Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/110306

A Monte Carlo Simulation Comparison of Methods of Detecting Outliers in Time Series Data

Egbo. Mary Nkechi [1], *Bartholomew. Desmond Chekwube*[2,*], *Obite. Chukwudi Paul*[2] *and Kiwu. Lawrence Chizoba*[2]

[1]Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Uli, Anambra State, Nigeria
[2]Department of Statistics, Federal University of Technology Owerri, Imo State, Nigeria

Received: 19 May 2021, Revised: 20 Aug. 2021, Accepted: 11 Sep. 2021.
Published online: 1 Sep. 2022.

Abstract: Observations that are considerably different from the rest of the data are referred to as outliers. Outliers in a dataset have a number of undesirable consequences for statistical analysis. The negative implications could include a decline in prediction quality and the inclusion of mistakes in model parameter estimates. Currently, just a few literature reviews have been done on these topics. As a result, four outlier detection methods that are specifically developed to find outliers in univariate standard normal time series datasets were compared. Comparative approaches that are simple like Mean Squared Error, Coefficient of Variation, Standard Error of Mean and Percentage Mean Success Rate, computed from outliers detected in a monte carlo simulation of samples of sizes 500 and 1,500 were proposed and used to select the best outlier detection method(s).

Keywords: Coefficient of Variation; Mean Squared Error; Outliers; Simulation; Standard Error of Mean.

1 Introduction

Any observation in a statistical data, collected for analysis and interpretation that differs obviously from the rest of the observations is called an outlier. Grubbs [1] remarked that an outlier is any member of the dataset that appears differently from the rest of the data in which it occurs. Outlier detection is the identification of these outliers. There are two types of outliers: univariate and multivariate. Univariate outliers can be detected in a single feature space, whereas multivariate outliers are frequently found in a q-dimensional space, where q is the number of features. Any of the following methods can be used to add outliers into a dataset: (1) by human errors during data entry (2) by experimental errors (3) by sampling errors (4) they can also occur naturally etc. Outlier detection is essential for almost any quantitative discipline where data analysis and visualization are carried out.

This is simply because the presence of outlier affects the quality of model used for prediction or classification especially in machine learning and any quantitative discipline like Physics, Finance, Statistics, Cyber Security etc. Outlier detection can be used to tackle a variety of problems, including quality control on a factory manufacturing line to detect flaws, detecting differences in usage patterns that could suggest fraudulent activity, such as a stolen automated machine (ATM) card, and so on. The study of outlier detection in a data set is one of the basic screening processes before formal analysis in statistics. Since the presence of outliers in a dataset could reduce the precision of parameter estimation and violation of some laid down assumptions, it is therefore necessary to consider methods of removal of outliers from time series data.

Some extant literatures exist on outlier detection methods in time series using different approaches – linear and Bayesian model-based approaches. The model-based approach which is often used when the series' structure is known was first studied by [2] in autoregressive models (AR). The approach was later extended to Autoregressive Moving Average (ARMA) models by [3, 4, 5, 6]. Secondly, Chen and Liu [6] suggested a linear model-based technique that uses an iterative procedure to detect outliers. Abraham and Box [7] proposed a Bayesian model-based technique for autoregressive models and Smith and West [8] proposed sequential judgments using dynamic linear systems. Some authors have also worked on comparison of outlier detection methods using simulation study [9, 10, 11, 12, 13, 14]. The impact of outliers on parameter estimation has been investigated by [15] considering autoregressive moving-average (ARMA) models and the implication on forecasts are addressed by [16] and [17]. When the time location of an outlier is

[*]Corresponding author e-mail: desmond.bartholomew@futo.edu.ng

known or missing value approaches, intervention analysis by [18] was developed to handle outlier detection and also [19, 20] may be useful. The case of the unknown time location is usually more controversial. Shittu [21] also detected outliers in discrete univariate samples with less attention paid on what to do with detected outliers. Ahmet [22] treated outliers in time series which have two special cases, Innovational outliers (IO) and additive outliers (AO). The presence of AO indicates that action is required, which could include adjusting the measurement device or correcting a mistake made by the person performing the observation and recording. If IO occurs, however, the measuring operation does not need to be adjusted. Seasonal outliers in time series were studied by [23] using a conventional approach for detecting and correcting outliers, such as additive, innovative, level shift, and transient change outliers. In linear regression, a new approach was utilized by [24] to detect outliers. The algorithm utilized in least trimmed squares (LTS) estimation was a non-iterative robust covariance matrix and concentration steps. Arimiyaw et al. [25] investigated the detection of influential outliers in linear regression analysis using one artificial and one real data set, and proposed the coefficient of determination ratio (CDR) as a metric for detecting influential outliers in linear regression analysis.

Depending on the circumstances, such as the underlying distribution of the research data and the type of the data, the performance of many strategies in outlier detection may vary (discrete or continuous). However, the questions that may be raised are: is there any approach among the existing ones that can be deemed to be the best in terms of detecting outliers in some way? Are there any strategies that work better than others at detecting outliers? The main purpose of this research is to examine and compare some typical outlier identification strategies in univariate time series data, which may be useful in determining the best method for a given situation. We plan to (i) look at some existing outlier detection algorithms for univariate time series data that presume the dataset is normally distributed. (ii) Identify the effect of sample sizes on the performance of the approaches, and (iii) determine the optimum method of outlier detection (based on certain comparative criteria).

Because all of the outlier identification approaches considered in this study require that the data follows normal distribution, this paper is limited to a Monte Carlo simulation that follows standard normal distribution. The four existing outlier detection methods in a time series data considered in this study are – (i) Standard deviation method (two standard deviations and three standard deviations approaches), (ii) Z-score method (the modified Z-score method), (iii)Median method and (iv) the box plot method.

The rest of this paper is divided into five sections namely: materials and methods, results and discussion, summary, conclusion and recommendation, references and appendix.

2 Experimental Sections

This study uses Monte Carlo simulated univariate data with a mean of 0 and a standard deviation of 1 which is a standard normal distribution. Sample 1 (with a size of 500) and sample 2 (with a size of 500) will be screened using four outlier detection methods (of size 1,500). The reason for the different sample sizes is to investigate if there is any effect of sample size on the selected methods. The results from the outlier detection methods will be compared using the methods of comparison listed in section 2.2 of this study.

2.1 Materials

Some of the existing outlier detection methods considered in this paper will be reviewed and the outliers detected will be displayed graphically. The Figure which shows the graphical display shall have three lines – lower bound, mid-point and upper bound. Any point in the time series data less than the lower bound or greater than the upper bound is considered as an outlier.

2.1(a) Standard Deviation (SD) Method Test

The simple classical approach to screening outliers is to use SD (Standard Deviation) method. It is defined for 2 SD and 3SD methods, respectively as

$$\bar{x} \pm 2SD \quad (1)$$

$$\bar{x} \pm 3SD \quad (2)$$

where

$\bar{x}$ is the sample mean and SD is the sample standard deviation

The lower, mid-point and upper bound are specified as:

Lower bound $= \bar{x} - 2SD$ or $\bar{x} - 3SD$; Mid-point $= \bar{x}$; Upper bound $= \bar{x} + 2SD$ or $\bar{x} + 3SD$

2.1(b) Z-Score Test

Another method used to screen data for outlier that uses the mean and standard deviation is the Z-Score test, which is expressed as

$$Z = \frac{x_i - \bar{x}}{\sigma} \tag{3}$$

where

$X \sim N(\mu, \sigma^2)$, x_i is the ith simulated value and σ is the standard deviation of data.

The lower, mid-point and upper bound are specified as:

Lower $= -3$; Mid-point $= \bar{x}$; Upper $= 3$

When data follows a normal distribution, it gives a fair criterion for identifying the outlier. According to [26], the maximum possible - score is determined by the sample size and is calculated as $(n-1)\sqrt{n}$. It should be said here that z-score method is not very good for outlier labeling, particularly in small data sets.

2.1(c) The Modified Z-Score Test

The sample mean and sample standard deviation are two estimators used in the Z – score outlier detection method. Even a single wild number can have an impact on these estimators. Using the median and median of absolute deviation (MAD) instead of the sample mean and standard deviation, the modified z-score test method solves this problem [27]. Let MAD $= |x_i - \tilde{x}|$, where $\tilde{x}$ is the sample median. Thus, the modified Z-score (M_i) is computed as

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}, \tag{4}$$

where

$$E(MAD) = 0.675\sigma \tag{5}$$

For large normal data, [27] suggested that observations are labeled outliers when $|M_i| > 3.5$.

The lower, mid-point and upper bound are specified as:

Lower $= -3.5$; Mid-point $= \bar{x}$; Upper $= 3.5$

2.1(d) Box Plot

This is also known as Tukey's technique, which was developed by [28]. It imposes no distributional assumptions on the data; therefore, it is a good method for spotting outliers in univariate data. The quartiles are used to calculate it. To use this method to discover outliers, the following basic steps must be followed.

Step 1: Calculate the upper quartile Q_3

Step 2: Calculate the lower quartile Q_1

Step 3: Calculate the inter-quartile range H ; $H = Q_3 - Q_1$ (6)

Step 4: The fence is given by $[Q_1 - 1.5 * H,\ Q_3 + 1.5 * H]$ (7)

Step 5: A value between intervals of the fence is called a possible outlier and data outside the outer fences are extreme outliers. This method is good when data is symmetric. It does not also depend on mean and standard deviation.

The lower, mid-point and upper bound are specified as:

Lower $= Q_1 - 1.5 * H$; Mid-point $= \tilde{x}$; Upper $= Q_3 + 1.5 * H$

2.1(e) Median Rule

The following basic steps are followed to achieve outlier detection using this method.

Step 1: Calculate the median of the data set

Step 2: Calculate the median absolute deviations as

$$\text{MAD} = |x_i - \tilde{x}| \tag{8}$$

Step 3: Find the median of the MADs in step two above as MMAD

Step 4: A point is called outlier if

$$|x - \tilde{x}| \ > \ 2 * (\text{MMAD}/0.6745) \tag{9}$$

The lower, mid-point and upper bound are specified as:

Lower $= \tilde{x} - 2 * (\text{MMAD}/0.6745)$; Mid-point $= \tilde{x}$; Upper $= \tilde{x} + 2 * (\text{MMAD}/0.6745)$

where

x is the time series data point and $\tilde{x}$ is the median of the series.

2.1(f) Median Absolute Deviation Rule (MADe)

Ratcliffe [29] established one of the most basic robust outlier detection algorithms, which is essentially unaffected by the existence of extreme values in the data. The standard deviation method is comparable to this strategy. In this method, however, the median and median absolute deviation is frequently used instead of the mean and standard deviation. The MADe method is defined for 2 MADe and 3 MADe methods respectively as follows:

$$Median \pm 2MADe \tag{10}$$

$$Median \pm 3MADe \tag{11}$$

where

$$MADe = 1.483\text{MAD} \tag{12}$$

for large normal data.

This is because when it is scaled by a factor of 1.483, it is similar to the standard deviation method in a normal distribution.

The lower, mid-point and upper bound are specified as:

Lower $=\tilde{x}-2*(\text{MADe})$ or $\tilde{x}-3*(\text{MADe})$; Mid-point $=$ $\tilde{x}$; Upper $=$ $\tilde{x}+2*(\text{MADe})$ or $\tilde{x}+3*(\text{MADe})$

where

x is the time series data point and $\tilde{x}$ is the median of the series.

The data set for this method must be large and must be approximately normal.

2.2 Methods of Comparison (Efficiency of outlier detection methods)

With the above reviewed outlier detection methods, it is important to know which of these methods under study is most efficient in screening outlier. We are going to use four different methods of comparison to test for the more efficient method namely:

Standard Error of Mean (SEM): The standard error of the mean is commonly defined as the difference between the estimated sample mean and the population mean. The sample will be more representative of the total population if the standard error is modest. However, the smaller the standard error tends to be the more data points used in the sample mean computation. It is mathematically given as:

$$\text{SEM} = \frac{\sigma}{\sqrt{n}} \tag{13}$$

where σ is the standard deviation of the points detected as outlier by the method and n is the number of points detected as outlier by the method.

Mean Square Error (MSE): The mean square error is a statistic that measures how near forecasts or predictions (the values of outliers) are to the actual values. The mean square error is a statistical tool that calculates the average squares of mistakes or deviations using the following formula:

$$\text{MSE} = \frac{\sum_{\text{i}=1}^{\text{n}} (x_i - \hat{x}_i)^2}{n} \tag{14}$$

where x_i is the upper or lower bound value depending on the sign of the outlier detected and $\hat{x}_i$ is the i$^{\text{th}}$ outlier detected by the method.

Percentage Mean Success Rate (PMSR): This metric represents the percentage of outliers found using a method in which the numerator is the number of outliers found and the denominator is the sample size of simulated time series data. The PMSR (percentage mean success rate) is calculated as follows:

$$\text{PMSR} = \frac{\text{Number of outliers detected}}{\textit{sample size of the simulated data}} \times 100 \tag{15}$$

Coefficient of Variation (CV): It is technically defined as the ratio of a sample's standard deviation to its mean for the univariate case, and it is one of the applications of standard deviation under the measures of dispersion in Statistics. It is significant since the standard deviation doesn't tell us anything about a single set of data's variabilities. The coefficient of variation increases as the level of dispersion around the mean increases, but the smaller the coefficient of variation, the more precise the estimate (outlier). The mathematical expression of the coefficient of variation is given by

$$CV = \frac{S}{\bar{x}} \tag{16}$$

where

S is the sample standard deviation of the points detected as outlier and $\bar{x}$ is the mean of the points detected as outlier.

Consider Table 1, the standard error of the mean number of outliers detected by the various outlier detection methods were displayed and ranked in ascending order. The standard errors were computed using equation (13). The smaller the standard error, the better the method used. The median, two median absolute deviations and the two standard deviation methods were the best methods when comparing the methods with standard error criterion irrespective of the sample sizes. The N/A means not applicable and this is because no outliers were detected by the method.

3 Results and Discussion

Table 1 Comparison of outlier detection methods using Standard error of mean outliers detected.

Sample Size	Outlier Detection Method	SEM	Rank (Ascending)
500	Two Standard Deviation	0.461219	3
	Three Standard Deviation	2.980994	6
	Z-score	2.980994	6
	Modified Z-score	N/A	N/A
	Median	0.405016	1.5
	Two Median Absolute Deviation Rule	0.405016	1.5
	Three Median Absolute Deviation Rule	2.980994	6
	Box plot	1.216410	4
1,500	Two Standard Deviation	0.288889	3
	Three Standard Deviation	1.924822	6.5
	Z-score	1.924822	6.5
	Modified Z-score	N/A	N/A
	Median	0.245998	1
	Two Median Absolute Deviation Rule	0.267141	2
	Three Median Absolute Deviation Rule	1.430479	5
	Box plot	0.852974	4

Table 1 displays the coefficient of variation values of the outlier detection methods considered in this work. The coefficient of variation was computed using equation (16). Again, for this criterion we are interested in the methods with the lowest coefficient of variation value. The box plot, three median absolute deviations, three standard deviations and z-score methods were the three best methods in the 500 sample size while two standard deviations, three median absolute deviation and three standard deviation methods were the best when the sample size was increased from 500 to 1,500. The implication is that the efficiency of the box plot is affected by sample size.

Table 3 shows how the mean square errors were computed. The lower and upper values that determines if a data point is an outlier were used as the actual values and the data points detected as outlier were considered as the predicted values.

Table 2 Comparison of outlier detection methods using coefficient of variation of outliers detected.

Sample Size	Outlier Detection Method	CV	Rank (Ascending)
500	Two Standard Deviation	-3.738643	5
	Three Standard Deviation	-76.51674	3
	Z-score	-76.51674	3
	Modified Z-score	N/A	
	Median	-2.753979	6.5
	Two Median Absolute Deviation Rule	-2.753979	6.5
	Three Median Absolute Deviation Rule	-76.51674	3
	Box plot	-137.4133	1
1,500	Two Standard Deviation	-34.56066	1
	Three Standard Deviation	-12.93339	3.5
	Z-score	-12.93339	3.5
	Modified Z-score	N/A	N/A
	Median	-6.397830	6
	Two Median Absolute Deviation Rule	-8.323819	5
	Three Median Absolute Deviation Rule	-17.72162	2
	Box plot	3.754439	7

Table 3 Computation of the Mean square errors.

Sample Size of 500			Sample size of 1,500		
Box plot method					
Outlier	Upper or Lower	MSE	Predicted	Actual	MSE
-3.03609	-2.50649	0.0727962	-3.03609	-2.56321	0.3033862
-2.56286	-2.50649		-2.6058	-2.56321	
-2.6058	-2.50649		2.559552	2.530759	
2.559552	2.549602		2.589201	2.530759	
2.589201	2.549602		2.925898	2.530759	
2.925898	2.549602		3.096502	2.530759	
			-2.89691	-2.56321	
			2.707111	2.530759	
			2.901192	2.530759	
			2.699391	2.530759	
			2.681024	2.530759	
			-4.17692	-2.56321	

The sample size (n) is the number of outliers detected by such method, for instance in Table 3, the sample sizes are 6 and 12 for 500 and 1,500 sample sizes respectively. The mean square errors were computed using equation (14).

Table 4 Comparison of outlier detection methods using Mean square error of outliers detected.

Sample Size	Outlier Detection Method	MSE	Rank (Descending)
500	Two Standard Deviation	0.223237883	5
	Three Standard Deviation	0.007418984	2
	Z-score	0.003396797	1
	Modified Z-score	N/A	N/A

	Median	1.399866416	7
	Two Median Absolute Deviation Rule	0.251389185	6
	Three Median Absolute Deviation Rule	0.040926548	3
	Box plot	0.072796158	4
1,500	Two Standard Deviation	5.580947065	7
	Three Standard Deviation	0.385533942	5
	Z-score	0.350309964	4
	Modified Z-score	0.458216625	6
	Median	0.248395324	1
	Two Median Absolute Deviation Rule	0.287264316	2
	Three Median Absolute Deviation Rule	0.320580088	3
	Box plot	0.303386161	6

The mean square errors displayed in Table 4 suggested that box plot and standard deviations rule were the best methods for small sample sizes while median methods were better for large sample sizes (1,500).

Table 5 Comparison of outlier detection methods using success rate of outliers detected.

Sample Size	Outlier Detection Method	Number of outliers detected	PMSR
500	Two Standard Deviation	25	5.00%
	Three Standard Deviation	2	0.40%
	Z-score	2	0.40%
	Modified Z-score	0	0.00%
	Median	29	5.80%
	Two Median Absolute Deviation Rule	29	5.80%
	Three Median Absolute Deviation Rule	2	0.40%
	Box plot	6	1.20%
1,500	Two Standard Deviation	67	4.47%
	Three Standard Deviation	4	0.27%
	Z-score	5	0.33%
	Modified Z-score	1	0.07%
	Median	84	5.60%
	Two Median Absolute Deviation Rule	79	5.27%
	Three Median Absolute Deviation Rule	6	0.40%
	Box plot	12	0.80%

Using the percentage mean success rate in equation (15), we are interested in an outlier detection method detecting as much outliers as possible. Table 5, therefore suggests that two standard deviations, median and two median absolute deviation rules were the best outlier detection methods for all sample sizes (small and large samples).

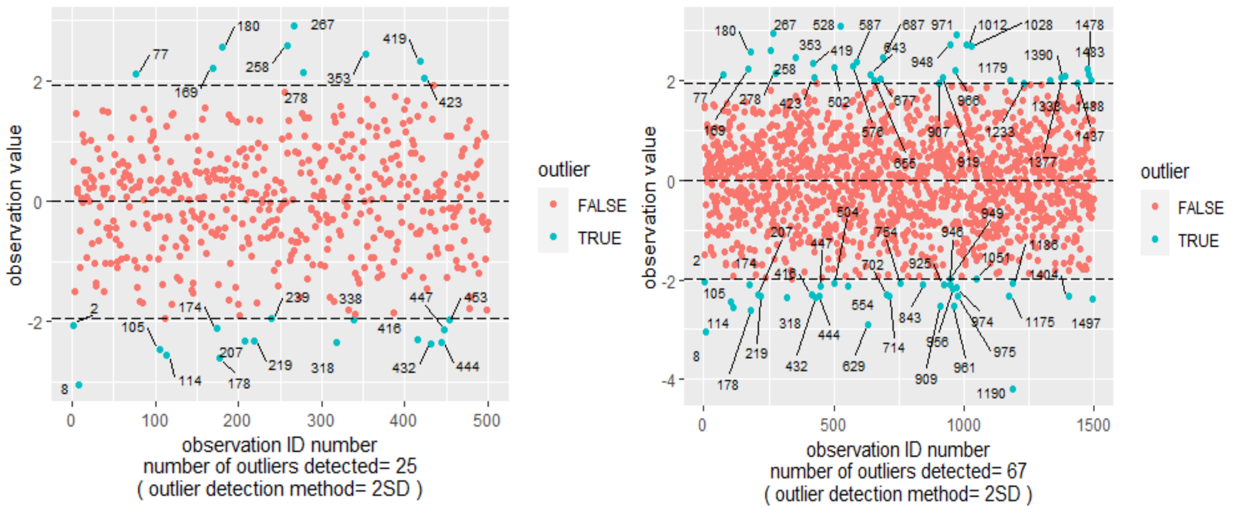

Fig.1: Outlier detection output for Two Standard deviation method.

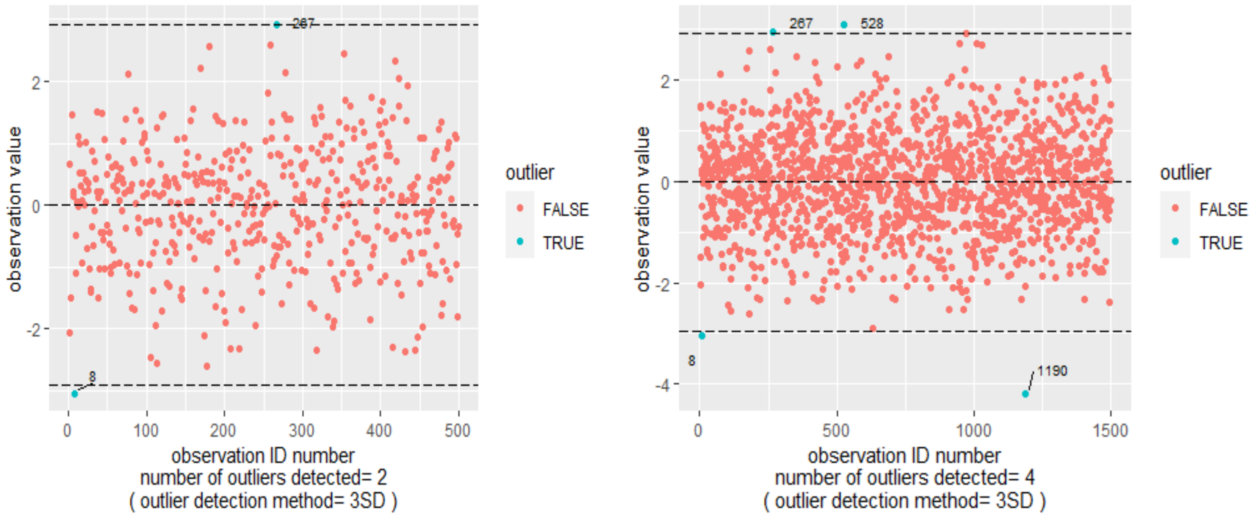

Fig.2: Outlier detection output for Three Standard deviation method.

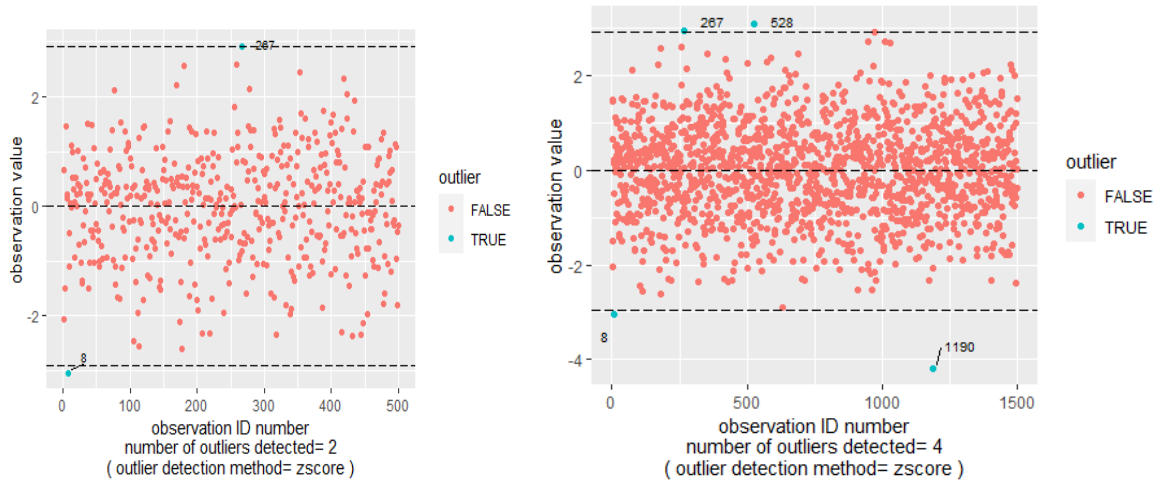

Fig. 3: Outlier detection output for zscore method

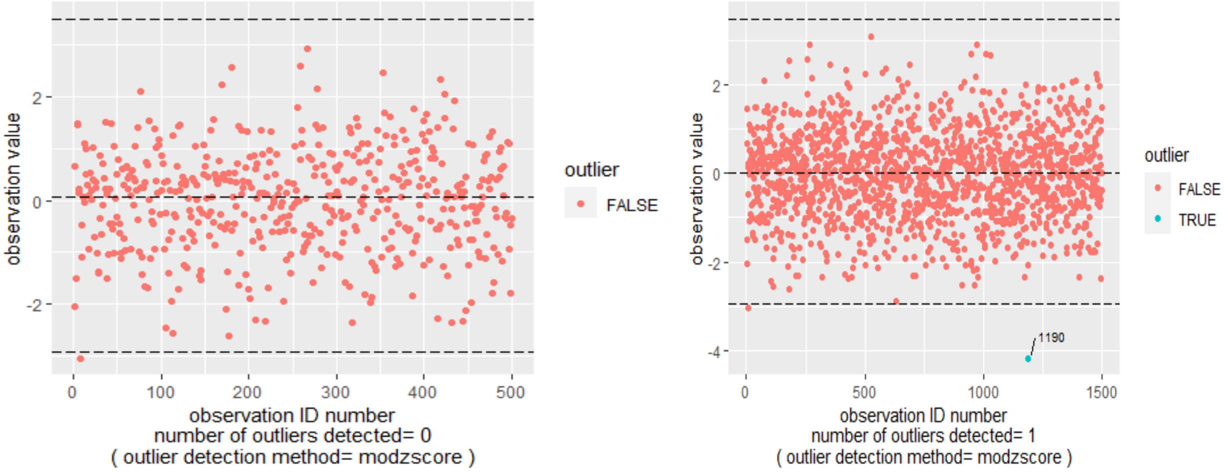

Fig. 4: Outlier detection output for modified z-score method.

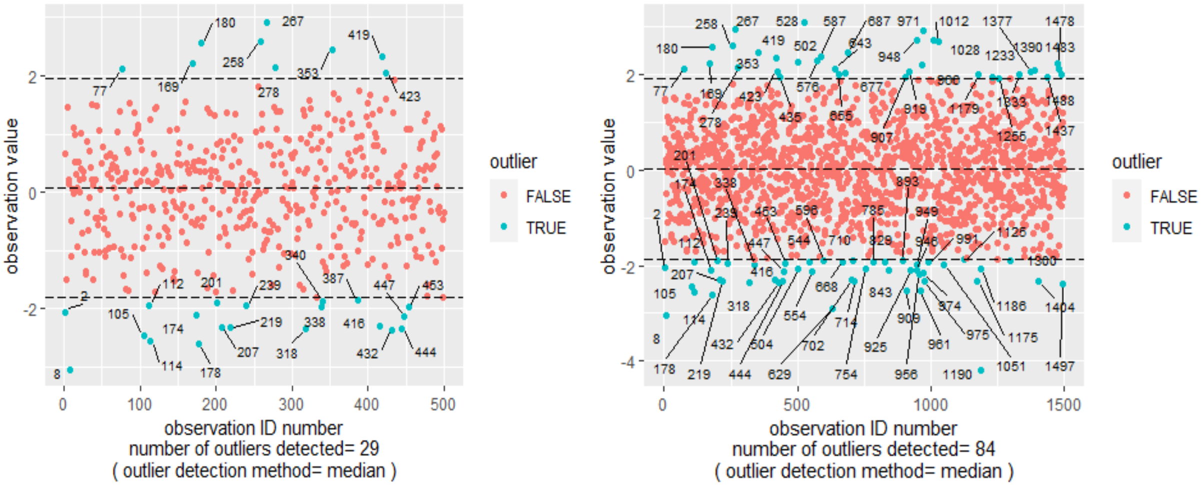

Fig. 5: Outlier detection output for median method.

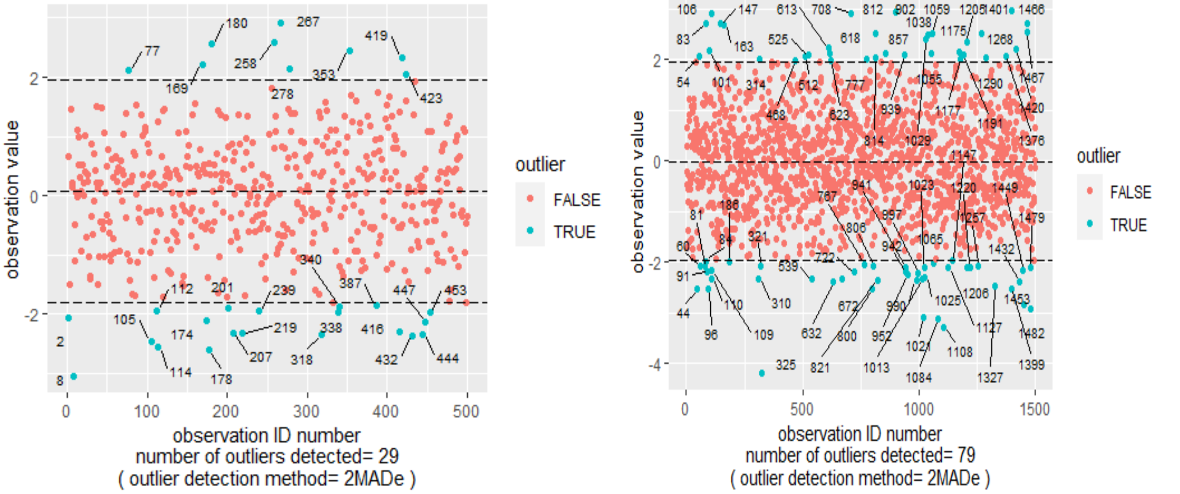

Fig. 6: Outlier detection output for 2 median absolute deviation rule method.

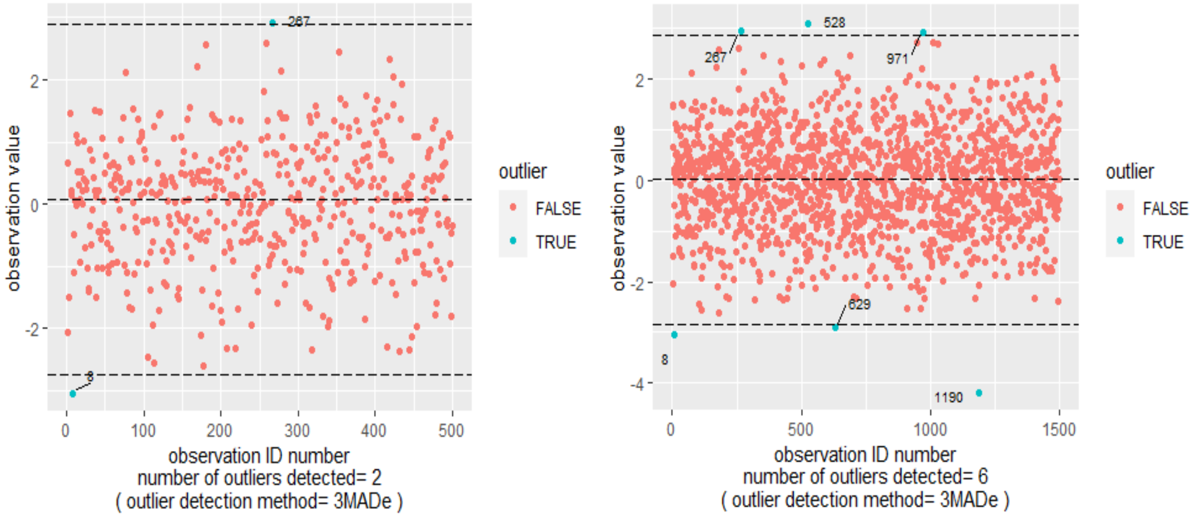

Fig. 7: Outlier detection output for 3 Median absolute deviation rule method.

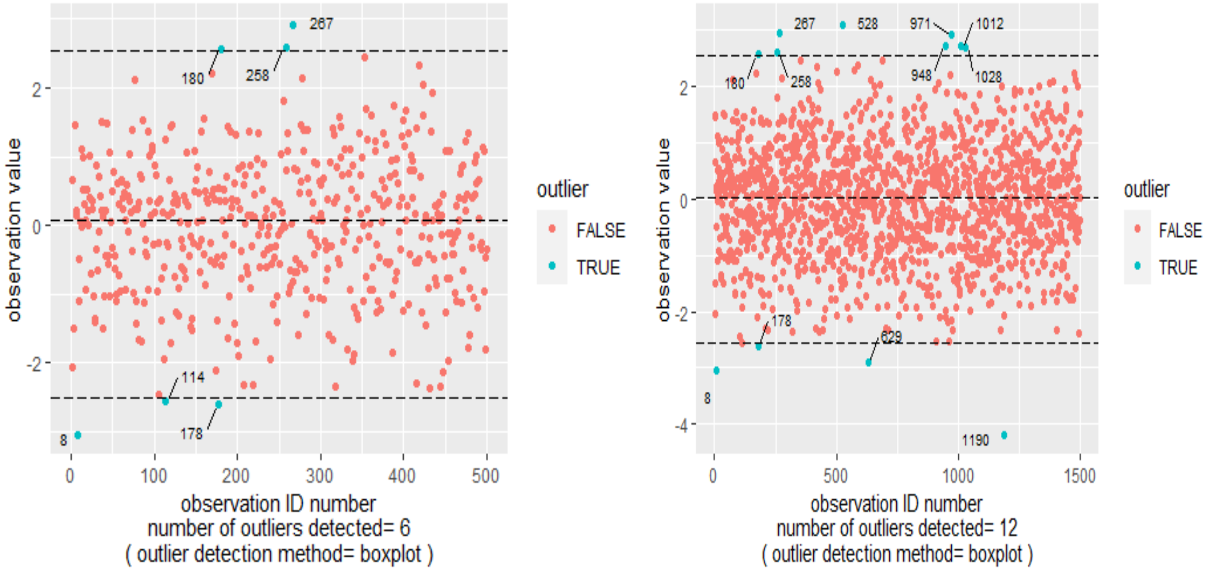

Fig. 8: Outlier detection output for Box plot method.

The findings of this research were summarized in Table 6 and we therefore recommend any of the median methods (median, two absolute median deviation and three absolute deviation methods) or the standard deviation methods (two standard deviation and three standard deviation methods) as the best methods for detecting outliers in a univariate time series data that follows the normal distribution. This means that the median or standard deviation methods of detecting outlier should be considered often than box plot and the z score methods. However, it is also clear that for both sample sizes and based on the best method detected by the comparison methods, the median method of detecting outlier is better than the standard deviation method.

4 Conclusions

The summary of the findings is displayed using Table 6 and the best methods with methods affected by sample size variations are also shown.

Table 6 Summary of the findings.

Sample Size	Three recommended methods				Best Method	Methods affected by sample size
	SEM	CV	MSE	PMSR	Overall	
500 (Small sample)	The median method. Two median absolute deviation method. Two standard deviation method.	Box plot method. Three median absolute deviation method. 3)Three standard deviation method	Z-score. Three standard deviation. 3) Three median absolute deviation rule.	Two standard error. Two median absolute deviation. 3)Median method	Any of the median methods or the standard deviation methods is recommended	Three standard deviation and z-score methods were detected to be affected by sample sizes using SEM. Box plot and two standard deviation methods were detected to be affected by sample size using CV. 3) z-score and median methods were detected to be affected by sample sizes using MSE.
1,500 (Large sample)	The median method. Two median absolute deviation method. Two standard deviation method.	Two standard deviation. Three median absolute deviation. 3)Three standard deviation methods	Median. Two median absolute deviation. 3)Three median absolute deviation	Median method. Two median absolute deviation method. 3)Two standard deviation method	Any of the median methods or the standard deviation methods.	

Availability of data and material

The data used in this research work was obtained through simulation in R program.

Code availability

The R programming codes used in this research are available in Appendix after the References. The program requires specific R libraries to operate. Some of the basic libraries used in the program are "ggplot2", "ggrepel" and "dplyr" packages.

Consent for publication

All authors consented to the publication of the research work.

Conflict of interest statement

The authors stated that there was no commercial or financial relationship that may be considered as a potential conflict of interest during the research.

Author contribution statement

EMN conceived the idea and drafted the first manuscript, BDC wrote the R program using R software version 4.10, BDC and OCP wrote the second draft of the manuscript while KLC reviewed the second draft of the manuscript. All authors certified the final manuscript.

Funding statement

The authors did not receive support from any organization for the work.

References

[1] F.E. Grubbs, "Procedures for detecting outlying observations in samples", *Technometrics*., **11**, 1-21, (1969).

[2] A.J. Fox, "Outliers in time series", *Journal of the Royal Statistical Society Series*, B **34**, 350–63, (1972).

[3] R.S. Tsay, "Time series model specification in the presence of outliers", *Journal of the American Statistical Association.*, **81**, 131–41, (1983).

[4] R.S. Tsay, "Outliers, level shifts and variance changes in time series", *Journal of Forecasting*., **7**, 1–20, (1988).

[5] I. Chang, G.C. Tiao and C. Chen, "Estimation of time series parameters in the presence of outliers". *Technometrics*., **30**, 193–204, (1988).

.[6] C. Chen and L.M. Liu, "Joint estimation of model parameters and outlier effects in time series". *Journal of the American Statistical Association.*, **88**, 284–97, (1993).

[7] B. Abraham & G. E. P. Box, "Bayesian analysis of some outlier problems in time series", *Biometrika*., **66**, 229 -236, (1979).

[8] A.F.M. Smith and M. West, "Monitoring renal transplants: an application of the multiprocessor Kalman filter", *Biometrics*., **39**, 67–78, (1983).

[9] R.E. McCulloch and R.S. Tsay, "Bayesian analysis of autoregressive time series via the Gibbs sampler", *Journal of Time Series Analysis*., **15**, 23–50, (1994).

[10] G. Barnett, R. Kohn and S. Sheather, "Bayesian estimation of an autoregressive model using Markov chain Monte Carlo", *Journal of Econometrics.*, **74**, 237–54, (1996).

[11] K.I. Penny and I.T. Jolliffe, "A comparison of multivariate outlier detection methods for clinical laboratory safety data", *The American Statistician*., **50**, 295-308, (2001).

[12] A. Pimpan, and S. Prachoom, "*A Comparative Study of Outlier Detection Procedures in Multiple Linear Regression*", Proceedings of the International Multi-Conference of Engineers and Computer Scientists., **1** (2009).

[13] S. Hekimoglu, R.C. Erenoglu and J. Kalina, "Outlier Detection by Means of Robust Regression Estimators for Use in Engineering Science", *Journal of Zhejiang University of Science A*., **10**, 909-921, (2009).

[14] E.E. Moawad, "An Alternative Outliers Detection Procedure in Linear Regression Analysis: A Comparative Study", *International Journal of Mathematics and Statistics*., **7**, 353-355, (2013).

[15] S.J. Deutsch, J.E. Richards and J.J. Swain, "Effects of a single outlier on ARMA economic problems", *Journal of the American Statistical Association.*, **70**, 70–79, (1975).

[16] J. Ledolter, "The effect of additive outliers on the forecasts from ARMA models", *International Journal of Forecasting*., **5**, 23–40, (1989).

[17] C. Chen and L.M. Liu, "Forecasting time series with outliers", *Journal of Forecasting*., **12**, 13–35, (1993).

[18] G. E. P. Box. and G.C. Tiao, "Intervention analysis with application to environmental and Environmental Problems", *Journal of American Statistical Association*., **70**, 70-79, (1975).

[19] G.M. Ljung, "A note on the estimation of missing values in time series. Communications in Statistics", *Simulation Computation.*, **18**, 45–65, (1989).

[20] S. Beveridge, "Least square estimation of missing values in time series", *Communications in Biometrika.*, **66**, 229–36, (1992).

[21] O.I. Shittu "Accommodation of Outliers in Time Series Data: An Alternative Method", *Asian Journal of Mathematics and Statistics*, (2008).

[22] K. Ahmet, "Statistical Modeling for Outlier Factors", *Ozean Journal of Applied Sciences.*, **3**, 1943–2429, (2010).

[23] K. Regina and M. Agustin, "Seasonal Outliers in Time Series", Partially supported by the Spanish grant, PB950299 of CICYT, (2001).

[24] H.S. Mehmet, “A New Algorithm for Detecting Outliers in Linear Regression”, *International Journal of Statistics and Probability.*, **2** (2013).

[25] Z. Arimiyaw, K. Nathaniel, B. Howard and N. Kwao, “On the detection of influential outliers in linear regression analysis”, *American journal of theoretical and applied statistics.*, **3**, 100–106, (2014).

[26] R.E. Shiffler, “Maximum Z Scores and Outliers", *The American Statistician.*, **1**, 79-80, (1988).

[27] B. Iglewicz & D. Hoaglin, “How to detect and handle outliers”, in Mykytka, E.F., Eds., vol.16, *ASQC Quality Press.*, Milwaukee, (1993).

[28] T.W. Tukey, “Exploratory Data Analysis”. series in behavioral science, Frederick Mosteller, Eds., *Addison-Wesley.*, Boston, (1977).

[29] R. Ratcliff, “Methods for dealing with reaction time outliers”, *Psychological Bulletin.*, **114**, 510–532, (1993).

Appendix (R 4.10 version)

```
set.seed(234)
y <- rnorm(1500,0,1)
outlierDetection <- function (x,method="2SD",addthres=FALSE){
  if (method=="2SD") {
    avrg <- mean(x)
    stdev <-sd(x)
    midp <<- avrg
    lower <<- avrg-2*stdev
    upper <<- avrg+2*stdev
    dtf <<- data.frame(ID=seq.int(length(x)), obs=x, outlier= x > upper | x < lower)
    outliern <<- length(which(dtf=="TRUE"))
    mydtfsub <- filter(dtf, dtf$outlier == "TRUE")
    View(mydtfsub$obs)
    print(upper)
    print(lower)
    mean_outlier <- mean(mydtfsub$obs)
    sd_outlier <- sd(mydtfsub$obs)
    standard_error_of_mean <- (sd_outlier/sqrt(outliern))
    coefficient_of_variation <- (sd_outlier/mean_outlier)
    print(list("standard_error_of_mean" = standard_error_of_mean, "coefficient_of_variation" = coefficient_of_variation, "number of
outliers detected" = outliern))
  } else {}

  if (method=="3SD") {
    avrg <- mean(x)
    stdev <-sd(x)
    midp <<- avrg
    lower <<- avrg-3*stdev
    upper <<- avrg+3*stdev
    dtf <<- data.frame(ID=seq.int(length(x)), obs=x, outlier= x > upper | x < lower)
    outliern <<- length(which(dtf=="TRUE"))
    mydtfsub <- filter(dtf, dtf$outlier == "TRUE")
    View(mydtfsub$obs)
    print(upper)
    print(lower)
    mean_outlier <- mean(mydtfsub$obs)
    sd_outlier <- sd(mydtfsub$obs)
    standard_error_of_mean <- (sd_outlier/sqrt(outliern))
    coefficient_of_variation <- (sd_outlier/mean_outlier)
    print(list("standard_error_of_mean" = standard_error_of_mean, "coefficient_of_variation" = coefficient_of_variation,"number of
outliers detected" = outliern))
  } else {}
  if (method=="zscore") {
    avrg <- mean(x)
    stdev <-sd(x)
    midp <<- avrg
    statistic <- abs((x-avrg)/stdev)
    dtf <<- data.frame(ID=seq.int(length(x)), obs=x, outlier= statistic > 3)
    outliern <<- length(which(dtf=="TRUE"))
    mydtfsub <- filter(dtf, dtf$outlier == "TRUE")
    View(mydtfsub$obs)
    mean_outlier <- mean(mydtfsub$obs)
    sd_outlier <- sd(mydtfsub$obs)
    standard_error_of_mean <- (sd_outlier/sqrt(outliern))
    coefficient_of_variation <- (sd_outlier/mean_outlier)
    print(list("standard_error_of_mean" = standard_error_of_mean, "coefficient_of_variation" = coefficient_of_variation, "number of outliers detected" = outliern))
  } else {}
  if (method=="modzscore") {
    avrg <- mean(x)
    stdev <-sd(x)
    midp <<- median(x)
    MD <- median(abs((x-midp)))
    statistic <- ((0.6745*(x-midp)/MD))
    upper <-  3.5
    dtf <<- data.frame(ID=seq.int(length(x)), obs=x, outlier= abs(statistic) > 3.5)
    outliern <<- length(which(dtf=="TRUE"))
    mydtfsub <- filter(dtf, dtf$outlier == "TRUE")
    View(mydtfsub$obs)
    mean_outlier <- mean(mydtfsub$obs)
    sd_outlier <- sd(mydtfsub$obs)
    standard_error_of_mean <- (sd_outlier/sqrt(outliern))
    coefficient_of_variation <- (sd_outlier/mean_outlier)
    print(list("standard_error_of_mean" = standard_error_of_mean, "coefficient_of_variation" = coefficient_of_variation, "number of outliers detected" = outliern))
  } else {}
    if (method=="median") {
    med <- median(x)
    MAD <-median(abs(x -med))
    dtf <<- data.frame(ID=seq.int(length(x)), obs=x, outlier=abs(x-med)>2*(MAD/0.6745))
    midp <<- med
    lower <<- med-2*(MAD/0.6745)
    upper <<- med+2*(MAD/0.6745)
    outliern <<- length(which(dtf=="TRUE"))
    mydtfsub <- filter(dtf, dtf$outlier == "TRUE")
    View(mydtfsub$obs)
    print(upper)
    print(lower)
    mean_outlier <- mean(mydtfsub$obs)
    sd_outlier <- sd(mydtfsub$obs)
    standard_error_of_mean <- (sd_outlier/sqrt(outliern))
    coefficient_of_variation <- (sd_outlier/mean_outlier)
    print(list("standard_error_of_mean" = standard_error_of_mean, "coefficient_of_variation" = coefficient_of_variation, "number of outliers detected" = outliern))
   } else {}
  if (method=="2MADe") {
    med <- median(x)
    MAD <-median(abs(x-med))
    MADe <- 1.483*MAD
    lower <<- med-2*(MADe)
    upper <<- med+2*(MADe)
    dtf <<- data.frame(ID=seq.int(length(x)), obs=x, outlier= x < lower| x > upper)
    midp <<- med
    outliern <<- length(which(dtf=="TRUE"))
    mydtfsub <- filter(dtf, dtf$outlier == "TRUE")
    View(mydtfsub$obs)
    print(upper)
    print(lower)
    mean_outlier <- mean(mydtfsub$obs)
    sd_outlier <- sd(mydtfsub$obs)
    standard_error_of_mean <- (sd_outlier/sqrt(outliern))
    coefficient_of_variation <- (sd_outlier/mean_outlier)
    print(list("standard_error_of_mean" = standard_error_of_mean, "coefficient_of_variation" = coefficient_of_variation, "number of
outliers detected" = outliern))
  } else {}
  if (method=="3MADe") {
    med <- median(x)
    MAD <-median(abs(x-med))
    MADe <- 1.483*MAD
    lower <<- med-3*(MADe)
    upper <<- med+3*(MADe)
    dtf <<- data.frame(ID=seq.int(length(x)), obs=x, outlier= x < lower| x > upper)
    midp <<- med
    outliern <<- length(which(dtf=="TRUE"))
    mydtfsub <- filter(dtf, dtf$outlier == "TRUE")
    View(mydtfsub$obs)
    print(upper)
    print(lower)
    mean_outlier <- mean(mydtfsub$obs)
    sd_outlier <- sd(mydtfsub$obs)
    standard_error_of_mean <- (sd_outlier/sqrt(outliern))
    coefficient_of_variation <- (sd_outlier/mean_outlier)
    print(list("standard_error_of_mean" = standard_error_of_mean, "coefficient_of_variation" = coefficient_of_variation, "number of
outliers detected" = outliern))
  } else {}
  if (method=="boxplot") {
    Q1 <- quantile(x, 0.25)
    Q3 <- quantile(x, 0.75)
    IntQ <-Q3-Q1
    dtf <<- data.frame(ID=seq.int(length(x)), obs=x, outlier=x<Q1-1.5*IntQ | x>Q3+1.5*IntQ)
    midp <<- median(x)
    lower <<- Q1-1.5*IntQ
    upper <<- Q3+1.5*IntQ
    outliern <<- length(which(dtf=="TRUE"))
    mydtfsub <- filter(dtf, dtf$outlier == "TRUE")
    View(mydtfsub$obs)
    print(upper)
    print(lower)
    mean_outlier <- mean(mydtfsub$obs)
    sd_outlier <- sd(mydtfsub$obs)
    standard_error_of_mean <- (sd_outlier/sqrt(outliern))
    coefficient_of_variation <- (sd_outlier/mean_outlier)
    print(list("standard_error_of_mean" = standard_error_of_mean, "coefficient_of_variation" = coefficient_of_variation, "number of
outliers detected" = outliern))
  } else {}
  if (addthres==TRUE) {
    p <- ggplot(dtf, aes(x=ID, y=obs, label=ID)) + geom_point(aes(colour=outlier)) + geom_text_repel(data = subset(dtf,
outlier=="TRUE"), aes(label = ID), size = 2.7, colour="black", box.padding = unit(0.35, "lines"), point.padding = unit(0.3, "lines")) +
labs(x=paste("observation ID number\n number of outliers detected=", outliern, "\n( outlier detection method=", method, ")"),
y="observation value") + geom_hline(yintercept = midp, colour="black", linetype = "longdash") + geom_hline(yintercept = lower,
colour="black", linetype = "longdash") + geom_hline(yintercept = upper, colour="black", linetype = "longdash")
  } else {
    p <- ggplot(dtf, aes(x=ID, y=obs, label=ID)) + geom_point(aes(colour=outlier)) + geom_text_repel(data = subset(dtf,
outlier=="TRUE"), aes(label = ID), size = 2.7, colour="black", box.padding = unit(0.35, "lines"), point.padding = unit(0.3, "lines")) +
labs(x=paste("observation ID number\n( outlier detection method=", method, ")"), y="observation value") #requires 'ggrepel'
  }
  return(p)
}
outlierDetection(x, method ="modzscore", addthres=T)
```