MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Review Article

Open Access

CrossMark

A review on the advances of outliers

Abstract

This study x-rays the perspectives, sources, types, incidence and consequences of outliers on some statistical methods and applications. Measures for detection, methods of handling/ testing and consequences of outliers are highlighted. Controversies and remedies associated with estimating or discarding outliers in their distributional or structural form are discussed.

Keywords: incidence, sources, detection, tests, methods.

Volume 12 Issue 1 - 2023

Ogoke Uchenna Petronilla, Nduka Ethelbert Chinaka

Department of Mathematics and Statistics, University of Port Harcourt, Nigeria

Correspondence: Ogoke Uchenna Petronilla, Department of Mathematics and Statistics, University of Port Harcourt, Nigeria; Email uchenna.ogok@uniport.edu.ng

Received: December 05, 2022 | **Published:** January 25, 2023

Outline

Concept of outliers

An experimenter may come across a batch of data that comprises a few observations that deviate from the pattern displayed by the bulk of the values. Outliers or wild shots are extremely questionable observations in a sample data. Several criteria have been proposed for finding outliers so that they can be investigated and potentially removed from the data. Outliers are defined as values that are greater than two and a half standard deviations from the mean, i.e. values below or over $\bar{x} \pm 2s$. When the distribution is normal, this criterion is probably the best fit. We know that when data is skewed, the median is a stronger indicator of central tendency than the mean.

For such skewed data, the criterion is to define outliers as any values outside the interval $\bar{x} \pm 2(x_{0.75} - x_{0.25})$ where $x_{0.75}$ is the 75^{th} percentile and $x_{0.25}$ is the 25^{th} percentile. This criterion, however, is sensitive to kurtosis. A more robust criterion therefore defines outliers as values are greater than $x_{0.75} + 1.5(x_{0.75} - x_{0.25})$ or less than $x_{0.25} + 1.5(x_{0.75} - x_{0.25})$.

Most empirical data sets include a certain amount of exceptional values or, some observations which deviate from the general trend to a certain degree. Such values are generally termed as "outliers." Sometimes an experimenter encounters a set of data, of which contains a few observations that fall outside the pattern exhibited by majority of the values. Such highly suspect observations in a sample data are known as 'Outliers or Wild Shots'. Data collation and analysis is apparently the core of the field of statistics. These data are collected and analysed with specific intent in the mind of the statisticians, some of them include; Survey studies, Time Series analysis, Experimental design, probability studies etc. There is a significant demand for precision to attain any of these aims, hence statisticians place a premium on the correctness of both the data and the method used to analyse the data. However, in practise, some observations appear to be intuitively incorrect following data collection. Bearing in mind the fragility and need for fidelity the statistician is under compulsion to take note of such observation. Questions as to whether such observations are "Bad data" will naturally become of necessity. What really worries is whether or not such observations are genuine members of the main population. If they are not, they may frustrate attempts to draw accurate conclusions from the data set. A suitable definition for outliers was given by Hawkins[1] his definition states. "Outliers can be seen as an observation(or subset of observations) which appears to be inconsistent so much from other observations as to arouse suspicions that it was generated by a different mechanism". He also defined Outliers simply to be extreme values that deviate from other observations in data. They may indicate a variability in a measurement, experimental errors or novelty. In other words, an outlier is an observation that lies in an abnormal distance from other values in a random sample from a population. The conclusion as to whether some set of the values will be picked out for scrutiny or not rest on the observer. Opinion is divided on precisely when it is justifiable to scrutinize suspected outliers.

Practically, nearly all experimental data samples are subject to contamination by outliers which theoretically reduce the reliability and efficiency of statistical methods. As an instance, consider an observation in which its standardized residual is relatively large compared to other observations in the data set, such observation is considered an outlier that lies at a distance from the rest of the data. Hence, the presence of outliers, which are data points that deviate remarkably from others, is one of the most trying methodological challenges in organizational and scientific research.[2] Outliers, according to Aggarwal,[3] can be characterised as noise points that lie outside of a set of defined clusters, or as points that lie outside of the set of clusters but are also isolated from the noise. These outliers act in ways that depart from the norm.

Historical perspectives

Outlier detection has been used to find and delete deviating observations from data for many years. When Boscovich attempted to estimate the elliptical form of the world in the mid-eighteenth century, he discovered the existence of outliers. He obtained ten distinct measurements, two of which were rejected due to their excessive readings, and then averaged the remaining eight data.[4] From then, several authors have contributed their own idea on possible remedy to this problem. The first published objectives test for anomalous observations was from the American astronomer Peirce[5]. Next author was also an American Astronomer Chauvenet[6] whose test has an attractive simplicity. Stone[7] published next in 1868 on the concept of a modulus of carelessness. An observation is to be discarded if it can be attributed more to the observer mistake than random variation. After this scientist began to consider alternatives to outright rejection of data. They did this by creating what can be considered to be a robust procedure for estimating a location parameter which secures the accommodation of outliers with the aid of weighting of observation. The first to publish a work with this concept was Glaisher[8]. This method developed by Glaisher had a quick criticism by Stone[9] in 1873 who also developed an alternative weighting procedure. Others who contributed to the robust procedure method where Edgeworth,[10] Newcomb,[11] Stigler[12]. However, some important procedure where still developed which supported outright rejection of outliers. Wrights' procedure[13] rejects any observation deviating from the mean more than three times the standard deviation, equivalently five times the probability error; Wright and Hayford[14] made some modification to Wright's procedure[13] and Goodwin's procedures.[15] The problem outliers is an age long and an unavoidable problem. Its prevalence in many branches of research has made it a relevant area of study.

Outliers can occur in any field that collects data and draws conclusions from it. Outliers can occur in any data set, including univariate, multivariate, and data from more organised statistical analysis such as regression. Outliers occur as a result of mechanical flaws, fraudulent activity, human error, changes in system behaviour, instrumental errors, or just natural variances in observed populations. Hence, their detection can discover system flaws and fraud before they have potentially disastrous implications. This type of detection can also discover errors and reduce their contaminating effect on the data collection, so purifying the data for processing. Previously, outlier detection approaches were haphazard, but now, logical and systematic methodologies derived from the entire spectrum of Computer Science and Statistics are applied.[16]

Sources of outliers

Unknown data structures and correlations can cause apparent outliers. 'Extreme' observations may also be due to sampling-related sources of error. Some subjects in an experiment, for example, may not be following the proper procedures. If observations are automatically recorded, it is possible that the recorder did not function properly for some of the observations. So human error in recording or negligence can be a source. However, incorrect assumptions about the data distribution can lead to them being labelled as outliers.[18]

Nevertheless, outliers exist in almost every real data set which the sources can be itemized and identified as follows:

i. Human Error: In experiments there are some observations that are clearly erroneous. Some of them may come from fatigue from human, poor recording, an automobile accident or a data reporting error, etc. All these errors are feasible in experiment and are sources of outliers.

ii. Instrument error

iii. Incorrect distribution assumption

iv. Natural deviations in populations,

v. Unknown data structure or just an incidental phenomenon

vi. Fraudulent behavior, and changes in behavior of systems or faults in systems.

vii. Malicious activity such as insurance or credit card or telecom fraud, a cyber intrusion, or a terrorist activity.

viii. Measurement Error: This is the most common cause of outliers. It occurs when the measuring tools employed become faulty.

ix. Intentional Error: These are dummy outliers made to test detection methods. It's commonly found in self-reported measurements that involve sensitive data.

x. Data Processing Error: This kind of error occurs during manipulation of data or data set unintended mutations.

xi. Sampling Error: These outliers occur during extraction or mixing of data from wrong or various sources.

xii. Instrumentation error such as defects in components of machines or wear and tear.

xiii. Change in the environment such as a climate change, a new buying pattern among consumers, and mutation in genes.

xiv. Outliers In Relation to Probability Models: Observations come from several probability models, some observations which are outliers in a certain probability model are clearly random variables from other models. As such note that observations are outliers in relations to a specific model. In a case of a wrong model wrong alarms of outliers could be raised.

xv. Natural Error: These types of outliers are normally called novelties because they are not products of errors.

xvi. Execution Error: Imperfect collection of our data maybe another source of unexpected variability in our data set. We may inadvertently choose a biased sample or include individuals/entities who may not truly be representatives of the population we aimed to sample. Sensible precautions may reduce such unwanted variability.

Types of outliers

Outliers are divided into three types based on their composition and relationship to the remainder of the dataset.

These are the classifications:

1. Individual outlier: This is referred to as a Type I outlier: A Type I outlier, for example, is a single outlying point in a given group of data points. This is the most basic type of outlier. The strategies for detecting Type I outliers examine the relationship of an individual point or instance to the rest of the data instances in the training or test datasets.

2. Type II Outlier: These are outliers induced by the appearance of a single data point in a specific context in the provided dataset. The context specifies the immediate surroundings of a specific data point or instance that is induced by the structure of the dataset and must be stated as part of the problem formulation. Type II outliers satisfy the property that the underlying data is spatial/sequential in nature, which means that each data point

is described as either contextual or behavioural qualities. The contextual properties define a data point's or instance's position and are used to identify the context for that instance. Contextual attributes in geographic data sets, for example, are a location's longitude and latitude. The behavioural attributes define an instance's non-contextual qualities. A behavioural property, for example, in a spatial data set describing the amount of rainfall at any location.

3. Type III Outliers: These arise when a subset of data points deviates from the overall data set. Individual data points in a Type III outlier are not outliers in and of themselves, but their occurrence as a substructure is unusual. Only when the data is spatial or sequential are type III outliers meaningful. These outliers are either diverging subgraphs or subsequences in the dataset.[19]

Consequences of outliers and its effects on analysis

When we construct confidence intervals on some parameters, such as the mean, from a relatively small sample in the presence of outliers, the resulting confidence bounds are likely to be unduly wide. Hence, there is a real need for procedures, which are efficient in guarding against the influence of outliers in relatively small samples. When we generate confidence intervals on some parameters, such as the mean, from a relatively small sample in the presence of outliers, the resulting confidence bounds are more likely to be excessively large. Hence, there is a real need for procedures, which are efficient in guarding against the influence of outliers in relatively small samples.[17] Depending on the number of outliers, eliminating an outlier reduces the degree of freedom in the analysis.

Consider the impacts of outliers on the following statistical techniques:

a. Bioassay

Data editing with elimination of "outliers" is commonly performed in the biomedical sciences. Hence, the effects of this type of data editing could influence study results, which results to misreporting and error analysis results. Also with the vast and expanding amount of research in medicine, these effects would be magnified.[18] The fundamental problem in outlier handling in bioassays is that root causes often cannot be determined with hindsight. It is uncertain if an extreme data point is the result of a technical malfunction or error, or of the assay's inherent variability. Many subsequent handling processes, for example, such as pipetting, are prone to outliers, and such experiments are used to confirm or reject a hypothesis, characterise a process, or evaluate the performance of a system. For the illustration, see Ogoke and Nduka,[19] Statistical Theory and Analysis in Bioassay https://www.amazon.com)

b. Correlation

An outlier will weaken the correlation, scattering the data and bringing the coefficient of determination, R^2, closer to zero. Hence removing the outlier raises the R^2 value (stronger correlation because the data is less scattered), which can have a considerable impact on the occurrence of type 1 error.

c. Regression

In regression analysis, an outlier is an observation with large residual. It is an observation whose dependent variable is unusual given its value on the predictor variables. When outliers are present in data analysis of Ordinary least squares estimation, it gives very misleading result. The problems of multiple linear regression models also arise when there is an outlier in the data. Because outliers can distort estimates of regression coefficients, and can produce misleading results, and the interpretation of the results may be in doubt.[20]

The existence of outliers and important examples can significantly alter the magnitude and direction of regression coefficients (i.e., from positive to negative or vice versa). When researchers disregard aberrant findings, particularly when it comes to dependent variables, their empirical results can be misleading. Because hypotheses frequently describe positive or negative causal links between dependent and independent variables, presenting estimated coefficients without outlier diagnostics might lead to inaccurate inferences and jeopardise otherwise useful scientific discoveries.[21]

d. T-test

Because the computations frequently rely on squared deviations from the mean, statistical inferential tests can be sensitive to outliers. Therefore, when comparing Means in a Student-t test, outliers may skew parameter estimates such as the mean and variance and lead to inaccurate test findings.

e. Time series

Certain forms of outliers are intrinsically more detrimental to parameter estimate and future forecasts for time series data than others, regardless of their frequency. Several academics have explored forecasting analysis utilising time series approaches ranging from the linear regression model to advanced techniques such as ARIMA, ARCH, and GARCH. The projected values have been estimated using these strategies. Consequently, identifying these outliers is a difficult process that may result in improved performance when estimating projected values.[22] Additive Outliers (AO), Innovational Outliers (IO), Temporary Change, and Level Shift are the most commonly encountered outlier types in time series. The additive outlier, also known as the Type I outlier, affects only a single observation that is either lower or greater than the predicted values in the data. After this disruption, the series resumes its regular course as if nothing had occurred. The influence of an additive outlier is bounded and independent of the ARIMA model. However, unlike the AO, innovational outliers are classified as Type II outliers, which influence many observations. A single residual is affected by an AO at the date of the outlier. The effect of the IO on an observed series consists of an initial shock that propagates with the weights of the moving average (MA) representation of the ARIMA model to subsequent observations.[22]

f. Quality control

Control charts have been widely utilised in fields such as manufacturing, public health, and financial services to analyse process performance and identify and investigate faults. For example, in the bottling of beer, data such as bottle weight is typically gathered and monitored to ensure that the process is working as planned and to detect underfilled or overfilled bottles.[23]

Detection of outliers

There have been numerous statistical tests and procedures proposed for detecting and rejecting outliers. However, the real-world relevance of outliers is a distinguishing characteristic of outlier identification that sets it apart from noise reduction and noise accommodation in the data. Noise in data has no practical relevance in and of itself, yet it hinders data analysis. The necessity of removing unwanted objects

prior to performing data analysis motivates noise removal. In outlier detection, the developer should choose an algorithm that is appropriate for their data set in terms of the correct distribution model, the speed, the correct attribute types, the scalability, any incremental capabilities that permit new exemplars to be stored, and the modelling precision.

Again, Quartiles can be used to create another measurement of variability, the interquartile range, which is defined as $H = Q_3 - Q_1$ This is used to measure the spread of the middle 50 % of the observations. Large values of this statistic indicate that the first and third quartiles are far apart, indicating a high level of variability.

Also, visual inspection of scatter plots and box-whisker-plots for detecting outliers is the most common approach to outlier detection. Different graphical procedures such as plot of residuals versus fitted values, plot of leverages versus standardized residuals are also used to detect outliers and leverage points.

Outliers can also be detected using Information Criteria. Furthermore, procedures that are resistant to the distorting influence of outliers—generally known as robust methods—are required when identifying outliers. A k-fold cross-validation is performed to assess the procedure's reliability and validity, as well as the results concerning the discovered outliers and in-control data.

Remedies and controversies

Outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behavior. These deviating patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains.

Consequently, statistical outlier detection techniques are essentially model-based techniques; i.e., they assume or estimate a statistical model which captures the distribution of the data, and the data instances are evaluated with respect to how well they fit the model. If the probability of a data instance to be generated by this model is very low, the instance is deemed as an outlier.

i. Data Trimming: One approach to dealing with this scenario is to eliminate the extreme observations, which is known as data trimming (or screening or cleaning). This process removes an equal number of lower or higher observations from the data, and the resultant reduced or (trimmed, screened, or cleaned) sample is used as the sample data-with a suitable adjustment made to correct the usual sampling distribution for the effects of the trimming.

ii. Winsorization: This is a way for dealing with extremely aberrant outliers. The high and low extreme values are replaced by the next to the highest and next to the lowest extremes in the simplest instance of winsorization ($g = 1$). The generated data is treated just like the original data. With $g = 2$, the highest two extremes are each replaced by the third highest, and a similar replacement is made at the ordered sample data's lower end. The technique is the same for $g = 3$, etc.[16]

If g represents the number of elements in the basic sample that have been replaced at either extreme under winsorization, then the number of observations remaining is $n - 2g$. Let $k = n - 2g$. If the sample observations are in order of magnitude, we have $x_1,\ldots,x_g, x_{g+1}\ldots x_{g+k},x_{g+k+1},\ldots,x_n$. The winsorized data would have the form $x_{g+1},\ldots x_{g+1},x_{g+1}\ldots x_{g+k}, x_{g+k},\ldots x_{g+k}$. A $100(1-\alpha)\%$ confidence interval for μ based on winsorized sample data has the form. A $100(1-\alpha)\%$ confidence interval for μ based on winsorized sample data has the form,

$$\Pr[[\bar{x} - A \le \mu \le \bar{x} + A] = 1-\alpha \text{ where } A = \frac{n-1}{k-1} t_{1-\alpha/2}(k-1) S_{\bar{x}}$$

; $s_x^2 = \frac{\sum_i x_i^2 - nx^{-2}}{n(n-1)}$

and $S_{\bar{x}} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n(n-1)}}$

Note: The typical t-statistic computed from the winsorized data behaves, to a satisfactory approximation, like student's with modified degrees of freedom when the basic population from which the sample data were drawn is normal. Thus, the processes for performing tests or establishing confidence intervals on population parameters are similar to those used with the standard t-test.

i. Range Test: When some assignable cause is discovered for the outlier being different from other observations, it may safely be discarded. If no obvious cause comes to light, a statistical test should be applied to confirm that a suspected outlier is as extreme as it appears. A number of these tests have been devised. One such test is the range test which is the quickest to calculate, and is given by,

$$R_T = \frac{ExtremeValue - OverallMean}{OverallS\tan dardDeviations}$$

We reject the outlier if its ratio exceeds a certain critical value, otherwise accept the outlier.

The hypothesis of interest would be,

H_0 : There is no outlier

H_1 : There is the presence of outlier

ii. Data transformation: Data transformation is one method for mitigating the impact of outliers because the two often used expressions, square root and logarithms, alter larger values considerably more than smaller values. Taking the log of a variable modifies the link between the original variable and the other variables in the model, which is more than just making the distribution less skewed.

Using log transformation i.e. $\log_e x$, this large data values can be modified to smaller values which are: 9.990, 10.017, 10.012, 10.056, 10.094, 10.123, 10.139, 10.158, 10.162 and 10.200.

iii. Construction techniques: A plot of values, with one variable an independent variable on the x-axis, and a dependent variable on the y-axis, can be constructed to identify potential outliers. Hence, the construction of Scatter plots can be used to identify a potential outlier by a data point lying far away from the centroid of data. Again, a Box plot can also be used to identify those points that lie beyond the plot's whiskers, that is, the smallest and largest values, excluding the outlier values. Box plot depicts a summary of the smallest values of a construct above the lower quartile (Q1), the median (Q2) and the largest values below the upper quartile (Q3), excluding outliers. Thus, this identified outliers can then be removed from the dataset.[2]

Recent advances

It has been discovered that outlier detection is directly applicable in a large variety of fields. This has led to a vast and quite diverse body

of work on outlier detection strategies. Several of these strategies were designed to handle specific difficulties in a particular application domain, while others were developed in a more general fashion. Numerous forms of the outlier identification problem have been investigated in several fields, including statistics, machine learning, data mining, information theory, and spectral decomposition. Recent developments in dealing with outliers include the following:

Tracking activity - identifying mobile phone fraud via monitoring phone activity or suspicious stock trades.

Monitoring the performance of computer networks, for example, to detect network bottlenecks, is an example of network performance monitoring.

Fault diagnostic – monitoring methods to discover flaws in, for instance, motors, generators, pipelines, or space shuttle sensors.

Monitoring safety-critical applications over time, such as drilling and high-speed milling. (Time series monitoring)

Detecting novelty in text - to detect the beginning of news stories, for subject detection and tracking, or for traders to identify equities, commodity, and foreign exchange trading stories, as well as outperforming or underperforming commodities.

Detecting unexpected database entries - for data mining to discover errors, frauds, or valid but unexpected entries, as well as detecting mislabeled data in a training data set.[16]

Novel domains like as data mining and machine learning have been employed to address a wide range of outlier-related issues. These approaches can manage massive amounts of data while making significantly fewer assumptions about the data set. In this subject, numerous methods for detecting outliers have been developed, and they are broadly classified into two classes.

Instance based method: In this method a training model is not developed up front. Rather, for a given test instance, one computes the most relevant (i.e., closest) instances of the training data, and makes predictions on the test instance using these related instances. Some important categorise in this method are;

i. K – nearest neighbour detectors (KNN)

ii. Local Outlier Factors (LOF)

iii. Local Correlation Integral (LOCI)

Explicit generalization method: In this method outliers are detected using already trained model. Some important categories in this method are;

i. Principal Component Analysis

ii. Expectation-maximization

iii. Mahala Nobis method

iv. Isolation Trees etc.

Outliers ensemble

Ensemble analysis is a well-known technique for improving the accuracy of various data mining methods. Ensemble methods integrate the results of several algorithms or base detectors to get a unified result. The approach's primary premise is that some algorithms perform well on a specific group of points while others perform better on other subsets of points. However, because of its ability to mix the outputs of numerous algorithms, the ensemble combination is frequently able to perform better robustly across the board.

Outlier ensembles are used in data mining to handle a variety of problems such as clustering, classification, and so on. It is not impossible to recognise outliers. Outlier ensemble has proven to be quite effective in detecting outliers.

There are two types of Outliers ensemble utilised for outlier ensemble:

i. Sequential Ensembles: a given algorithm or set of algorithms are applied sequentially, so that future applications of the algorithms are influenced by previous applications, in terms of either modifications of the base data for analysis or in terms of the specific choices of the algorithms. The final result is either a weighted combination of, or the final result of the last application of an outlier analysis algorithm.

ii. Independent Ensembles: different algorithms, or different instantiations of the same algorithm are applied to either the complete data or portions of the data. The choices made about the data and algorithms applied are independent of the results obtained from these different algorithmic executions. The results from the different algorithm executions are combined together in order to obtain more robust outliers.

Statistical packages

Several statistical packages have been developed to handle outliers. Some are used to detect/ analyse data in large range of analysis while others have been built for specific analysis.e.g. SPSS, Outlier Flag (Time Series), Regression Diagnostic (Experimental Design) etc. Some software's that can perform robust analysis have also been developed e.g. S-PLUS etc. In essence, outlier detection techniques traditionally employ unsupervised learning processes. Many clustering algorithms detect outliers as by-products.

Suggestions

Nevertheless, there are other techniques, methods and procedures that can be used to identify outliers and provide remedies to its incidence which include the use of Information Criteria to detect outliers. In case of univariate data changing the distribution to fit the data set is generally considered a good idea. In cases where the data set do not meet the condition of any distribution, the use of robust method which are immune to outliers can be used. For structured data the use of software's specialized to handle such case is an efficient way to handle them. Outliers can also be handled by deleting the observation if it occurred due to data entry error (human error) and data processing error. Faulty or worn-out Measuring instruments can also be repaired or replaced as the case maybe. Also for extremely large, multi-dimensional data generally complex data the use of data mining techniques is efficient.

Conclusion

Outlier analysis has been studied for several decades for the sake of data cleaning, fraud detection, gaining insights into the hidden patterns, etc. Numerous models and methods have been discussed. to detect outliers either for static data. Therefore, developing techniques to look for outliers and understanding how they impact data analyses are extremely important parts of a thorough analysis especially when statistical techniques are applied to the data. Outlier detection has necessitated either estimating or eliminating them. However, it is more plausible to estimate the correct value rather than discard because the latter reduces the number of degrees of freedom.

Acknowledgments

None.

Conflicts of interest

The authors declared that there are no conflicts of interest.

Funding

None.

References

1. D. Hawkins. Identification of Outliers. *Chapman and Hall*. London:1980.
2. Herman A, Ryan G, Harry J. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods*. 2013;*16*(2):270–301.
3. Aggarwal CC, Yu PS. 'Outlier Detection for High Dimensional Data'. In: *Proceedings of the ACM SIGMOD Conference:*2001.
4. Tara A, Ivan S. Detection and Treatment of Outliers in Data Sets. *Iraqi Journal of Statistical Science.* 2006;(9):58–74.
5. Peirce B. 'Criterion for the rejection of doubtful observations'. *Astr J*. 1852;1(2):161–163.
6. Chauvenet W. 'Method of least squares'. Appendix to Manual of Spherical and Practical Astronomy. 5th ed. New York; USA, 1863.
7. Stone EJ. 'On the rejection of discordant observations'. Monthly Notices. *Roy Astr Soc*. 1868;28:165–168.
8. Glaisher JWL. 'On the rejection of discordant observations'. Monthly Notices Roy. *Astr Soc*. 1872-73;33:391–402.
9. Stone EJ. 'On the rejection of discordant observations'. Monthly Notices. *Roy Astr Soc*. 1873;34:9–15.
10. Edgeworth FY. 'The method of least squares'. *Philosophical Magazine*. 2009;16:360–375.
11. Newcomb S. 'A generalized theory of the combination of observations so as to obtain the best result'. *Amer J Math*. 1886;8:343–366.
12. Stigler SM. 'Simon Newcomb, Percy Daniel, and the history of robust estimation 1885- 1920'. 1. *Amer Statist Ass*. 1973b;68:872–879.
13. Wright TW. *A Treatise on the Adjustment of Observations by the Method of Least Squares. New York: Van Nostrand. 1884.*
14. Wright TW, Hayford JF. Adjustment of Observations. *Van Nostrand*. New York:1906.
15. Goodwin HM. *Elements of the Precision of Measurements and Graphical Methods.* New York: McGraw-Hill. 1913.
16. Victoria J Hodge, Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review.* 2004;22(2)85–126.
17. Nduka EC, Ogoke UP. Principles of Applied Statistics. *Regression and Correlation Analyses*. 2nd ed. 2016.
18. Todd W, James D, Joseph I. Effect of removing outliers on statistical inference: implications to interpretation of experimental data in medical research. *Marshall Journal of Medicine*. 2018;*4*(2):1–18.
19. Ogoke UP, Nduka EC. *Statistical Theory and Analysis in Bioassay*. 2021.
20. Afrah Y. Effect of outliers on the coefficient of determination in multiple regression analysis with the application on the GPA for student. *International Journal of Advanced and Applied Sciences*. 2020;7(10):30–37.
21. Seung Whan C. The effect of outliers on regression analysis: regime type and foreign direct investment. *Quarterly Journal of Political Science.* 2009;4:153–165.
22. Deneshkumar V, Senthamarai K. Outliers in time series data. *International Journal of Agricultural Statistical Science.* 2011;7(2):685–691.
23. Jiayun J, Geert L. Identifying outliers in response quality assessment by using multivariate control charts based on kernel density estimation. *Journal of Official Statistics*. 2021;37(1):97–119.