

Note on the Comparison of Some Outlier Labeling Techniques

Olewuezi, N.P.

Department of Statistics, School of Science,
Federal University of Technology, Owerri, Nigeria

Abstract: Problem statement: Methods proposed for estimating and resolving outliers are compared. **Approach:** In this respect, we exploit three well-known classifiers for identifying outliers to establish guidelines for the choice of outlier detection methods. **Results:** It was shown that the standard deviation is inappropriate to use here because it is highly sensitive to extreme values. **Conclusion/Recommendations:** The result of these estimated outliers is a better way of resolving large population.

Keywords: Outlier detection, detection methods, extreme values, Standard Deviation (SD), Median Absolute Deviation (MAD), breakdown point, labeling methods

INTRODUCTION

In statistical theory, an outlier is an observation that is numerically distant from the rest of the data. Outlier is an observation that lies outside the overall pattern of a distribution. Some outlier labeling methods such as the Standard Deviation (SD), the MADe and the Median rule are commonly used. These methods are quite reasonable when the data distribution is normal. Accordingly Bain and Engelhardt (1992) if data follows a normal distribution, it helps to estimate the likelihood of having extreme values in the data so that the observation two or three standard deviations away from the mean may be considered as an outlier in the data. Outliers affect significantly the estimates of the Transfer Function models and this can jeopardize the functions as model identification tools Olewmezi (2008). In non random distributions, outliers can decrease normality . When data depart from a normal distribution, a transformation to normality is simply a common step in order to identify outliers using a method which is quite effective in a normal distribution. There is only about a 1% chance you will get an outlying data point from a normally distributed populations; this means that, on average, about 1% of your subjects should be 3 standard deviations from the mean Osborne (2002) and Olewuezi, 2008. Iglewicz and Hoaglin (1993) observed that such a transformation could be useful when the identification of outliers is conducted as a preliminary step for the analysis of data. This helps to make possible the selection of appropriate statistical procedures for estimating and testing as well. The cases of outliers using sub multihalver and improved way of performance Criteria for Outlier

Identification are available in literature Onoghojobi (2010a, 2010b)

Traditionally, the sample mean and the sample variance give good estimation for data location and data shape if it is contaminated by outliers. When the database is contaminated, those parameters may deviate and significantly affect the outlier detection performance. Hampel (1971) introduced the concept of the breakdown point as a measure for the robustness of an estimator against outliers. The breakdown point is defined as the smallest percentage of outliers that can cause an estimator to take arbitrary large values. Hence, the larger breakdown point an estimator has, the more robust it is. For instance, the sample mean has a breakdown point of $1/n$ since a single large observation can make the sample mean and variance cross any bound. Hamper suggested the Median and the Median Absolute Deviation (MAD) as robust estimates of the location and the spread.

Osborne and Overbay (2004) categorized the effects of outliers on statistical analyses:

- Outliers generally serve to increase error variance and reduce the power of the statistical tests
- Outliers can seriously bias or influence estimates that may be of substantive interest
- If non-randomly distributed, they can decrease normality

MATERIALS AND METHODS

Outlier labeling techniques: In this study, three labeling methods are selected namely the method of Standard Deviation (SD), the MADe and the Median rule.

The method of Standard Deviation (SD method): This is a simple classical approach to identify outliers in a data set. The SD methods use less robust measures such as the mean and standard deviation which are highly affected by extreme values. Hence, their intervals have a tendency to be inflated as the data increases in skewers. This method is defined as:

$$\begin{aligned} 2 \text{ SD Method} &= \bar{X} \pm 2 \text{ SD} \\ 3 \text{ SD Method} &= \bar{X} \pm 3 \text{ SD} \end{aligned}$$

Where:

\bar{X} = Sample mean
SD = Slandered deviation of the data set

Observations outside these intervals may be considered as outliers.

According to the Chebyshev's inequality, if a random variable X with mean μ and variance σ^2 exists, then for any $k > 0$:

$$P[|x - \mu| \geq k\sigma] \leq \frac{1}{K^2}$$

Or:

$$P[|x - \mu| < k\sigma] \geq 1 - \frac{1}{K^2}, K > 0$$

The inequality $[1 - (1/k)^2]$ enables us to determine what proportion of our data will be within k standard deviation of the mean.

The MADe method: This is one of the basic robust methods which are largely unaffected by the presence of extreme values of the data set. It is defined as follows Eq. 1-3:

$$2 \text{ MADe Method} : \text{Median} \pm 2 \text{ MADe} \tag{1}$$

$$3 \text{ MADe Method} : \text{Median} \pm 3 \text{ MADe} \tag{2}$$

$$\text{MADe} = 1.483 \times \text{MAD} \text{ (for large normal data)} \tag{3}$$

It should be noted that MAD is an estimator of the spread in a data similar to the standard deviation, but has an approximately 50% breakdown point like the median. Therefore Eq. 4:

$$\text{MAD} = \text{Median} \left(|X_i - \text{Median}(x)| \mid i = 1, 2, \dots, n \right) \tag{4}$$

where, the MAD value is scaled by a factor of 1.483. It is similar to the standard deviation in a normal distribution. This scaled MAD value is the MADe. This

approach uses two robust estimators having a high breakdown point, that is, it is not unduly affected by extreme values even though a few observations make the distribution of the data skewed, the interval is seldom inflated, unlike the SD method.

The Median rule: If X_1, X_2, \dots, X_n is a random sample of size n arranged in order of magnitude, then we define the median as Eq. 5:

$$\text{Median, } \tilde{x} = \begin{cases} x_m, & n \text{ odd} \\ (x_m + x_{m+1} / 2), & n \text{ even} \end{cases} \tag{5}$$

Where:

$$m = \text{round up} (n / 2)$$

Hence, the median is the value that falls exactly in the centre of the data when the data are arranged in order.

Carling (2000) introduced the median rule for identification of outliers through studying the relationship between target outlier percentage and Generalized Lambda Distributions (GLDs). GLDs with different parameters are used for various moderately skewed distributions.

The Median rule is a robust estimator of location having an approximately 50% breakdown point.

The method is given by obtaining the range:

$$[K_1, K_2] = Q_2 \pm 2.3 \text{IQR}$$

Where:

Q_2 = Sample Median,
IQR = Inter Quartile Range

K_1 and K_2 are the lowest and highest values in the Median Rule Interval beyond which outlier(s) is/are detected.

RESULTS

Data for this study was obtained by reviewing the economic data page of www.nigeriabusinessinfo.com which is on the Gradate output by discipline.

SD method: We adopt the following steps in this identification:

Obtain the frequency distribution table and hence the mean and the standard deviation:

$$\begin{aligned} \bar{x} &= 1451.413 \\ s &= 1758.13 \end{aligned}$$

Determine the range of values for the 2SD and 3SD approaches.

For the 2SD Method we obtained the interval:

[-2064.85, 4965.67]

Thus, the following outliers were obtained 5243, 5271, 5818 and 8962.

Also, for the 3 SD methods, we obtained the following interval:

[-3822.98, 6725.80]

The value 8962 is the only outlier obtained.

MADe method: In this method, we employ the median and Median Absolute Deviation (MAD) of the data set:

The median of the raw = 66705
From Eq. 4, MAD = 521
Also, from Eq. 3, MADe = 772.64

The range using the 2 MADe method is [-877.78, 2212.78], identifying 9 outliers which are:

2425, 2800, 2917, 3246, 3661, 5243, 5271, 5818 and 8962

The 3 MADe method gives us the range [-1650.42, 2985.42] and six outliers were identified as:

3246, 3661, 5243, 5271, 5818 and 8962

Median method: This identification procedure uses the median and the Inter Quartile Range (IQR):

The IQR = 1182.5 from Eq. 6, we obtained the:

$[K_1, K_2] = [-2052.25, 3387.25]$

The following outliers were obtained 3661, 5243, 5271, 5818 and 8962.

DISCUSSION

In this study, we have used several outlier labeling methods. Each method has different measures to detect outliers and shows different behaviours according to the skewness and sample size of the data.

The SD method identified only one outlier. Depending on the approach used in the MADe method, the 2 MADe approach detected nine outliers and six for the 3 MADe methods. The median rule identified five outliers. The 2 MADe methods classify more observations as outliers than any other method. It should be noted that the MADe and the Median rule increase in the percentage of outliers on the right side of the distribution as the skewness of the data increases while the SD method seldom change in each sample size.

CONCLUSION

In terms of a particular number of outliers identified by any of the methods of the same data set,

the three methods oppose each other. Most intervals used to identify possible outliers in outlier labeling methods are effective under the normal distribution. Although these methods are quite powerful with large normal data, it may be problematic to apply them to non-normal data, or small sample sizes without information about their characteristics in these circumstances. This is so since each labeling method has different measures to detect outliers. It should also be noted that the expected outlier percentages change differently according to the sample size or the distribution type of the data. It is worthy to note that all methods identified the value 8962 as an outlier. The SD method is inappropriate to use here because it is highly sensitive to extreme values. The data used in study is large and normally distributed, with large gap between the majority of the data and some extreme values. Hence, we cannot at this juncture recommend a unique method to detect outliers because some methods are efficient for detecting certain types of outliers but fail to detect others.

In a future work, we are planning to include more outlier labeling techniques and to consider larger data sets.

REFERENCES

- Bain, L.J. and M. Engelhardt.,1992. Introduction to Probability and Mathematical Statistics. 2nd Edn., Duxbury, Belmont, ISBN: 0534380204, pp: 644.
- Carling, K.,2000. Resistant outlier rules and the non-gaussian case. *Comput. Stat. Data Anal.*, 33: 249-258. DOI: 10.1016/S0167-9473(99)00057-2
- Hampel, F.R., 1971. A general qualitative definition of robustness. *Ann. Mathe. Stat.*, 42: 1887-1896.
- Iglewicz, B. and D.C. Hoaglin, 1993. How to Detect and Handle Outliers. 1st Edn. ASQC Quality Press., Milwaukee, ISBN : 087389247X , pp: 87.
- Onoghojobi, B. 2010a. Subsample Goal Model for Multihalver on outliers. *J. Math. Stat.* 6: 347-349. DOI:10.3844/jmssp.2010.347.349
- Onoghojobi, B. 2010b. An instant of performance criteria for outlier identification. *J. Math. Stat.*, 6: 325-326. DOI:10.3844/jmssp.2010.325.326
- Olewuezi, N.P., 2008. Estimation of the outlier free and outlier contaminated transfer function models. *Global J. Mathe. Sci.*
- Osborne, J.W. and A. Overbay, 2004. The power of outliers (and why researchers should always check for them). North Carolina State University.
- Osborne, J.W., 2002. Notes on the use of data transformations. *Practical Assess. Res. Evaluation.*