ON THE TRACTABILITY OF SOME DISCORDANCY STATISTICS FOR MODELLING OUTLIERS IN A UNIVARIATE DATASET

ISSN: 2141 – 3290
www.wojast.com

MOFFAT, I. U. AND OKON, R. A.
Department of Mathematics & Statistics
University of Uyo, Uyo.(moffitto2011@gmail.com)

ABSTRACT

This paper compares the tractability of four discordancy statistics for modelling outliers based on extremeness. They are the Generlaized Extreme Studentized Deviate (ESD), Grubb's test, Hampel's method and the quartile method. However, the last two methods are seen to detect outliers even for datasets that are not approximately normal, although Hampel out performs the quartile method in some cases. However, a multiplier effect of 2.2 is proposed for the quartile method in addition to the robust statistics for accommodating the outliers.

INTRODUCTION

Outlier modelling has become a critical aspect of time series as they can lead to model mis-specification, testing, biased parameter estimation, inference, poor forecasts and inappropriate decomposition of the series (Barnett and Lewis, 1994; Moffat and Etuk, 2007). Although their detection uses mathematical methods, the way they are dealt with may depend on reasoned but ultimately subjective judgement. When outliers are found there are three methods of dealing with them: correction, omission and accommodation. Firstly, if the outlier has been generated by a mistake in data entry or in the construction of the data set (e.g. when merging files) then it may be possible to correct it. Sometimes this is the result of transcription errors and in others it might be possible to check the data with the respondents. If it cannot be corrected then it must be omitted as the second approach. This can also be applied to contaminants. Accommodations of outliers maintain the sample size but require additional parameters to be fit while omission reduces the sample size but might lead to a simple model. The decision on which method is better depends on a number of factors. If the contaminating process is very different from the main process, then the final model may be extremely complex and proves very difficult to report. Secondly, the contaminating process might not be relevant to the main purpose of the study and so modelling it is inappropriate. In both cases, omission may be the best solution. However, if the accommodation involves only adding a few additional parameters and/or is of some other relevance to the analysis then the method that accommodates outliers should be used (Bell, 2004). Thus, an *outlier* is considered as a data point whose response doesn't follow the general trend of the remaining dataset of the population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled-out, it is necessary to characterize normal observations.

Generally, there are two main reasons for outlier analysis. Firstly, outliers bias our estimates and we would like to prevent this. Secondly, we want to find potential causes of extreme scores, for example, to some subgroup of our data set (Osborne and Amy, 2004; Dehon *et al*., 2009a; Kaya, 2010; Gumedze *et al*., 2010; Moffat, 2011). In economic theory, this separation of the two goals of outlier analysis has an interesting interpretation. First of all, outliers can help in fitting imperfect theories to complex real world phenomena. This is the traditional usage, implicit in the use of dummy variables. But on the other hand, outliers can also be used to reveal where a theory does not work, or to check what aspects of it need refining in order to better describe the real world. The examination of outliers can therefore be justified not only from the traditional data analysis perspective, but also by appealing to the interaction of theoretical and empirical economics (Zellner, 1981). Outliers can take several forms in time series. There are additive and innovational outliers (Fox, 1972). An additive outlier affects a

single observation, which is smaller or larger in value than expected. In contrast an innovational outlier affects several observations. Three other types of outliers can be defined, namely level shifts, transient changes and variance changes. A level shift simply changes the level or mean of the series by a certain magnitude from a certain observation onwards. A transient change is a generalisation of the additive outlier and level shift in the sense that it causes an initial impact like an additive outlier but the effect is passed on to the observations that come after it. A variance change simply changes the variance of the observed data by a certain magnitude. Outliers affect the autocorrelation structure of a time series, and therefore they also bias the estimated Autocorrelation (ACF), Partial Autocorrelation (PACF) and the Extended Autocorrelation Functions (EACF). The exact results of the effects are complicated and require lengthy computations (Tsay, 1986a). In this work, two activities are adopted and considered essential for characterizing a set of data:

1. Examination of the overall shape of the graphed data for important features, including symmetry and departures from assumptions.
2. Examination of the data for unusual observations that are far removed from the mass of data. These points are often referred to as outliers.

DATA AND METHODOLOGY

The data comprises the number of students that graduated under normal four years duration from the Department of Statistics, University of Uyo, Uyo, from 2000/2001 to 2009/2010 session and the yearly inflation rates in Nigeria between 1961 and 2013.

Generalized Extreme Studentized Deviate (ESD):
This test (Rosner, 1975) is used to detect one or more outliers in a univariate dataset that follows an approximately normal distribution. Given the upper bound, *k*, the generalized ESD test essentially performs *k* separate tests: a test for one outlier, a test for two outliers, and so on, up to *k* outliers.
The hypothesis under consideration is:

H_0: There are no outliers in the dataset

H_1: There are up to *k* outliers in the dataset

Test Statistics: Compute

$$R_i = \frac{MAX|X_i - \bar{X}|}{S}$$

where $\bar{x}$ and *s* denote the sample mean and sample standard deviation, respectively. Here, we first remove the observation that maximizes $|x_i - \bar{x}|$ and then recompute the above statistic with *n*-1 observation. The process is repeated until *k* observations have been removed. This results in the *k* test statistics $R_1, R_2, ..., R_k$.

Significance Level: α

Critical Region: Corresponding to the *k* test statistics, we compute the following *k* critical values:

$$\lambda_i = \frac{(n-i)\, t_{p,n-i-1}}{\sqrt{(n-i-1+t^2_{p,n-i-1})(n-i+1)}}, \; i = 1, 2, \ldots, k$$

where $t_{p,v}$ is the $100p$ percentage point from the *t* distribution with *v* degrees of freedom and $p = 1 - \frac{\alpha}{2(n-i+1)}$. The number of outliers is determined by finding the largest *i* such that $R_i > \lambda_i$.

Grubbs' Test: Given a dataset that is approximately normal, Grubbs' (1969) test detects a single outlier in a univariate dataset by considering the following hypotheses:

H_0: There is no outlier in the data set

H_1: There is at least single outlier in the data set

The general formula for Grubbs' test statistic is defined as:

$G = \frac{Max|Y_i - \bar{Y}|}{S}$, where y_i is the i^{th} element of the data set, $\bar{y}$ is the sample mean and s denotes the standard deviation. The test statistic is the largest absolute deviation from the sample mean (in units) of the sample standard deviation. The calculated value of parameter G is compared with the critical value for Grubb's test. When the calculated value is larger or smaller than the critical value of choosing statistical significance, then the calculated value can be accepted as an outlier. The statistical significance level (α) describes the maximum probability of committing a Type I error.

Hampel's Test: In the calculation of Hampel's test, statistical tables are not necessary. Theoretically, this method is resistant, as it is not sensitive to outliers. It also has no restrictions as to the abundance of the dataset. The steps are as follows:

i. Compute the median (Me) for the total dataset.

ii. Compute the value of the deviation r_i from the median value, and this is done for all elements from the data set:

$r_i = x_i$ — Me, where, x_i is the sample data from the data set, $i = 1, \ldots, n$

n is the number of all elements in the set while Me– is the median.

iii. Calculate the median for deviation Me$|r_i|$

iv. Check the conditions: $|r_i| \geq 4.5\text{Me}|r_i|$

If the condition is executed, then the value from the dataset can be accepted as an outlier.

Quartile Test: To detect outliers using the quartile method, the following steps are considered:

Step 1: Calculate the upper quartile, Q_3.
Step 2: Calculate the lower quartile, Q_1.
Step 3: Calculate the gap between the quartiles: $H = Q_3 - Q_1$. A value lower than $Q_1 - 1.5H$ and higher than $Q_3 + 1.5H$ is considered to be a mild outlier (influential observation). A value lower than $Q_1 - 3H$ and higher than $Q_3 + 3H$, is considered to be an extreme outlier.

EMPIRICAL RESULTS

As a first step, a normal probability plot, histogram and sequence chart was generated

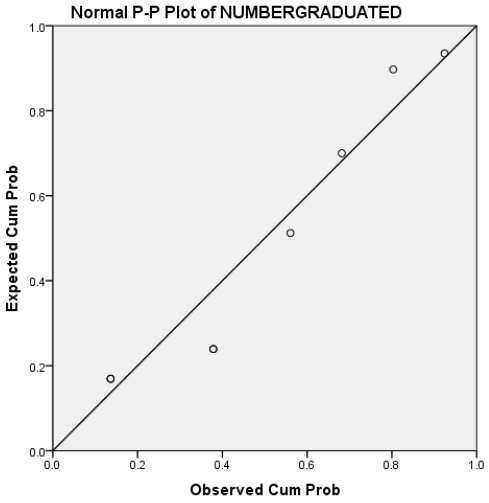


Figure 1:

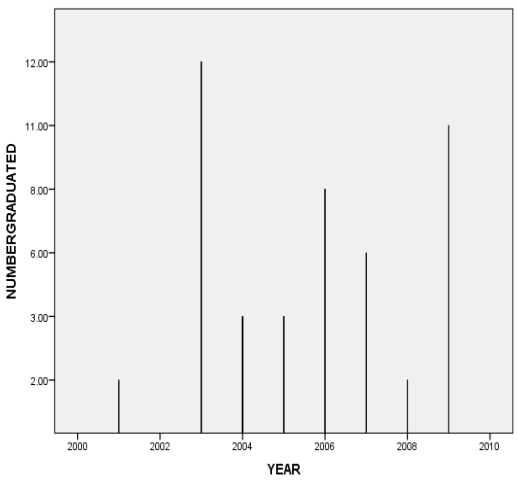
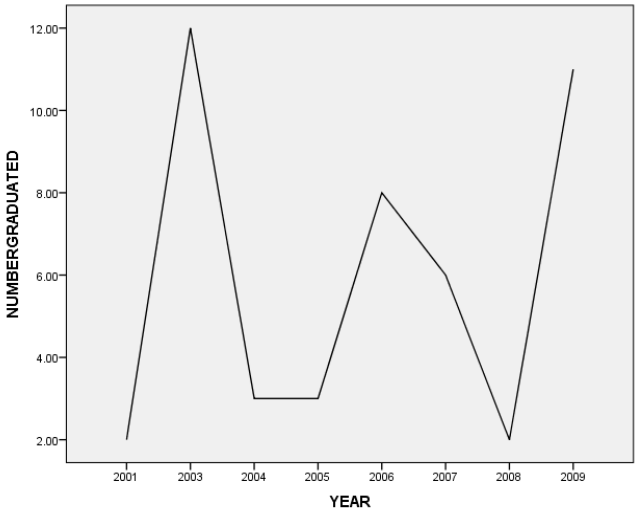


Figure 2:



In these experiments, Grubbs' test has given the same results in repeated cases. The three other methods did not detect additional outlier when the experiment was repeated unlike the Grubbs' test. Note that the experiment is only repeated after an outlier is detected.

Table 1: Comparison of the number of outliers detected using various discordancy tests.

Outliers test (two tailed test)	Number of outliers detected(test without outliers)	Number of outliers detected(Test with outliers(1st test)	Number of outliers detected(Test with outliers(2^{nd} test)	Total number of outliers detected	Sig(α)
Grubbs test	1	1	0	2	0.5%
Hampel	3	0	0	3	0.5%
Quartile	3	0	0	3	0.5%
Generalized ESD	3	0	0	3	0.5%

Considering data on the inflation rate in Nigeria from 1961 to 2013, a test of normality was first carried out using normal probability plot (Figure 3), histogram (with normal curve, Figure 4) and Kolmogorov-Smirnov test (Table 2).

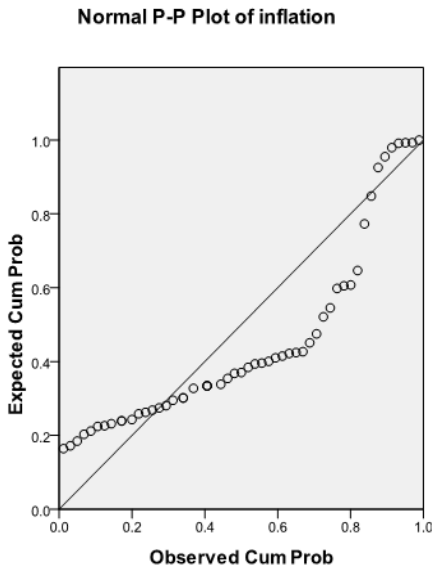


Figure 3:

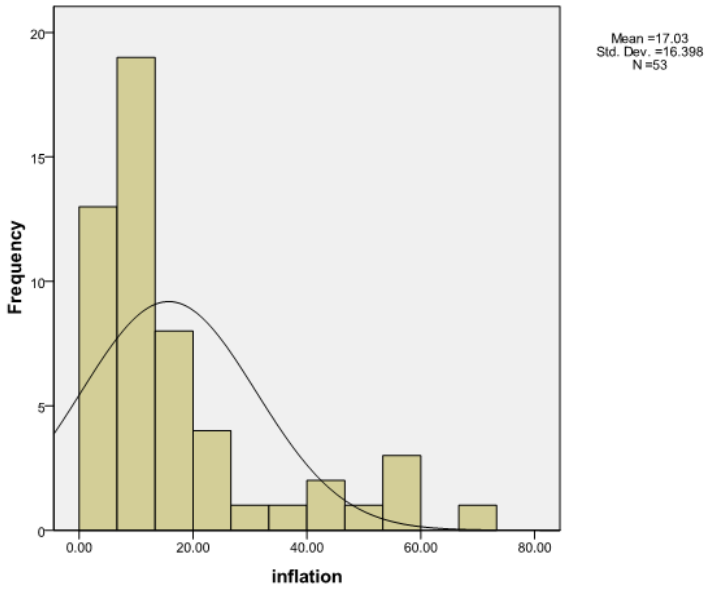


Figure 4: Histogram with Normal curve

Table 2: Tests for normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	Df	Sig.
Inflation	.253	53	.000	.758	53	.000

a. Lilliefors Significance Correction

The normality tests however revealed that the dataset does not approximately follow a normal distribution. Hence, Grubbs' test and the generalized Extreme Standardized Deviate (ESD) test are not appropriate for this detection. However, Hampel's method and the quartile method were used to detect the outliers as these methods have no restriction on the distribution of the dataset. Hampel's method detected 7 outliers while the quartile method detected 5 outliers. In a way of handling the outliers, robust statistics as already discussed in this was employed. The robust estimators of the mean and standard deviation were 11.6 and 7.8 respectively.

Specifically, inflation rates for year 1984, 1988, 1989, 1992, 1993, 1994, and 1995 were detected as outliers. The effect of these outlying points could be seen in the structure of the autocorrelation function as compared to the structure of the auto correlation function when these outliers are removed.

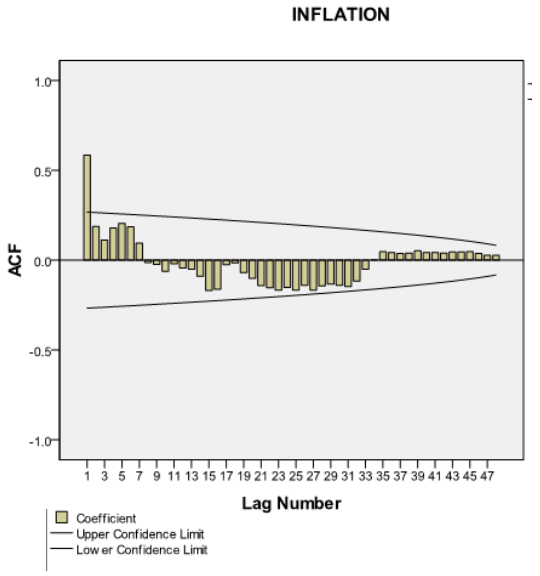


Figure 5: Autocorrelation Structure of Inflation Rates with Outliers

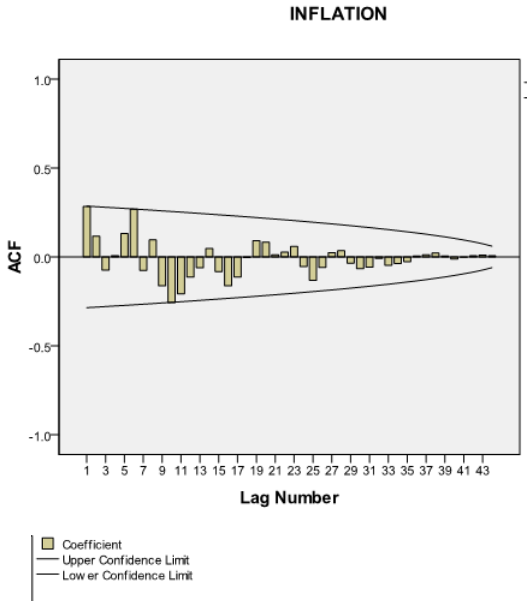


Figure 6: Autocorrelation Structure with Outliers removed

CONCLUSION

From our results in Table 1, it was revealed that Grubbs' test had low sensitivity for outlier detection. The other three methods are better than Grubbs as they could identify the maximum outliers. The charts and graphs reveal that the number of students that graduated under normal 4 years duration was very low in 2001/ 2002 and 2008/2009 admitted session and it was high in 2003/2004 and 2009/2010 admitted session. As our contribution, for any dataset that is approximately normal, it is recommended that the multiplier used in the quartile method which is conventionally given as: Upper quartile = $Q_3 + 1.5(Q_3 - Q_1)$ and Lower quartile = $Q_1 - 1.5(Q_3 - Q_1)$ be upgraded to: Upper quartile = $Q_3 + 2.2(Q_3 - Q_1)$ and lower quartile = $Q_1 - 2.2(Q_3 - Q_1)$. It has been observed that for an approximately normal data, the use of 1.5 as the multiplier reduces the normality of the data, whereas when 2.2 is used as the multiplier, it maintains the normality of the dataset.

From the two ACF plots above (Figures 5 & 6), it can be observed that the ACF with the outliers decayed very slowly indicating a high non-stationary series while the ACF when the outliers are removed decayed more rapidly indicating a seemingly stationary process. Again, the ACF with the outliers showed some cut-offs beyond the confidence interval while the ACF when the outliers are removed had all the points within the confidence interval.

REFERENCES

Barnett, V. And Lewis, T. (1994). Outliers in statistical data (3rd edition). New York: John Wiley.

Bell, J. F. (2004). Outliers and Multilevel models. 6th International Conference paper on social research methodology in Amsterdam.

Dehon, C., Gassner, M. and Verardi, V. (2009a). Beware of good outliers and overoptimistic conclusions. Oxf. Bull. Econ. Stat., 71(3): 437-452.

Fox, J. A. (1972). Outliers in time series. J. Royal Stat. Soc., Series B, 34: 350-363.

Grubbs, F. E. (1969). Procedures for detecting outlying Observations in Samples. American Statistical Association and American Society for Quality. Technometrics, 11, 1-21.

Gumedze, F. N., Welham, S. J., Gogel, B. J., Thompson, R. (2010). A variance shift model for detection of outliers in the linear mixed model. Comput. Stat. Data Anal., 54: 2128-2144

Kaya, A. (2010). Statistical modelling for outlier factors. Oze. J. Appl. Sci., 3(1): 185-194.

Moffat, I. U. & Etuk, E. H. (2007). Relationship between an innovative outlier model and the intervention effects model. Research Journal of Applied Sciences, 2(5): 574 – 578.

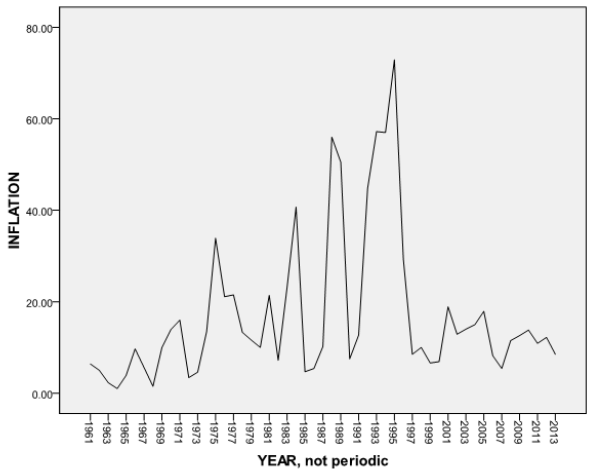
Moffat, I. U. & Etuk, E. H. (2011). On the biases of model parameters by aberrant observations. Journal of Mathematical Sciences, 22(2): 95-104.

Osborne, J. W., Amy, O. (2004). The power of outliers (and why researchers should always check for them). Practical Assessment, Res. Eval., 9(6). Available online at: http://PAREonline.net/getvn.aspv=9&n , accessed 03 February 2012.

Rosner Bernard(1975), On the Detection of many outliers, Technometrics, 17(2), 221-227.

Tsay R (1986a). Time series model specification in the presence of outliers. J. Am. Stat. Soc., 81: 132-141.

Zellner, A. (1981). Philosophy and objectives of econometrics. In :Currie D, Nobay R, Peel D (1981).Macroeconomic analysis: Essays in Macroeconomics and economics. Croom Helm, London. pp. 24-34.



Appendix A: A Time plot of the original data (1961 – 2013)