Determining Outliers in Given Observations using Linear Regression Model

Shubham Kunhare
M Tech 4th (C.S.E.)
S.I.R.T. Indore M.P. India

Sachin Patel, PhD
Asso. Prof., C.S.E. Department
S.I.R.T. Indore M.P. India

ABSTRACT
Outliers is a type of sample which are specially very far from the all other object. There is no scientific definition for an outlier. Defining whether or not an observation is an outlier it is a particular matter. It can also be defined and clarified as a member of data that really differs from the given data set. Outlier discovery is the method of identifying and eliminating outliers from a given set of data. There are no identical approaches are available to outlier, these are mainly reliant upon the data set. Outlier discovery is division of data mining and it has many applications in data study. It is significant to keep outliers in mind when observing at pools of data because they can occasionally affect how the objects are look on the whole. They can extremely change the results of the data analysis and numerical modeling. In the proposed work we found the outliers for linear data set. We used linear regression method to found the outliers. We used the value of correlation coefficient to check correctness of the proposed work. First we compute the value of correlation coefficient with outlier and then also check after eliminating the outlier, how the value is affected.

Keywords
Outlier, Linear Regression, Correlation, Data Mining, Discovery

1. INTRODUCTION
Outlier discovery is one of the significant parts of in data mining. Outliers are objects which can be considered uncommon value due to some reasons. Outlier discovery techniques are used to reduce the impact of outliers in pre-processing phase before the evidence is processed. Outliers are more exciting than the common examples and outliers discovery methods are used to search for them.

No one can identify outliers while gathering data. We will not distinguish what standards are outliers until and except we begin examining the data. So many arithmetical tests are complex to outliers and they have the capability to detect them as a significant part of data analytics. The outlier model has the capability to throw an outlier must some rationale. Occasionally Outliers can be helpful pointers. In some applications of data analytics similar credit card scam discovery, outlier inspection becomes very important because here, the exclusion somewhat the rule may be of fear to the analyst [11,12].

Traditional outlier discovery approaches can be classified into four main groups:

1. Based on Distance
2. Based on Density
3. Based on Clustering
4. Based on Distribution.

Each of these methods has benefits and limits. Traditional outlier discovery approaches are oftennot proper to treat some specific databases. The procedures like artificial intelligence, genetic algorithms, rough set and image processing must use in command to develop new effective outlier's detection approaches.

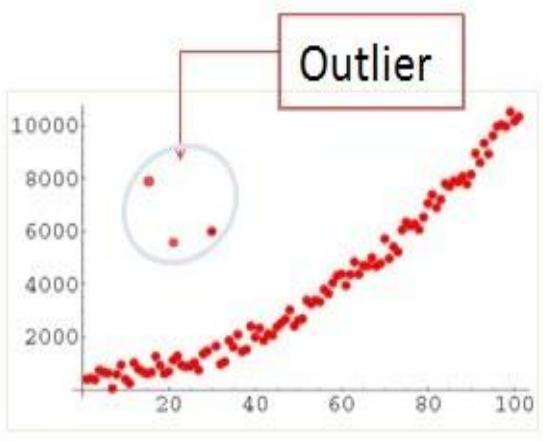


Fig 1 group of outliers

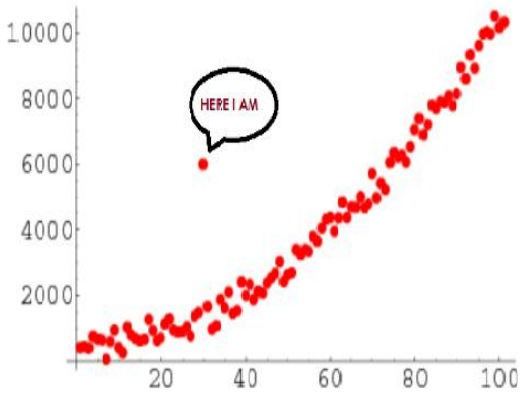


Fig 2 Single Global outlier

2. IMPACTS OF OUTLIERS
There are frequent negative impacts of outliers in the data set[12,13,14]

1. It rises the error adjustment and decreases the power of statistical tests
2. If the outliers distributed non-randomly, they can reduction regularity
3. They can bias or effect approximations that may be of practical interest
4. They can also influence the basic hypothesis of Regression, ANOVA and other numerical model norms.

5. They may basis a important impact on the mean and the normal deviation.
6. They can bias or influence approximations that may be of practical interest.

Outliers are uncommon values in the dataset, and they can deceive statistical examines and disrupt the assumptions. Eliminating outliers is suitable only for precise reasons. Outliers can be very supportive about the subject-area and data gathering process. It's important to recognize how outliers occur and whether they strength occur again as a normal portion of the procedure or study area. Unfortunately, repelling the invitation to remove outliers unsuitably can be difficult. Outlier's growth the inconsistency in data, which decreases statistical power. Consequently, without outliers can reason results to convert statistically important.

3. METHODS TO IDENTIFY OUTLIERS

There are numerous ways to organize outliers in a dataset, following are some of them [9,15,17]

1. Categorization the data
2. Graphical Technique
3. Z score Technique
4. IQR interquartile range
5. Probabilistic and Statistical Technique
6. PCA and LMS
7. Non-parametric Technique

Arranging the dataset is the simplest and actual method to check uncommon value. Scatter plots is used in graphical technique which often have a pattern. We call an object an outlier if it doesn't fit the pattern. There is no distinct rule that expresses us whether or not a point is an outlier in a scatter plot.

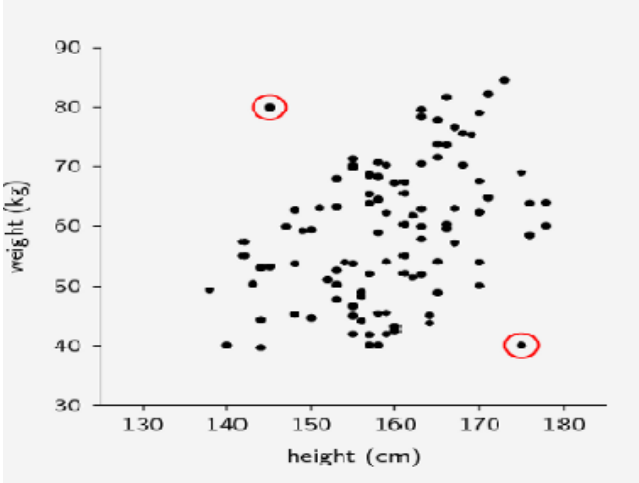


Fig 3 outliers in linear data set

Z-score offers an idea of how several standard deviations missing a data point is from the mean. But more precisely it's a quantity of how many standard deviations under or above the population mean a raw mark is. IQR is just the size of the box in the box-plot which can be used as a quantity of how ranges out the values are. An outlier is any value that lays more than one and partial times the distance of the box from also end of the box.

Probabilistic and Statistical Models undertake specific supplies for data. They make use of the expectancy expansion methods to approximation the constraints of the model. They calculate the probability of affiliation of each data point to intended distribution. The points with a low probability of affiliation are marked as outliers.

4. LITERATURE SURVEY

In 2013 Lakshmi Sreenivasa et al proposed "Outlier Analysis of Categorical Data using Fuzzy AVF". They recommend an algorithm AVF to discover outliers in definite data. The algorithm uses the recurrent pattern data mining method. It avoids difficult of giving k-outliers to get optimal correctness in any organization models in preceding work like Greedy, AVF, FPOF, and FDOD while discovery outliers. The algorithm is practical on UCI ML Source datasets like Nursery, Breast cancer expand dataset by without numerical qualities. The trial results show that it is well-organized for outlier detection in definite dataset. The recommended method gives the best number of outliers KN. In current models it is compulsory to give the number of outliers to find them. While taking the number of outliers occasionally the original data may be missed [1] .

In 2014 VarunChandola et al proposed "Outlier Detection: A Survey". Outlier discovery has been studied within numerous application domains and information disciplines. The survey provides a complete overview of current outlier detection methods by categorizing them along different extents. Categories and subject descriptors H.2.8. They discoursed the different ways in which the problem has been expressed in works. Numerous techniques have been suggested to target a specific request area. The survey can positively allow mapping such existing methods to other request domains. The perception of using a situation to detect Type II outliers has not been totally implicit. Several methods unknowingly have adopted a Type II outlier detection method. Song et al. have exposed that by a framework expands the outlier detection fitness of a method [2].

In 2015 P. Patel et al proposed "A Survey of Outlier Detection". Maximum of real world data set must outliers. Outlier discovery shows a vital role in data mining. They surveyed on dissimilar Outlier discovery methods, which are based on statistical method, deviation approach, distance approach, density. In order to contract with outlier, clustering method is used. K-mean is generally used to cluster the data set before we can relate any method for discovery outliers. They conferred different technique used for clustering the data set. They determined that k-mean procedure is most generally used for clustering the data set. They also designate and compare diverse methods of outlier discovery which are statistical method, distance approach, density approach, and deviation approach [3].

In 2016 CharuAggarwal et al proposed "Outlier Analysis". They demonstrated that the difficulty of outlier detection discoveries applications in frequent areas, where it is needed to determine exciting and rare events in the fundamental generating process. They presented that the core of all outlier discovery approaches is the formation of a probabilistic, numerical or algorithmic model that describes the usual data. The nonconformities from this classical are used to classify the outliers. A good area exact knowledge of the fundamental data is often crucial scheming simple and precise models that do not over fit the fundamental data. The difficult of outlier detection develops particularly stimulating, when important relations exist among the dissimilar data points [4].

In 2017 Zeeshan Ahmad et al proposed "A survey on machine learning and outlier detection techniques". The machine learning methods try to recognize the dissimilar data sets which are given to the machine. The data which originates inside can be separated into two kinds i.e. labeled data and the

unlabeled. These must to challenge both of the data. Those methods have been observed upon as fine. Then the notion of outlier comes into image. Outlier discovery is one of the major issues in Data Mining; to discovery an outlier from a collection of patterns is a famed problem in data mining. They discourses and it tries to clarify some of the methods which can help us in classifying or discovering the opinion which show such kind of irregular behavior, and in technical terms named as outlier discovery techniques[5].

In 2018 C. Leela et al proposed "Outlier Discovery Using Association Rule Mining and Cluster Analysis". The current outlier discovery algorithms are able to identify outliers only in static data sets, but are create to be unsuitable, when it creates to active data sets where data arrive unceasingly in a stream creased fashion like sensor data. They suggested two dissimilar approaches for outlier discovery. They measured outlier discovery difficult in two varieties of data bases, one including data streams, where data arrives unceasingly and also based in a time order and the other connecting static data bases. They delivered two procedures for the difficult. For streamed data, a sliding window is measured to create the data items in a database bounded [6].

In 2019 Hongzhi Wang et al proposed Progress in Outlier Detection Techniques: A Survey". They offered a complete and ordered review of the improvement of outlier discovery methods from 2000 to 2019. They explained essential ideas of outlier discovery and then classify them into different methods from varied outlier discovery techniques, such as distance, clustering, density and ensemble. They described state-of the- art outlier discovery procedures and further debate them in factor in terms of their presentation. They define their pros, cons, and contests to deliver investigators with a concise impression of each method and commend explanations and likely research guidelines. They also gave existing progress of outlier discovery methods and delivers a better considerate of the dissimilar outlier detection approaches [7].

In 2020 Ziyu Wang et al proposed "Further Analysis of Outlier Detection with Deep Generative Models". They offered a possible clarification for occurrence, starting from the comment that a model's characteristic set and great density area may not concur. They also conduct extra investigates to help separate the impact of small level surface versus great level semantics in distinguishing outliers. They discover the tones of applying DGMs to outlier discovery, with the goal of accepting the limits of current methods as well as applied workarounds. The examination here in could more reasonably be practical to uncovering and justifying such algorithmic biases [8].

In 2020 Omar Alghushairy et al proposed "A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams". They addressed native outlier discovery. The best known method for native outlier discovery is the Local Outlier Factor, a density created technique. There are numerous LOF procedures for a static data situation; however, these measures cannot be applied straight to data streams, which are a significant kind of big data. In general, local outlier discovery procedures for data streams are still lacking and better procedures need to be advanced that can efficiently examine the high velocity of data streams to notice local outliers. They presented a works evaluation of local outlier discovery procedures in static and stream situations, with a stress on LOF procedures. They deliberated the benefits and limits of those procedures and suggest several hopeful instructions for developing better local outlier detection approaches for data streams [9].

In 2020 Harry Bhagat et al proposed "Outlier Detection Based on Machine Learning Techniques". They examined and bring composed various outlier discovery methods. The goal of the scheme was to perceive the outliers of the housing values in Melbourne (Australia), by statistical and Machine Learning likelihood models. The type of Machine Learning applied was unsupervised knowledge for all models. The reproductions used were separation Forest, Elliptic Wrapper, Density Spatial Clustering of Presentations with Noise (DBSCAN) and Native Outlier Factor. The results of each model were imagined for multivariate data to detect outliers [10].

4. PROBLEMSTATEMENT

These are the some of the difficulties whichneed to considerwhileconductofoutliers.

1. How we could estimate at outliers by looking at a diagram of the throw plot and best fit-line.
2. We would essential some guideline as to how far away a point requirements to be in order to be careful an outlier.
3. In the linear data set, how to find fresh line which is a superior fit to the outstanding data values.
4. There are several approaches and practices are available it is difficult to choose a method for a given data set.

5. PROPOSED ALGORITHM

Proposed algorithm has following steps

1. Draw the scatterplot.
   1) Linear or non-linear pattern of the data
   2) Deviations from the pattern (outliers).
2. Fit the least-squares regression line to the data and check the assumptions of the model by looking at the Residual Plot and normal probability plot (for normality assumption). If the assumptions of the model appear not to be met, a transformation may be necessary.
3. Fit the least-squares regression line data by calculating these values

$$m = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b = \bar{y} - m \times \bar{x}$$
$$y = m \times x + b$$

4. Calculate SSE Sum of Squared Errors.
5. Calculate correlation coefficient.
6. Calculate *s*, the standard deviation by using formula

$$s = \sqrt{\frac{SSE}{n-2}}$$

7. Measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance is at least 2s.
8. The data point which do not satisfy the condition declared as outliers

$$\|y - \bar{y}\| \geq 2(s)$$

6. IMPLEMENTATION

We have taken more than 500 records with two field attendance and marks. Attendance value is taken out 100 and a mark of student is taken out of form out of 200. We need to

find good fitting line and also find out the correct outliers. After finding outlier it also found that when we delete outliers what the value of correlation coefficient is. We used VB dot net 2010 as front end to design user interface. We used SQL server 2010 R2 to store data set.

Figure 4 display number of outlier and required calculations. These forms also display value of coefficient (b=-35.932) and value of (m=2.879) intercept and also display value of standard deviation(s=18.86). This form also displays the upper and lower regression line which displays the lower and upper boundary for outliers

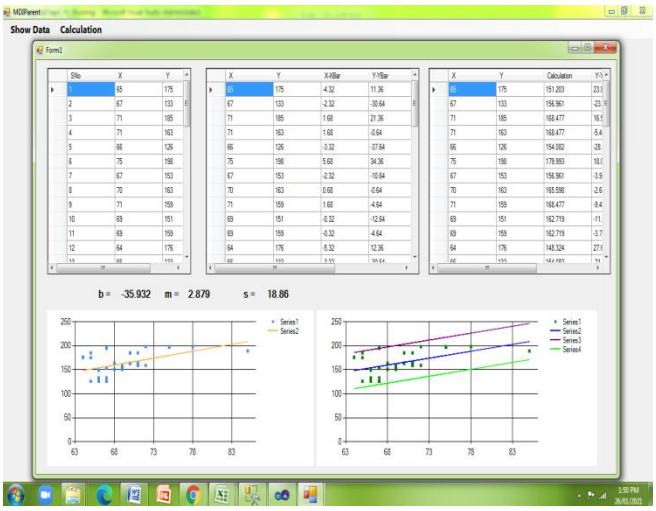


Fig 4 Outliers with 500 records

7. COMPARISON BASED ON VALUES CORRELATION COEFFICIENT

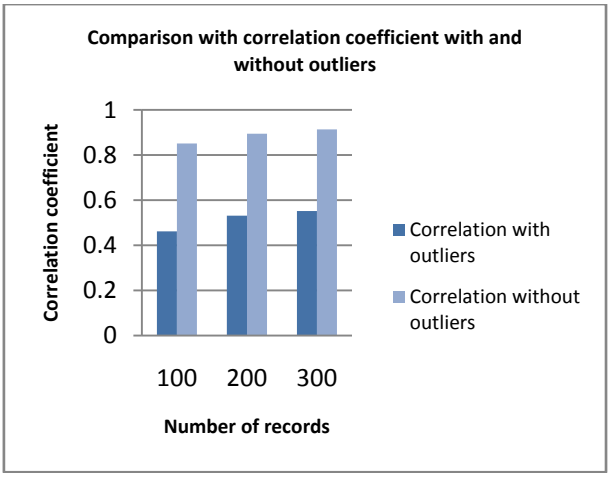
We compare the value of correlation coefficient with and without outliers. With 100 record the value of correlation coefficient 0.462 with outliers and when we removed the outlier the value of correlation coefficient is 0.851. With 500 records the value of correlation coefficient 0.531 with outliers and when we removed the outlier the value of correlation coefficient is 0.894. With100 record the value of correlation coefficient 0.552 with outliers and when we removed the outlier the value of correlation coefficient is 0.9134.

Table 1 Number of records and correlation coefficient with and without outliers

Number of Records	Correlation with outliers	Correlation without outliers
100	0.462	0.851
500	0.531	0.894
1000	0.552	0.9134

8. CONCLUSION AND FUTURE WORK

We proposed an approach to identify outliers in linear data. Proposed approach is simpler and easy to interpret. In the proposed work we used the value correlation coefficient with and without outliers, and show that how the effected in the presence of outlier and without outliers. The correlation coefficient ought to be closer to 1 or -1. The value of correlation coefficient near to 1 show, that there is strong correlation between the given dataset.



Graph 1 Number of records and correlation coefficient with and without outliers

3. REFERENCES

[1] Lakshmi Sreenivasa Reddy. D Dr B. RaveendraBabu Outlier Analysis of Categorical Data using FuzzyAVF2013 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2013].

[2] VarunChandolaArindam Banerjee Outlier Detection : A Survey Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications| Data Mining General Terms: Algorithms Additional Key Words and Phrases: Outlier Detection, Anomaly Detection

[3] Shivani P. Patel Vinita Shah A Survey Of Outlier Detection In Data Mining National Conference on Recent Research in Engineering and Technology (NCRRET -2015) International Journal of Advance Engineer ing and Research Development (IJAERD)

[4] Charu C. Aggarwal Outlier Analysis Second Edition Charu C. Aggarwal IBM T. J. Watson Research Center Yorktown Heights, New York November 25, 2016 rd.springer.com.

[5] Zeeshan Ahmad Lodhia and AkhtarRasool A survey on machine learning and outlier detection techniques IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.5, May 2017.

[6] C. Leela Krishna, C. Kala Krishna Outlier Detection Using Association Rule Mining and Cluster Analysis International Journal of Computer Sciences and Engineering Vol.6(6), Jun 2018, E-ISSN: 2347-2693.

[7] Hongzhi Wang and Mohamed Hammad Progress in Outlier Detection Techniques: A Survey Received July 14, 2019, accepted July 29, 2019, date of publication August 2, 2019.

[8] Ziyu Wang, Bin Dai and Jun Zhu "Further Analysis of Outlier Detection with Deep Generative Models" arXiv:2010.13064v1 [stat.ML] 25 Oct 2020.

[9] Omar Alghushairy and RaedAlsini "A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams" Big Data Cogn. Comput. 2021,

[10] Harry Bhagat, S.Priya and, K. Aditya Outlier Detection Based on Machine Learning Techniques International

Journal of Advanced Science and Technology Vol. 29, No. 6, (2020),

[11] Denis Cousineau "Outliers detection and treatment: a review" International Journal of Psychological Research, 2010. Vol. 3. No. 1.

[12] Prasanta Gogoi1, D K Bhattacharyya Survey of Outlier Detection Methods in Network Anomaly Identification" Received 27 September 2010;

[13] Arturo Elías1, Alberto Ochoa-Zezzatti Outlier Analysis for Plastic Card Fraud Detection aHybridized and Multi-Objective Approach E. Corchado, LNAI 6679, Springer-Verlag Berlin Heidelberg 2011

[14] V. Ilango and R. Subramanian A Five Step Procedure for Outlier Analysis in Data Mining European Journal of Scientific Research ISSN 1450-216X Vol.75 No.3 (2012), Karanjit Singh and Dr. ShuchitaUpadhyaya Outlier Detection: Applications And Techniques IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012ISSN (Online): 1694-0814www.IJCSI.org

[15] Kamal Malik H. Sadawarti, Member IEEE, 3Kalra G.S., Member IEEE Comparative Analysis of Outlier Detection Techniques International Journal of Computer Applications (0975 – 8887) Volume 97– No.8, July 2014.

[16] Zuriana Abu Bakar, RosmayatiMohemad, Akbar A Comparative Study for Outlier Detection Techniques in Data Mining Conference Paper · July 2006 IEEE Xplore University College of Science and Technology21030 Kuala Terengganu, Malaysia

[17] KamaljeetKaurAtulGargComparative Study of Outlier Detection Algorithms International Journal of Computer Applications (0975 – 8887) Volume 147 – No. 9, August 2016.

[18] DipannitaKar, Mr. HareshChande, Mr. RajendraGaikwad A Study Paper on Outlier Detection on Time Series Data www.ijcrt.org © 2017 IJCRT | Volume 5, Issue 4 December 2017 | ISSN: 2320-2882.

[19] RemiDomingues, Maurizio Filippone A comparative evaluation of outlier detection algorithms: experiments and analyses a Department of Data Science, EURECOM, Sophia Antipolis, France Amadeus, Sophia Antipolis, France Preprint submitted to Elsevier August 20, 2018.