efficiency does not put into consideration the probability of obtaining significant difference or detecting significant difference if they exist between the treatments. RCBD is said to be more efficient to complete randomized design (CRD) based on the comparison of their error variance since the error variance of RCBD is always smaller than that of complete randomized design (CRD). There is a decrease in the error degree of freedom of RCBD compare to CRD and a decrease in the error degree of freedom leads to an increase in the tabulated value thereby reducing the probability of obtaining a significant result since the decision rule is always to reject the null hypothesis if F-calculated is greater than F-tabulated. Based on this assessment which is sensitivity analysis, CRD is said to be more efficient than RCBD; in other words, the sensitivity of RCBD is decreased. From above, it can be clearly seen that the relative efficiency of any two designs cannot be best judged by considering the ratio of their error