DETECTION AND TREATMENT OF OUTLIERS IN DATA SETS

Tara Ahmed H. Chawsheen* Ivan Subhi Latif **

ABSTRACT
In this paper, we shall try to determine outliers and pinpoint its source of existence by using Box-Whisker plots technique which is an effective approach to detect and treat outliers. Thus the researchers prove that the Box-Whisker-Plot is the most effective method among other methods used in this research which is the hypothesis of the paper.

*Assistant Lecture-Department of Statistics, College of Administration and Economics, University of Salahaddin.
**Assistant Lecture-Department of Mathematics, College of Education, University of Salahaddin.

Received: 10/ 6 /2005 ____________________ Accepted: 30/ 8 / 2005
1- Introduction

Most empirical data bases include a certain amount of exceptional values, generally termed as "outliers." The isolation of outliers is important both for improving the quality of original data and reducing the impact of outlying values in the process of analyzing databases, because many statistical data may include some observations which deviate from the general trend to a certain degree and are referred to as "outliers."

Practically, nearly all experimental data samples are subject to contamination by outliers which theoretically reduce the efficiency, and reliability of statistical methods.

The idea of the existence of outliers was first realized in the mid eighteenth century when Boscovich attempted to determine the elliptical shape of earth. He was able to obtain ten different measurements two of which were discarded for their extreme values, and then calculated the average of the remaining eight data.

Outliers are investigated to see if a reason for their unusual behavior can be found. Sometimes outliers have "bad" values occurring as a result of unusual but explainable events.

Examples include faulty measurement, incorrect recording of data or failure of a measuring instrument. This being the case, the outlier is corrected, or deleted from the data set. [7]

Mark and Kandel (2004) [6], emphasized the importance of outlier frequency in a slightly different definition. High and Robin (2004) show that it is a fact of life that data are not well-behaved. 'Outliers'—unusual data values—pop up in most research projects involve data collection.
This is especially true in observational studies where data may naturally have unusual values, even if they come from reputable sources.[3]

Developing techniques to look for outliers and understanding how the impact data analyses are extremely important parts of a thorough analysis especially when statistical techniques are applied to the data. In this paper, we try to determine outliers and specify the sources of its existence.

2- Fundamental Concepts

The fundamental concepts that are embodied in this subject can be outlined as below.

2-1 Outliers

An observation in which its standardized residual is large relative to other observations in the data set, it is considered an outlier that lies at a distance from the rest of the data as shown in Figure No. (1).[7]
**Figure No. (1)**

Examples of outliers

![Graph showing outliers](image)


2-2 **High – Leverage Point**

A large leverage that lies away from the center of points in the X space \((X\text{Axis})\), is regarded as an outlier in the X space.

2-3 **Good Leverage Point**

The data that usually lies on the linear trend set by the majority of the data is called a good leverage point.

2-4 **Bad Leverage Point**

If unusual points appear in both \((x_i, y_i)\) axis, they are called bad leverage points.

2-5 **Influential Observation**

A point is defined to be influential if it doesn't conform with the rest of the data. Another way to look at it is: do your results change substantially when computing them with or without such an observation? If so, it is considered an influential observation.

* In general, if the observation diverges significantly in the X- direction it is called a high leverage point and if it diverges in the Y- direction it is known as an influential observation.

2-6 **Extreme Value**

It is not certain that those values do not belong to the others because they are either far too big or far too small, Figure No. (2).
3- Source of Outliers

The possible source of outliers can be summarized by recording and measuring errors, incorrect distribution assumption, unknown data structure or just an incidental phenomenon.

Recording and measuring errors are often the first source of suspected outliers. Incorrect assumption about the data distribution can lead to mislabeling them as outliers. The data which do not fit well into the assumed distribution may fit well into a different set of distributions, see [2][4][8].

Unknown data structure and correlations can cause apparent outliers. A data set could be made up of subsets with two different mechanisms; the first one is that data come from some heavy tailed distribution such as student's t. As such any observation in such a distribution is in no way erroneous. The second mechanism is that data arise from two distributions. One of these, the 'basic distribution', generates 'good' observation, while, the other one 'contaminating distribution', generates 'contaminants.' If the contaminating distribution has tails which are heavier than those of the basic distribution, then there will be a tendency for contaminators to be outliers – that is separable from the good observations, which will then constitute the "inliers". The data set indicated above should be analyzed independently of each other.
4- Outlier Detection

Visual inspection of scatter plots is the most common approach to outlier detection.

Making an analogy between unsupervised and supervised methods of machine learning, two types of detection methods can be distinguished: univariate methods, which examine each variable individually, and multivariate methods, that take into account associations between variables in the same dataset. In both methods a value is considered to be an outlier if it is far away from other values of the same attribute.[1]

The information in a frequency distribution is often graphed more easily. That is if the distribution is graphed the presentation will be more appealing. One very common graphical presentation is the histogram. Tukey [4] introduced a method of organizing interval–scaled data and called it "stem-and-leaf display". This is viewed as an alternative to the Histogram and it is useful in preliminary analysis only, for more information see [1][8].

The Mahalanobis Distance [10] is another method which is used for detecting outliers, and the main reasons for using it is its sensitivity to inter-variable changes in the training data. In addition, it is measured in terms of standard deviations from the mean of the training sample. This method is not perfect though, and in fact there are a number of drawbacks. In this respect, Rouseeuw & Leroy (1987), introduced robust distance to detect outliers since it is very sensitive to outliers.[11]

As mentioned earlier, visual inspection is not a viable method for actual discriminate analysis applications and detecting outliers. After that, many statistical measures are developed in order to determine the magnitude of outliers, Standardized residuals and DIFFTS (different fits), that is why they are so widely recommended.[7],[9]

Visual detection of outliers suffers from two basic limitations subjectiveness and poor scalability. Analysts have to apply their own subjective perception in order to determine the parameters such as the "very far away" and "low frequency." manual inspection of scatter plots for every variable is also an extremely time-consuming task and, not suitable for most commercial databases containing hundreds of numeric and nominal attributes.
An objective and quantitative approach to unsupervised detection of numeric outliers is described in [6]. It is based on the graphical technique of constructing a box plot, box-whisker-plots, which represent the median of all the observations and two hinges, (whisker), or medians of each half of the data set.

Most values are expected in the inter quartile range (IQR) or located between the two hinges. Values lying outside the ±1.5 H or 1.5 (IQR) are called "mild outliers" and values outside the boundaries of 3H or 3(IQR) are termed "extreme outliers." While this method represents a practical alternative to manual inspection of each box plot, it can deal only with continuous variables characterized by unimodel probability distributions. The other limitation is imposed by the ternary classification of all values into "extreme outliers", "mild outliers" and "non-outliers." The classification changes abruptly with moving a value across one of the 1.5 H, or 3H boundaries.

An information-theoretic approach to supervised detection of erroneous data has been developed by Guyon et al. The method requires building a prediction model by using one of data mining techniques (e.g., neural networks or decision trees). See [6].

The patterns having the lowest probability to be predicted correctly by the model are unreliable and should be treated as outliers.

However, this approach ignores the fact that data conformity may also depend on the inherent distribution of database attributes and some subjective, user – dependent factors, see [4] and [6].

5- Interquartile Range

The quartiles can be used to create another measurement of variability, the interquartile range, which is defined as follows:

\[ H = Q_3 - Q_1 \]

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.
And these quartiles are used in box-whisker-plot for detecting outliers and depends on five statistics: the minimum, the maximum observations, the first, second, and third quartiles. It also depicts other features of a set of data, as shown in Figure (3).

The three vertical lines of the box are first, second and third quartiles. The lines extending to the left and right are called whiskers. Any point that lies outside the whiskers is called an outlier.

The whiskers extend outward to the smaller of 1.5 times the interquartile range or to the most extreme point that is not an outlier.

In other words, if any observation lies outside the range of $Q_1 - (1.5 \times IQR)$ and $Q_3 + (1.5 \times IQR)$ it will be defined as potential outliers and if it lies outside the range $Q_1 - (3 \times IQR)$ and $Q_3 + (3 \times IQR)$ it is defined as problematic outliers.\[5\]

6- Treatment of outliers

Effectively working with outliers in numerical data can be a training experience. Various statistical tests have been proposed for detecting and rejecting outliers, because they can cause potential computational and inference problems. A few possible approaches to treat outliers are listed below:-
6-1 Transformation
Transformation data is one way to soften the impact of outliers since the most commonly used expressions, square root and logarithms, modify the larger values to a much greater extent than they do the smaller values. However, transformations may not fit into the theory of the model or they may affect its interpretation. Taking the log of a variable does more than make a distribution less skewed; it changes the relationship between the original variable and the other variables in the model. In addition, most commonly used transformations require non-negative data or data that is greater than zero, so they do not always provide the answer.

6-2 Deletion
When there are legitimate errors and cannot be corrected, or lie so far outside the range of the data that they distort statistical inferences the outliers should be deleted. When in doubt, we can report model results both with and without outliers to see how much they change.

Data transformation and deletion are important tools, but they should not be viewed as an all-out for distributional problems associated with outliers. Transformations and/or outlier elimination should be an informed choice, not a routine task.

6-3 Accommodation
One very effective plan is to use methods that are robust in the presence of outliers. Non-parametric statistical methods fit into this category and should be more widely applied to continuous or interval data. When outliers are not a problem, simulation studies have indicated their ability to detect significant differences is only slightly smaller than corresponding parametric methods. There are also various forms of robust regression models and computer intensive approaches that deserve further consideration.

7- A Numerical Example
In order to compare the visual inspection of scatter plots and box-whisker-plots for detecting outliers, a clinical data sets taken from the Rizgari hospital, Erbil, Iraq, about
eye illness in past year (2003). These data are presented in Table (1) below:

Table (1): The Data Under Study

<table>
<thead>
<tr>
<th>Y</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
</tr>
</thead>
<tbody>
<tr>
<td>914</td>
<td>62</td>
<td>358</td>
<td>232</td>
<td>160</td>
<td>260</td>
</tr>
<tr>
<td>423</td>
<td>109</td>
<td>305</td>
<td>28</td>
<td>18</td>
<td>82</td>
</tr>
<tr>
<td>230</td>
<td>11</td>
<td>143</td>
<td>87</td>
<td>14</td>
<td>119</td>
</tr>
<tr>
<td>831</td>
<td>48</td>
<td>246</td>
<td>227</td>
<td>106</td>
<td>149</td>
</tr>
<tr>
<td>748</td>
<td>64</td>
<td>348</td>
<td>148</td>
<td>189</td>
<td>254</td>
</tr>
<tr>
<td>908</td>
<td>47</td>
<td>322</td>
<td>150</td>
<td>134</td>
<td>138</td>
</tr>
<tr>
<td>790</td>
<td>57</td>
<td>344</td>
<td>243</td>
<td>122</td>
<td>250</td>
</tr>
<tr>
<td>848</td>
<td>91</td>
<td>385</td>
<td>223</td>
<td>162</td>
<td>314</td>
</tr>
<tr>
<td>740</td>
<td>88</td>
<td>375</td>
<td>576</td>
<td>105</td>
<td>264</td>
</tr>
<tr>
<td>560</td>
<td>96</td>
<td>389</td>
<td>212</td>
<td>164</td>
<td>254</td>
</tr>
<tr>
<td>911</td>
<td>80</td>
<td>358</td>
<td>232</td>
<td>118</td>
<td>301</td>
</tr>
<tr>
<td>760</td>
<td>3</td>
<td>333</td>
<td>226</td>
<td>151</td>
<td>155</td>
</tr>
</tbody>
</table>


A multiple linear regression model which is based on the data in Table (1) above is expressed in the following form:

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i \\
\text{............................................... (7.1),}
\]

Where
\[
\beta_j = \text{Unknown parameters (regression coefficients), } j = 1, 2, 3, \ldots, 5 \\
Y_i = \text{Blindness and vision weakness, } i = 1, 2, 3, \ldots, 12 \\
X_{i1} = \text{Trachoma} \\
X_{i2} = \text{A pinkeye} \\
X_{i3} = \text{Illness of tear gland (body lachrymal)} \\
X_{i4} = \text{Strabismus} \\
X_{i5} = \text{Glaucoma}
\]

For detecting outlier in the data under study, the researchers begin with the standardized Residual. According to staticgraphics program package, there is one outlier among the entire set of observation; Row (3) belongs to Y, X, X, and X are as shown in table (2). This Table contains one
observation which has a standardized residual value which is greater in absolute value than 2 which measures a number of standard deviations with which each observed value of \( Y \) deviates from the fitted model. In this case, there is one standardized residuals greater than 2;\(^9\) thus the value of 3\(^{rd} \) row represents an influential point and extreme value simultaneously.

**Table (2): Potential Outlier Observations by Using STATICGRAPHICS Program Package, Outlier Greater than 2**

<table>
<thead>
<tr>
<th>Row No.</th>
<th>( Y )</th>
<th>Predicted ( Y )</th>
<th>Residual</th>
<th>Standardized Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>230</td>
<td>382.712</td>
<td>-152.712</td>
<td>-2.85</td>
</tr>
</tbody>
</table>

Source: Compiled and organized by the researchers.

As for influential points, Table (3) shows four rows: 3, 4, 9, and 12. They are influential points because their values are greater than \(|\text{DEFFIT}| > 2\) or their leverage values are greater than three times that of an average data point. Here, the DFFITE has a limit by which it can tell the existence of influential points. This is called a "cutoff" having a magnitude of 2, but its recommended magnitude is: \( 2 \sqrt{(k/2)} \), and if \(|\text{DFFITS}| > \text{cutoff} \) and it is considered as an influential point.

Whereas a leverage is a statistics which measures how influential each observation is in determining the coefficients of the estimated model. In this case, an average data point would have a leverage value equal to 0.5 and there are no data points with more than 3 times the average leverage. Here we find just four data points with unusually large values of DFFITS.
Table (3): Potential Influential Observations by Using \textit{STATICGRAPHICS} Program Package, DFFITS > 2

<table>
<thead>
<tr>
<th>Row</th>
<th>Y</th>
<th>X_2</th>
<th>X_3</th>
<th>X_4</th>
<th>Leverage</th>
<th>DFFITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>230</td>
<td>143</td>
<td>87</td>
<td>14</td>
<td>0.827813</td>
<td>-6.24907</td>
</tr>
<tr>
<td>4</td>
<td>831</td>
<td>246</td>
<td>227</td>
<td>106</td>
<td>0.564614</td>
<td>1.91636</td>
</tr>
<tr>
<td>9</td>
<td>740</td>
<td>375</td>
<td>576</td>
<td>105</td>
<td>0.905815</td>
<td>-3.64831</td>
</tr>
<tr>
<td>12</td>
<td>760</td>
<td>333</td>
<td>226</td>
<td>151</td>
<td>0.765609</td>
<td>-1.56324</td>
</tr>
</tbody>
</table>

Average leverage of single data Point = 0.5

Source: Compiled and organized by the researchers.

Table No. (4) Shows the extreme values in the stated data.

Table (4): Extreme Value, Using \textit{STATICGRAPHICS} Program Package

<table>
<thead>
<tr>
<th>Row</th>
<th>Y</th>
<th>X_1</th>
<th>X_2</th>
<th>X_3</th>
<th>X_4</th>
<th>X_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>82</td>
</tr>
<tr>
<td>3</td>
<td>230</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>576</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Source: Compiled and organized by the researchers.

Figure (4) Confirms the existence of outliers which causes the data to digress the dispersion of the data from normality.
Figure (4): Examples of outliers

Note: Outliers are indicated in bold

Residual Plot

With the use of MINITAB program package, the smallest value of Y is equal to 230 and the largest one is 914. The first, second, and third quartiles are 605, 775 and 893, respectively. The inter-quartile range is equal to 288. One and one half times the inter quartile range is \(1.5(288) = 432\). Outliers are defined as any observations that are less than \(605-432= 173\) and any observations that are larger than \(893+432= 1325\). The whisker to the left extends to zero, which is the smallest observation that is not an outlier. The whisker to the right extends to 914, which is the largest observation that is not an outlier. Therefore, there are no outliers.

By looking at the Figure No. (5), the data appear to be uniformly distributed. Now for \(X_{i1}\), Figure No. (6) the smallest value is 3 and the largest one is 109, and the first, second and third quartiles are 47.250, 63, and 90.250, respectively. The IQR is equal to 43 and outliers are apparently defined because observations are less than -17.25 and otherwise are larger than 154.75. The whisker to the left extends to 3, which is the smallest observation that is not an outlier. The whisker to the right extends to 109, which is the largest observation for \(X_{i1}\), and it is not an outlier. That means the data are devoid of outlier. Concerning \(X_{i2}\), which is graphed in Figure No. (7), Shows that the smallest value is
143 and the largest one is 389, and the first, second, and third quartiles are 309.250, 346, and 370.750. The IQR is equal to 61.5. Here we notice that the value of third row is less than 217 and it lies outside the whiskers (217 and 432.25) is considered as an outlier.

While Figure No. (8), for \(X_{i3}\), the IQR is equal to 83.5, any observation that lies outside the range of 23.25 and 357.25 is considered to be an outlier. Here we have a value which belongs to row 9 with a magnitude of 576 being greater than 357.25 which is considered an outlier on the right hand side of box-whisker-plot. And it appears to be \(\chi^2\) distributed. Now for \(X_{i4}\), its box-whisker-plot shows that there are two outliers see Figure No. (9), the smallest value is equal to 14 and the largest one is equal to 189, while its first quartile is equal to 105.250, second quartile equals 128 and third one is 161.500, IQR is of the magnitude 56.25. If there were any observation that is less than 20.875 and greater than 248.875 it is considered an outlier, since there are two values 18 and 14 belonging to second and third rows of \(X_{i4}\) the observations are less than 20.875 so they are considered outliers.

Finally, Figure No. (10) indicates that the data is normally distributed and there is not any outlier, because all observations are less than 446.375 and are greater than -42.625.
Figure No. (9)

Figure No. (10)

Source: Originated by the researchers.
8- Conclusions

Box-Whisker-Plot shows that there are outliers in the data under study (Figure 11), while other measures do not. This explains that it is not necessarily correct that every extreme value to be an influential observation and that not every influential observation is necessarily an outlier.

An influential observation can be recognized in any data if its deletion from the rest of the data changes the result of the analysis. It is also true to say that any point in the data is extreme value if its magnitudes were twice its smallest value.

It is to note that the shape of a box plot is not heavily influenced by a few extreme observations because the median and the other quartiles are not unduly influenced by extreme observations. That is in contrast with the situation in means and variances. Also, outliers affect the real distribution of any data if that data contains influential points. Thus the researchers prove the research hypothesis.

![Box-and-Whisker Plot](image)

**Source:** Originated by the researchers.

![Figure (11)](image)

9- Recommendations

In order to enquire further on the merits of the outliers in the future the researchers recommend:

1- The use of Information Criteria to detect outliers.

2- The use of Robust Methods especially Forward Search Algorithm of Least Trimmed Square for finding outliers in any data sets.
10- References


