University of Bahrain
College of Science
Department of Mathematics

جامعة البحرين
المكتبة
LIBRARY
University of Bahrain

DETECTION AND HANDLING OF OUTLIERS

By
SULAFA DEEB OBAID

A Thesis submitted to the Department of Mathematics as a partial fulfillment of the Requirements for the degree

Of
MASTER OF SCIENCE

Under the supervision of
Dr. Abdul Latif Al Siddiqi
Dr. Sandor Szabo

January 2002

VERSITY OF BAHRAIN
College of Science
Department of Mathematics

جامعة البحرين
كلية العلوم
قسم الرياضيات

12th January 2003

The Thesis Defence Committee considers the thesis entitled:

"Detection and Handling of Outliners"

Submitted by *Sulafa Deeb Obaid* (I.D. No. 19809041) to the Department of Mathematics is satisfactory and acceptable for the Master of Science Degree in Mathematics.

Thesis Defence Committee:

Dr. Abdullatif AlSiddiqi & Dr. Sandor Szabo Supervisor	Department of Mathematics University of Bahrain
Dr. Mohammed Akkas Ali Internal Examiner	Department of Mathematics University of Bahrain
Prof. Robert Michael Loynes External Examiner	School of Mathematics and Statistics Sheffield, U.K.
Dr. A. Abusafiya Chairman	Department of Mathematics University of Bahrain

Box: 32038 - State of Bahrain
Phone: 782311 / 782349
582582

صندوق بريد: ٣٢٠٣٨ - دولة البحرين
هاتف: ٧٨٢٣١١ / ٧٨٢٣٤٩
فاكس: ٦٨٢٥٨٢

TO MY DEAR HUSBAND, PARENTS, BROTHERS, SISTERS AND CHILDREN

Acknowledgments

I would like to express my deepest gratitude and very special thanks to my supervisors Dr. Sandor Szabo and Dr. Abdul Latif Al Siddiqi.

Their advise, support and encouragement during this research were invaluable.

I would also like to thank Dr. M. Akkas for his instructive courses in statistics, to Dr. Scott Bailey for his enlightening course in operations research.

My thanks also goes to our chairman, Dr. A. Abu Safeya.

Last but not least I am gratitude to my husband and my children as without their support and understanding I would not have been able to complete this thesis.

Abstract

Data sometimes include observations that are clearly separated from the bulk of the data.

If we can determine the source that produces such anomalous results, we might be able to give an explanation for those values, such results may have arisen from a reading, recording, or calculating error in data, an instrument may have not been adjusted properly.

If we are sure that a mistake had happened while collecting or recording the data, then the solution is simply to discard those anomalous values, however in many cases the sample outliers may contain important information and thus data should be investigated carefully before deciding whether to include or to exclude any anomalous values.

This dissertation studies sources, nature, effects associated problems, ways of detection, accommodation and transformation of outliers present in any sample data, also the effect of outliers in regression.

Generally speaking the aim is to address the following three major questions:

1- Do outliers exist in a sample data or not?
2- If yes, do we have to include or to exclude them?
3- If we have to include them, how to handle them? And how to minimize their effect on the results and on the conclusions.

In chapter 1, we present a background of outliers together with several definitions of outliers as well as their sources. Also we discuss why outliers cause problems and how to detect them.

Chapter 2 lists down all definitions and notations used in this dissertation.

Chapter 3 presents methods for labeling of outliers, such as the z-score, the modified Z-score and the box plots method, as well as methods for outliers identification such as ESD single and many outlier procedures, Dixon type test, box plot rule, the L_r test, E_k - statistic, sample kurtosis and the W- statistics.

Chapter 4 discusses ways for accommodation of outliers in normally distributed samples. In accommodation of outliers we present the trimmed mean, the winsorized mean, the scale estimators and the confidence intervals.
We also go beyond the normally distributed samples to handle data that follow and we presented ways for transforming data to resemble a normal distribution such as the logarithm transformation and the square root method.

In chapter 5, the effect of outliers on the fitted regression models was discussed and ways for identifying outliers from graphs such as scatter plots and residual plots and numerically were presented.

All above methods were applied on sets of hypothetical data and the effect of outliers in each case was highlighted.

Finally, we conclude that exploring of outliers is important and fundamental for any data analysis in all areas of study and that nearly all scientists and statisticians should be aware about outliers and shall pay particular attention to it, as it distort the results of any experiment and accordingly the conclusions which may be derived from it.

Contents

1 **Introduction** 2

1.1 Background 2
1.2 Definition of outliers 4
1.3 Sources of outliers 6
1.4 Why outliers cause problems? 7
1.5 Detecting outliers 7
1.6 How to handle outliers 8

2 **Preliminaries** 9

2.1 Definitions 9
2.2 Notations 15

3 **Labeling and identification of outliers** 18

3.1 Labeling of outliers 18
3.1.1 The Z- Scores method 19
3.1.2 Modified Z- scores 21
3.1.3 Box plots 25
3.2 Identification of outliers 30
3.2.1 ESD single - outlier procedure 30
3.2.2 ESD many - outlier procedure 32
3.2.3 Dixon type test 35
3.2.4 Box plot rule 38
3.2.5 L_r test 39
3.2.6 The E_k - statistic 41
3.2.7 Test based on the sample kurtosis 43
3.2.8 The Shapiro- Wilk W test 45

4 **Accommodation and Transformation of outliers** 52

4.1 Accommodation of outliers 53
4.1.1 Estimation of location 54
4.1.2 Trimmed mean 54
4.1.3 Winsorized mean 55
4.1.4 Median 57
4.1.5 Scale estimators 58

4.1.6 Confidence intervals..... 62
4.2 Transformation..... 67
4.2.1 Transforming data by taking logarithm..... 68
4.2.2 Combining data..... 72
4.2.3 Ranking data..... 74
4.2.4 Square root..... 76
4.2.5 Deletion of outliers..... 77

5 Regression **78**
5.1 Scatter plots..... 78
5.2 Jackknife method..... 79
5.3 High leverage and high influence point..... 81
5.4 The residual..... 88
5.4.1 Standardized residuals..... 91
5.4.2 The residuals in matrix notation 93
5.5 Simple regression in matrix term..... 97
5.6 Leverage and the hat matrix..... 99
5.7 Multiple regression..... 105
5.8 Studentized residuals..... 106
5.9 The influence function..... 107
5.10 The empirical influence function..... 108
5.11 Testing for outliers..... 108

6 Conclusion **115**

Bibliography **118**

Appendix A
Figures 121
Tables..... 133

Chapter 1

Introduction

1.1 Background

The concept of outliers has fascinated experimentalists since the earliest attempts to interpret data were made.

In the beginning many people argued whether and on what basis to exclude odd (or unrepresentative) observations from a set of data, nowadays a vast array of specific techniques and procedures are commonly used for processing outliers.

In general, outliers are defined as sample data points that appear to be inconsistent with the majority of the data.

If there exist an outlier, we have to decide how to deal with such occurrence:

1. Should we assume that the outlier is just as the data points and calculate

statistics based on all results? OR

2. Should we discard the odd value, as being unrepresentative.

Initially the best action is to try to discover any special circumstances which affected the outlying observation, if we discover some assignable cause of the outlier then we can safely discard the outlier, however if cause of the outliers was not discovered, then we have to apply the statistic tests.

In analyzing any real set of data, a Statistician may come across several abnormal values, the fact which is commonly referred to as outliers.

Outliers may be highly significant to the interpretation of the results of any research, especially the one which involves the analysis of real data, such as pure statistics, Chemical, Biological, Medical and Earth sciences.

Our subject matter is to investigate the nature, the aspects and the treatment of such outliers. In this regard we shall present several definitions reflecting different views of different resources.

Generally speaking, the term outlier is used to describe a discordant or a contaminant observation, where the terms discordant and contaminant observation may be defined as :

> "A discordant observation is the one that appears surprising or discrepant to the investigator." (*Iglewicz*).
> "A contaminant observation is the one from different distribution than the rest of the data, it may or may not be noted by the investigator." (*Barnett*)

However, regardless of the observation being a discordant or a contaminant, outliers have an important effect on the conclusion of the related study and accordingly it is very important to identify them at the early stages of analysis with a view to either include

or exclude them or at least to perform alternative analysis with or without them.

It is important to note here that the impact of outliers is mainly significant on the mean and on the standard deviation of the population and is not proportional to its frequencies. For instance outliers do not exist in bell shaped populations . Moreover nowadays there are several computer packages, which automatically delete outliers from the sample.

Despite the above, the fact remains that outliers constitute a problem, which needs to be addressed properly by any researcher or investigator.

The question to whether to adopt or ignore them is the main concern of this study.

1.2 Definition of Outliers

We shall list some of the definitions of outliers from different point of view and we shall rely upon these definitions only:

> "Values that are very small or very large relative to the majority of the values in data set."
>
> (*Mann*)
>
> "Those values that deviate appreciably from most of the measurements in data set."
>
> (*Wayne*)
>
> " Some of the observations are too extreme (high or low or both) to be consistent with the assumption that they have all been generated by the same parent."
>
> (*Kendall*)
>
> "Are observations with a unique combination of characteristics identifiable as distinctly different from the other observations"
>
> (*Hair*)
>
> "Are values that seem inconsistent with rest of the points in the sample."
>
> (*Rosner*)

" Outliers are values in the population that have a small probability and are far from regions with the greatest concentration of density."

(*Berry*)

"An outlier in a set of data (to be an observation or subset of observations) which appears to be inconsistent with the remainder of that set of data".

(*Barnett*)

" An observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism."

(*Hawkins*)

"A collective to refer to either a contaminant or a discordant observations"

(*Iglewicz*)

The following example may help in clarifying the differences between the above last three definitions

Example : Consider the following random sample

4.42	4.32	4.63	4.57
4.73	4.43	4.41	4.62
4.39	4.87	14.64	4.92
4.81	4.54	4.56	4.39

At the first instance, the number 14.64 will be considered as outlier while the rest of the values seem to be selected from a single source. However, upon detailed checking of the data we may find that as a result of incorrect setting of the instrument four-measurement errors exist (measurement errors are underlined).

Three of the four underlined values $4.81, 4.87$ and 4.92 may not be discovered immediately from errors in the instrument setting, and thus were not considered outliers. For this reason Barnet and Hawkins would consider only 14.64 as the only outlier in the above data, however Iglewicz would consider the four underlined values as outliers as they don not come from the same distribution.

The above example highlights the fact that we should be careful in distinguishing between discordant observations and outliers.

1.3 Sources of outliers

What are the sources of outliers?

Outliers may be classified in four major categories as follows:

1. Outliers arising from procedural errors such as data entry error, mistakes in recording or coding of data, etc. Such type of outliers needs to be identified early in the beginning. However if they were over looked then we have to eliminate them or treat them same as the missing values are treated.

2. Outliers arising from explainable extra ordinary events. In this case there is an explanation for the uniqueness of the observations and the decision to retain or delete this type of outliers depends on whether the outlying value is representing an observation in the population or not.

3. Outliers arising from unexplainable extra ordinary events. In this case there is no explanation for the extra ordinary observations and usually these values are ommited, however if they represent a valid segment of the population, then it might be retained.

4. Outliers falling in the ordinary range but are unique from the rest of the population. In this case the observation need to be retained unless a clear evidence to exclude them from the population is available.

The main aim of classifying the outliers is to help us in deciding whether to include or to exclude them.

1.4 Why Outliers cause problems?

The question of outliers is one of the earliest statistical interests, and is fundamental in the sense to overcome the difficulties that may occur later on when statistical techniques are applied for the data analysis.

In general, the presence of outliers in any population will:

- Distort any statistical test based on sample means and variances.

- Lead to incorrect conclusions.

- Give bias or distortion of estimators.

- Distort P-values (statistical significance which are affected by the presence of one or more of unusual data)

- Inflate sums of squares (which make it unlikely for any one to partition sources of variation in the data into meaningful components.)

- Affect the regression coefficients estimated by minimizing the sum of squares for errors which are very sensitive to outliers

- Give wrong conclusions in case unusual observations were totally ignored.

1.5 Detecting outliers

The way of handling outliers is depending on their type, nature and how we look at .

Basically there are two possibilities: either the outliers were due to chance or due to a mistake in recording or in labeling the data. In the first case outliers should be maintained, as it could possibly indicate a very important characteristic of the population.

However in the second case they might give an invalid results and accordingly could be easily ignored.

The problem is that we may never know for sure which possibility is more correct.

Moreover, the mathematical calculation exactly identify whether the outlier is coming from the same population as the rest of the data or not.

However, the statistical calculation can tell us if the values are sampled from Gaussian distribution and we find one value far a way from the other values with a small probability. And thus we may conclude that this value is a mistake in recording the data and we can ignore it from the data.

1.6 How to handle the outliers

There are two ways in managing data with outliers:

1. Keeping good and accurate records for every experiment conducted in the laboratory, complete with all possible explanation or any additional information which might be useful in analyzing the observations.

2. Using robust statistical methods in analyzing the data, because the effect of outliers in those methods is relatively minimum.

Chapter 2

Preliminaries

2.1 Definitions

Definition 1 *Underlying distribution*

"The distribution of the measurements in the original population is called the underlying or parent distribution"

Definition 2 *Statistical inference*

"Is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample that has been drawn from that population."

Definition 3 *Target population*

"The target population is the population which one wishes to make an inference."

Definition 4 *Unbiased estimator*

"An estimator is unbiased if the average value of the estimator over all possible

samples equal the true population value. In other words an estimator T of the parameter θ is said to be unbiased estimator of θ if $E(T) = \theta$ "

Definition 5 *Parameter*

"A parameter is a numerical summary measure of a population distribution"

Definition 6 *Statistic*

"A statistic is a numerical quantity calculated from the observations in a sample"

Definition 7 *Skewness*

"Measure of the symmetry of the distribution, in most instances the comparison is made to a normal distribution. A distribution is positively skewed when it contains relatively few large values and tails off to the right. A negatively skewed distribution has a relatively few small values. Skewness values falling outside the range of -1 to +1 indicates a substantially skewed distribution."

Definition 8 *Kurtosis*

"Measure of peakedness or flatness of a distribution when compared with a normal distribution. A positive value indicates relatively peaked distribution, and negative value indicates relatively flat distribution."

Definition 9 *Null hypothesis*

"A null hypothesis is a claim (or statement) about a population parameter that is assumed to be true until it is declared false"

Definition 10 *Alternative hypothesis*

"An alternative hypothesis is a claim about a population parameter that will be true if the null hypothesis is false"

Definition 11 *Type I error*

"Probability of incorrectly rejecting the null hypothesis in most cases also termed alpha (α). Typical levels are 5 or 1 percent, termed 0.05 or 0.01 level respectively"

Definition 12 *Type II error*

"Acceptance of the null hypothesis when it is false, the probability of committing a type II error is denoted by β "

Definition 13 *Normal distribution*

"Purely theoretical continuous probability distribution in which a horizontal axis represents all possible values of a variable and the vertical axis represents the probability of those values occurring. The scores on the variable are clustered around the mean in a symmetrical uni model pattern known as the bell shaped or normal curve."

Definition 14 *Normality*

"Degree to which the distribution of the sample corresponds to a normal distribution."

Definition 15 *Response variable*

"A response variable, also called a dependent variable, is the variable we wish to predict or describe based on the values of another variable "

Definition 16 *Predictor variable*

"The predictor variable, also called the independent variable, is the variable that is used to predict the response variable"

Definition 17 *Scatter plots*

"Representation of the relationship between two metric variables portraying the joint values of each observation in a two dimensional graph."

Definition 18 *Breakdown point*

"The breakdown point of an estimator is defined as the largest proportion of the data that can be replaced by an arbitrary values without causing the estimated value to become infinite"

Definition 19 *Order statistics*

"Given N observations $X_1, X_2, \cdots, X_N$ *of a random variable X, the order statistics are obtained by sorting the* $\{X_i\}$ *in an ascending order. This procedures* $\{X_{(i)}\}$ *satisfying* $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}$. *The* $\{X_{(j)}\}$ *are the order statistics of the N observation."*

Definition 20 *Missing data*

"Information not available for a subject or case about whom other information is available . Missing data often occur when a respondent fails to answer one or more questions in a survey."

Definition 21 *Missing data process*

"Any systematic event external to the respondent (such as data entry errors or data collection problems) or any action on the part of the respondent, such as refusal to answer a question) that leads to missing data".

Definition 22 *Discordancy test*

"It is a statistical method for handling outliers and it is defined as testing an outliers with the possibility of rejecting it from the data or of identifying it as a feature of special interest."

Definition 23 *Extreme observation*

"Consider a random univariate sample of size n , $X_1, X_2, \cdots, X_n$ from a distribution, if the data is ordered $X_{(1)}, X_{(2)}, \cdots, X_{(n-1)}, X_{(n)}$ then $X_{(1)}$ and $X_{(n)}$ are called the sample extreme. These observations whether we declare either of them to be an outlier depends on how they appear in relation to the model . Extreme values may or may not be outliers, any outlier however are always extreme or relatively extreme values in the sample."

Definition 24 *Masking*

"Masking exists when discordant observation cancel the influence of more extreme observations, some times suspected observations from subgroup that several values are closer to each other than the bulk of the observations. This phenomenon affects the procedure for identifying outliers, even a single outlier in the presence of several suspected values. When masking exists , multiple outliers make it difficult to identify the single most extreme sample point as an outlier, and multiple outliers artificially inflate the standard deviation."

Definition 25 *Swamping*

"Swamping affects a block discordancy test of a group of two or more outliers. Swamping occur when we have r suspected observations contain r^* true outliers and r-r^* values not outliers. The r^* true outliers can make the test procedure that all the r suspected observations appear to be outliers."

Definition 26 *Studentized residuals*

"Are the actual residuals divided by their standard errors, values exceeding 2.5 in magnitude may be used to indicate outliers"

Definition 27 *Simple regression*

"It is an equation that describes the relation between two variables x and y where x is independent variable and y is a dependent variable."

Definition 28 *Linear regression models*

"A simple regression model that gives a straight line, that describes the relationship between two variables x and y is called a linear regression model. $Y = \beta_0 + \beta_1 X + \varepsilon$ $[\beta_0$ *is Y intercept and* β_1 *= slope ,* $E(\varepsilon) = 0$, $var(\varepsilon) = \sigma^2$ *constant]"*

Definition 29 *Non Linear regression*

"If it is not a straight line then it is called a non linear regression."

Definition 30 *Least squares method*

"In regression analysis we find the best line that fits the points . The line obtained by the least square method is called least squares regression line . We choose $\hat{\beta}_0$ *and* $\hat{\beta}_1$ *to minimize the quantity* $\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left[\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right]^2$

where

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

$$S_{xy} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Definition 31 *High leverage points*

"High leverage points are points that have very high or very low values of the independent variables -outliers in the x direction. High leverage are often called leverage points"

Definition 32 *High influence point*

"High leverage point that also happens to correspond to a y outlier is a high influence point"

Definition 33 *The Pth quartiles* [13]

1-The $(K+1)th$ *largest sample point if* $\frac{np}{100}$ *is not an integer (where* K *is the largest integer less than* $\frac{np}{100}$ *)*

2-The average of the $(\frac{np}{100})th$ *and* $(\frac{np}{100}+1)th$ *largest observation if* $\frac{np}{100}$ *is an integer. The upper and lower quartiles can be thought of conceptually as the approximate 75th and 25th percentiles of the sample, that is the points* $\frac{3}{4}$ *and* $\frac{1}{4}$ *along the way in the ordered sample"*

2.2 Notations

Z	Units of standard normal distribution

μ	(Greek letter mu) population mean
σ^2	Population variance
$\bar{x}$	Sample mean
S^2	Sample variance
S	Sample standard deviation
n	Number of observation in a data set.
MAD	Median of the absolute deviation about the median.
x	Independent variable
y	Dependent variable
$\tilde{x}$	median
Q_1, Q_2, Q_3	First, second and third quartiles respectively.
ESD	Extreme studentized deviate.
H_0	Null hypothesis.
H_1	Alternative hypothesis.
$x_{(i)}$	The ith observation in data set when data ranked in an ascending order.
$r_{j,k-1}$	Dixon's r statistic.
f_U	Upper fence.
f_L	Lower fence.
$x^{(i)}$	The most outlying data point after the i most extreme have been deleted.
Z_i	$Z-$ Score of the ith distribution

$\overline{x}$	Sample mean (sample average for the x_i's)
$\overline{y}$	Sample average of the y_i's.
S_{xx}	Corrected sum of squares for the x_i's.
S_{xy}	Corrected sum of cross products.
S_{yy}	Corrected sum of the y_i's
$\hat{y}$	Estimated or predicted value of y, using regression model.
t	The t distribution.
df	Number of degrees of freedom.
$\hat{e}$	Random error $y - \hat{y}$ for a regression model.

Chapter 3

Labeling and Identification of outliers

3.1 Labeling of Outliers

Usually the analyst inspects the data visually before starting the detailed statistical analysis. However, we all know that detection of outliers by visual inspection (without using graphical tools or analysis) is quite difficult.

Accordingly, it would be better for any analyst to start by labeling the suspected outliers before actually making any further studies.

To illustrate the concepts of outlier labeling we will use a hypothetical data set below:

Consider the ten numbers

$3.2, 3.3, 8.1, 3.2, 2.9, 3.7, 3.1, 3.5, 3.3, 9.2$

Generally, there are three different methods to label the outliers for a normal distribution, namely:

- The Z-scores method.

- Modified Z- scores method.

- Box plot method.

In the following pages we shall try to illustrate how each of the above methods may be applied to label the outliers.

3.1.1 The Z- scores method

Z- scores method is normally used for screening data for outliers.

let $x \sim N(\mu, \sigma^2)$

then

$$z = \frac{x - \mu}{\sigma} \sim N(0, 1)$$

For each data points we find the Z- Score using the formula

$$Z_i = \frac{x_i - \bar{x}}{S}$$

Where

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

Any observation which has a Z-score greater than 3 in absolute value will be labeled as outliers.

Most statistical software packages and simple calculator perform Z-scores procedures.

For instance with reference to the above example.

$$\begin{aligned}
\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \frac{43.5}{10} \\
&= 4.35 \\
S^2 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \\
&= \frac{47.245}{9} \\
&= 5.24944 \\
S &= 2.29
\end{aligned}$$

For the hypothetical data set, the Z-scores are:

-0.50, -0.46, 1.64,-0.50,-0.63,-0.28,-0.55,-0.37,-0.46, 2.12

From our calculations the absolute value of the $Z-$scores, nothing exceeds 3.

So what is wrong?

The answer is that subtracting $\bar{x}$ from x_i and dividing by S will not produce large value of Z_i especially when S is large, accordingly the concept of modified Z-Scores is introduced.

Iglewicz mentioned that the absolute value of Z-score is at most $Z = \frac{(n-1)}{\sqrt{n}}$,

for some selected value of n the maximum value of Z-scores which is denoted by Z_{max} , as shown in the following table.

n	Z_{max}
5	1.79
10	2.85
11	3.02
15	3.61
20	4.25

For a small sample size as $n = 10$. It can be seen from the above table that the criterion for labeling x_i as an outlier if $|Z_i| > 3$ will not work because $|Z_i| \leq Z_{\max} = 2.85$ and thus this method will not give accurate labeling because the mean and the standard deviation affected by outliers especially in small data sets and are sensitive to outliers and especially if the data sets are small.

we conclude that Z-scores are not good method for labeling, particularly when the data sets are small.

3.1.2 Modified Z-scores:

The sample mean and the standard deviation are effected even by a single outlier, which means that should an outlier exist then, the Z-scores method would not be helpful .

The estimators $\bar{x}$ and S are not resistant

To overcome such problems, the concept of modified Z-scores method is introduced. Basically, the modified Z-scores method is based on the so called resistant estimators which are not effected by the unusual observations (outliers).

The following example illustrates how the sample mean and the standard deviation are affected by the presence of outliers in the observations.

Example : Consider n observations, let $x_1, x_2, \ldots, x_{n-1}$ have similar values and x_n different from the $n-1$ observations, this situation is $x_1 = x_2 = \cdots = x_{n-1} = x$

$$x_n = x + ny$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{(n-1)\,x + x + ny}{n}$$

$$= \frac{nx - x + x + ny}{n}$$

$$= \frac{n\,(x+y)}{n}$$

$$= x + y$$

$$S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$$

$$= \frac{(n-1)(x-\bar{x})^2 + (x_n - \bar{x})^2}{n-1}$$

$$= \frac{(n-1)\,(x-(x+y))^2 + (x+ny-(x+y))^2}{n-1}$$

$$= \frac{(n-1)\,y^2 + (n-1)^2 y^2}{n-1}$$

$$= \frac{(n-1)\,y^2\,[1+n-1]}{n-1}$$

$$= ny^2$$

$$S = \sqrt{n}\,y \qquad \text{for } y > 0$$

If y approaches infinity, then $\bar{X}$ and S approaches infinity

The largest Z- score is

$$Z_n = \frac{x_n - \bar{x}}{S} = \frac{x+ny-(x+y)}{\sqrt{n}y} = \frac{n-1}{\sqrt{n}}$$

Which is the same as the Z_{max}.

The maximum Z- score depends on the number of observations not on the values of the data.

Instead of the Z-score we will use resistant estimators which are not affected by the changes in the proportion of the sample, these estimators are called high breakdown point. The term breakdown bound sometimes is used instead of breakdown points.

For an example, the sample mean, the standard deviation and the range have breakdown points of zero, if one observation approaches infinity then the estimators will be infinite.

If n is odd, the breakdown point of the median is $50(1-\frac{1}{n})\%$

If n is even, the breakdown point of the median is $50(1-\frac{2}{n})\%$

The median is a robust estimator of the mean

Since the median has a high breakdown point. We replace the sample mean by the sample median which is denoted by $\tilde{x}$.

Also we replace the standard deviation by the estimator MAD (the median of the absolute deviations about the median).

$MAD = median_i \{|x_i - \tilde{x}|\}$

The MAD has 50% breakdown point

The modified Z-scores is defined as

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

since $E(MAD) = 0.6745\ \sigma$, for large n, we put the constant $= 0.6745$

Observations will be labeled outliers if $|M_i| > D$ where the value of D is given in **table- 1**

This test is reliable because the parameters which we use are less effected by outliers.

For the hypothetical data set we compute the modified Z-score which is shown in the table below

i	$Data\ x_i$	$ordered\ x_i$	$ordered\ x_i - \tilde{x}$	$ordered \mid x_i - \tilde{x} \mid$	$M_i = \frac{o.6745(x_i - \tilde{x})}{0.2}$
1	3.2	2.9	−0.45	0.05	−0.51
2	3.3	3.1	−0.25	0.05	−0.17
3	8.1	3.2	−0.15	0.15	16.02
4	3.2	3.2	−0.15	0.15	−0.51
5	2.9	3.3	−0.05	0.15	−1.52
6	3.7	3.4	0.05	0.25	1.18
7	3.1	3.5	0.15	0.35	−0.84
8	3.5	3.7	0.35	0.45	0.51
9	3.4	8.1	4.75	4.75	0.17
10	9.2	9.2	5.85	5.85	19.73

To compute the sample median of the data , we find the median position

1- We arrange the data in an ascending order

2-Since n is even the sample median is the average of $(\frac{n}{2})th$ and $(\frac{n}{2}+1)th$ largest observations

$$\frac{n}{2} \quad , \quad \frac{n}{2}+1$$

$$5 \quad , \quad 6$$

then we find the average of the fifth and sixth value of the ordered data

$$median = \frac{3.3+3.4}{2} = 3.35$$

The $MAD = med_i \{|x_i - \tilde{x}|\}$

$$= \frac{0.15 + 0.25}{2} = 0.2$$

From the above **table** M_i for 8.1 and 9.2 are 16.02 and 19.73 respectively, from **table-1, D = 3.41**

These values are greater than 3.41, so we label the two values 8.1 and 9.2 as outliers.

The modified Z-score is very efficient in labeling outliers, and gives a great help to the statistician to decide the appropriate test for identifying outliers.

3.1.3 Box Plots

Box plots method is a graphic display method which can be used for:

1-Describing the symmetry and the skweness of a distribution

The relationship among the median, upper quartile Q_3 and the lower quartile Q_1 can describe the symmetry and the skweness of a distribution as follows :

a) If Q_1 and Q_3 are approximately equally spaced from the median then the distribution is symmetric.

b) If Q_3 is farther from the median than Q_1, then the distribution is positively skewed.

c) If Q_1 is farther from the median than Q_3, then the distribution is negatively skewed.

2-Labeling the outlying value:

The value x may considered as an outlying value if:

I) $x > Q_3 + 1.5$ (Q-spread)

OR

II) $x < Q_1 - 1.5$ (Q-spread)

where Q-spread $= Q_3 - Q_1$

3-Labeling the extreme outlying value:[9]

The value x is considered as an extreme outlying value if either

1) $x > Q_3 + 3$ ($Q-$spread)

OR

2) $X < Q_1 - 3$ (Q−spread)

Construction of the box plot:

The construction of the box plot is based on the information of the quartiles (including the median) , the highest and the lowest scores.

Box plots are usually constructed as follows:

1) Draw a number line that includes the range of observations.

2) Above that line, draw a box starting from Q_1 and ending at Q_3

3) Inside the box , draw a line at the median.

4) To label the outliers, we find the lower fence (f_L), and the upper fence (f_U), as follows:

$f_L = Q_1 - 1.5$ (Q−spread)

$f_U = Q_3 + 1.5$ (Q−spread)

5) Sample data located beyond the fences are classified as outliers.

Where an extreme outliers is an outlier that is more than 3(Q−spread), beyond the Quartiles.

6) If there are no outliers, extend the vertical line segments from the end of the box to the smallest and largest value of the observations.

7) If there are outliers, extend the vertical line segments from the end of the box to the smallest and largest non-outliers.

Example

Consider the hypothetical data

$$2.1, 2.6, 2.4, 2.5, 2.3, 2.1, 2.3, 2.6, 8.2, 8.3$$

We want to label the outliers.

1- Order the data

$$2.1, 2.1, 2.3, 2.3, 2.4, 2.5, 2.6, 2.6, 8.2, 8.3$$

$n = 10$

2- Find Q_1 (position value)

$$\frac{np}{100} = \frac{10(25)}{100} = 2.5$$

The largest integer less then 2.5 is 2

so Q_1 is $(2+1)\,th$ largest sample point

The third observation $= 2.3$

$Q_1 = 2.3$

3- Find Q_3

$$\frac{np}{100} = \frac{10(75)}{100} = 7.5$$

The largest integer less than 7.5 is 7 , so Q_3 is $(7+1)th$ largest sample point

The 8th observation $= 2.6$

$Q_3 = 2.6$

$$(Q\text{−spread}) = Q_3 - Q_1 = 2.6 - 2.3 = 0.3$$

$$
\begin{aligned}
f_L &= Q_1 - 1.5\ (Q - \text{spread}) \\
&= 2.3 - 1.5(0.3) \\
&= 1.85 \\
f_U &= Q_3\ + 1.5\ (Q - \text{spread}) \\
&= 2.6 + 1.5(0.3) \\
&= 3.05
\end{aligned}
$$

Outlying observations are $8.2, 8.3$

Extreme outlying observations

$x > Q_3\ + 3\ \ (Q-\text{spread})$

OR

$x < Q_1 - 3\ \ \ (Q-\text{spread})$

$x > 2.6 + 3(0.3)$

$x\ \ > 3.5$

$x < 2.3 - 3(0.3)$

$x < 1.4$

8.3 and 8.2 are extreme outlying values

Median $= \dfrac{2.4 + 2.5}{2} = 2.45$

For the hypothetical data set the box plot label the two largest observations as

outliers.

From **figure- 1** box plot for the hypothetical data set, the observation at 8.2 and 8.3 are labeled as outliers.

The box plot labels outliers as part of the data summary it encourages the experimentalist to reconsider and try to explain the observations labeled as outliers.

3.2 Identification of outliers

There are several ways to detect outliers, the common tests used to detect outliers in normal distribution are the Extreme Studentized Deviate and the L_r test.

The extreme studentized deviate (ESD) is an easy method which may be applied even without using a calculator as shown below:

3.2.1 ESD single outlier procedure[13]

If we have a sample $x_1, x_2, \ldots, x_n \sim N(\mu, \sigma^2)$ and suppose we want to test the hypothesis that there is no outlier against there exists outlier i.e. :

H_0 : that no outliers exist

H_1 : that a single outlier exist

With type I error of α

ESD procedure is as follows:

1- find $ESD = max_{i=1,2,\ldots,n} \dfrac{|x_i - \bar{x}|}{S}$ where $S^2 = \dfrac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$

let the sample value x_i such that

$$ESD = \frac{|x_i - \bar{x}|}{S} \quad \text{referred as } x^{(n)}$$

2- Obtain the critical value

$= ESD_{n,1-\alpha}$ from **table-2**

3- $If \;\; ESD > ESD_{n,1-\alpha}$, we reject H_0 and declare that $x^{(n)}$ is an outlier

$If \;\; ESD \leq \; ESD_{n,1-\alpha}$, we declare that no outliers exist.

Example:Consider the hypothetical data set

2.1, 2.6, 2.4, 2.5, 2.3, 2.1, 2.3, 2.6, 8.2, 8.3

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = 3.540$$

$S = 2.489$

$Data$	$\mid x_i - \bar{x} \mid$	$\mid x_i - \bar{x} \mid / S$
2.1	1.44	0.579
2.6	0.94	0.378
2.4	1.14	0.458
2.5	1.04	0.418
2.3	1.24	0.498
2.1	1.44	0.579
2.3	1.24	0.498
2.6	0.94	0.378
8.2	4.66	1.872
8.3	4.76	1.912

As shown above, we are testing for only one outlier $r = 1$ and $\alpha = 0.05$

the critical value from **table- 2** $ESD_{10,0.95} = 2.29$

$ESD = 1.912$ with 8.3 being the most extreme value

$= 1.912 < 2.29$

$ESD < ESD_{10,0.95}$

So we accept the hypothesis and we declare that there are no outliers.

Question : We have seen in the previous example that the values $8.2, 8.3$ appear as outlier, however the single outlier procedure did not identify any outlier. So what went wrong?

The answer to the above question is what is known as "the masking problem" . The value 8.2 hide the effect of the value 8.3 because it makes the standard deviation too large.

Masking problem exists whenever discordant observations cancel the effect of more extreme values and declare that no outlier exist when the detecting procedure is used.

To solve the masking problem, we have to utilize a flexible procedure which can identify either single or multiple outliers.

3.2.2 ESD many - outlier procedure.[13]

This procedure is used for detecting outliers from 1 to k unless the number of outliers is larger than k.

For a large majority of the sample points from experience, a reasonable upper bound is min $\left(\left[\frac{n}{10}\right], 5\right)$

where $\left[\frac{n}{10}\right]$ is the largest integer $\leq \frac{n}{10}$

If the number of outliers is more than five, then the distribution is not normal, unless the sample size is very large.

Let the sample data be $x_1, x_2, \ldots, x_n \sim N\left(\mu, \sigma^2\right)$

Suppose we suspect that there are k outliers, where

$k = min\left(\left[\frac{n}{10}\right], 5\right)$ where $\left[\frac{n}{10}\right]$ *is equal to* the largest number $\leq \frac{n}{10}$

We want to test the hypothesis

H_0 : there are no outliers

H_1 : there exist outliers between 1 and k with a type I error of α

1- Calculate ESD for all the sample data

$$ESD = max_{i=1,2,...,n} \frac{|x_i - \bar{x}|}{S} \text{ where } S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$$

we denote it by $ESD^{(n)}$ and the most extreme value by $x^{(n)}$

2- Remove the value $x^{(n)}$ from the sample data and calculate the mean and the standard deviation for the $n-1$ observations

ESD for the $(n-1)$ observations and we denote it by $ESD^{(n-1)}$ and we denote the most outlying value by $x^{(n-1)}$.

Remove the value $x^{(n-1)}$ from the data and calculate the mean and the standard deviation for the $(n-2)$ sample data and compute $ESD^{(n-2)}$ and find the outlying value and denote it by $x^{(n-2)}$ and so on.

3- Apply the same procedure to remove the most outlying value until we compute k ESD.

We denote k ESD by $ESD^{(n)}$, $ESD^{(n-1)}$, ..., $ESD^{(n-k+1)}$

we denote the most outlying value by $x^{(n)}$, $x^{(n-1)}$, ..., $x^{(n-k+1)}$

4-From **table -2** we find the critical values of

$ESD_{n,1-\alpha}$, $ESD_{n-1,1-\alpha} \ldots$, $ESD_{n-k+1,1-\alpha}$

5- We use the following decision rule for identifying outliers

If $ESD^{(n-k+1)} > ESD_{n-k+1,1-\alpha}$

we declare that there exist k outliers

$x^{(n)}, \ldots, \ldots\ x^{(n-k+1)}$

else, if $ESD^{(n-k+2)} > ESD_{n-k+2,1-\alpha}$ we conclude that there are $(k-1)$ outliers $x^{(n)}, \ldots\ldots, x^{(n-k+2)}$

$\vdots$

else, if $ESD^{(n)} > ESD_{n,1-\alpha}$ then we declare that one outlier $x^{(n)}$

else, if $ESD^{(n)} \leq ESD_{n,1-\alpha}$ then we declare that there is no outliers present.

Computing ESD for a hypothetical data set

Example: Consider the hypothetical data set:

2.1, 2.6, 2.4, 2.5, 2.3, 2.1, 2.3, 2.6, 8.2, 8.3

Suppose we want to test the hypothesis for the hypothetical data set

H_0: There are no outliers

H_1:There are two outliers with a tytpe I error α

where $\bar{x}$= 3.54 , $S = 2.489$

$ESD^{(10)} = 1.91$, $x^{(10)} = 8.3$

We remove 8.3 from the data and we calculate the mean and the standard deviation

$$\bar{x}= \frac{27.1}{9} = 3.01,\ S = 1.95$$

$Data$	$\mid x_i - \bar{x} \mid$	$\mid x_i - \bar{x} \mid / S$
2.1	0.91	0.47
2.6	0.41	0.21
2.4	0.61	0.31
2.5	0.51	0.26
2.3	0.71	0.36
2.1	0.91	0.47
2.3	0.71	0.36
2.6	0.41	0.21
8.2	5.19	2.66

$ESD^{(9)} = 2.66$ $x^{(9)} = 8.2$

We find from **table-2**

$ESD_{10,0.95} = 2.29$, $ESD_{9,0.95} = 2.21$

$ESD^{(10)} = 1.91$, $ESD^{(9)} = 2.66$

Since $ESD^{(9)} > ESD_{9,0.95}$

we declare the two values $x^{(10)}$, $x^{(9)}$ are outliers

Therefore 8.2 and 8.3 are outliers

3.2.3 Dixon - Type Tests[7]

Dixon's test is generally used for detecting a small number of outliers. This test may be used when the number of observations are between 3 and 30, we rank the data in an ascending order, if we have n observations we rank them as $x_{(1)}, x_{(2)}, ..., x_{(n-1)}, x_{(n)}$

We want to test the hypothesis that

H_0: there are no outliers

H_1: there exist outliers

We start by finding the $r_{j,k-1}$ statistic and it is compared to a critical value $\lambda_{j,k-1}$ for a chosen value of α as given in **table- 3**

If the statistic is less than the critical value we accept the null hypothesis and we conclude that no outliers are present.

If the statistic is greater than the critical value we reject the null hypothesis and we conclude that there exist outliers.

$$r_{j,k-1} = \frac{x_{(n)} - x_{(n-j)}}{x_{(n)} - x_{(k)}} \text{ for some } j \text{ and } k \text{ not too much greater than 1}$$

Test for the existence of a single outlier on the right

We choose the value of k as 2 in order to minimize the potential effects of masking which might be caused by an outlier at the left end of the sample.

We find $r_{11} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}$

If $r_{11} > \lambda_{11}$, then, we reject the null hypothesis and we declare that $x_{(n)}$ is an outlier.

If $r_{11} < \lambda_{11}$, then we accept the null hypothesis and we conclude that $x_{(n)}$ is not an outlier.

If we want to test a single outlier on the left then we multiply each observation by $(-)$ and we apply the same procedure.

Test for two outliers on the right

We find $r_{21} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}}$

If $r_{21} > \lambda_{21}$ we reject the hypothesis and we declare that $x_{(n)}, x_{(n-1)}$ are outliers

If $r_{21} < \lambda_{21}$ we accept the null hypothesis and we conclude that there are no outliers.

If we want to test for two outliers on the left we multiply each observation by (-)and apply the same procedure.

Test for one outlier on either side

We want to test if the largest or the smallest observation is an outlier if :

$$\acute{r}_{10} = max\left[\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}, \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}\right] > \acute{\lambda}_{10}$$

Example

Consider the hypothetical data set 2.1, 2.6, 2.4, 2.5, 2.3, 2.1, 2.3, 2.6, 8.2, 8.3

The hypothetical data set in an ascending order

$2.1, 2.1, 2.3, 2.3, 2.4, 2.5, 2.6, 2.6, 8.2, 8.3$

8.2 and 8.3 seem to be outliers , we want to test the hypothesis

we use the test of two outliers on the right

$$r_{21} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}} = \frac{8.3 - 2.6}{8.3 - 2.1} = \frac{5.7}{6.2} = 0.92$$

λ_{21} from **table-3** at $\alpha = 0.05$

$\lambda_{21} = 0.612$

$r_{21} > \lambda_{21}$ so we reject H_0 and accept the H_1 hypothesis that 8.3 and 8.2 are outliers

The Dixon-type tests are easy to apply but it depends on choosing the exact number and location of all suspected outliers. The Dixon- type tests are susceptible of either swamping or masking and accordingly it is used for very small sample size when one or at most two outliers need to be tested.

3.2.4 Box plot rule

Labeling outliers using box plot has upper and lower fences, defined as.

Upper fence $= f_u = Q_3 + 1.5$ (Q–spread)

Lower fence $= f_l = Q_1 - 1.5$ (Q–spread)

The values outside the fences are considered as outliers , but this rule may give incorrect detection of outliers.

Iglewicz developed an appropriate constants to be used instead of the constant 1.5

For large values of n , the appropriate constant is 2.3.

Those constants are given in **table -4** .

Example

Consider the hypothetical data set

$2.1, 2.6, 2.4, 2.5, 2.3, 2.1, 2.3, 2.6, 8.2, 8.3$

$n = 10$ constant from **table-4** is 2.4

we replace 1.5 by 2.4 in f_u and f_l

$$\begin{aligned} f_u &= Q_3 + 2.4(Q - spread) \\ &= 2.6 + 2.4(0.3) \\ &= 3.32 \\ f_l &= Q_1 - 2.4(Q - spread) \\ &= 2.3 - 2.4(0.3) \\ &= 1.58 \end{aligned}$$

Any observation outside the interval $(1.58, 3.32)$ will be considered as outlier

Therefore 8.2 and 8.3 are identified as outliers.

The box plot is easy to use and doesnot need previous knowledge of the number of outliers.

3.2.5 The L_r test [7]

The L_r test can be used for detecting r upper outliers, an analogous test may be used for r lower outliers

$$L_r = \frac{\sum_{i=1}^{n-r} (x_{(i)} - \overline{x}_r)^2}{\sum_{i=1}^{n} (x_{(i)} - \overline{x})^2}$$

where $\overline{x}_r$ is the sample mean obtained after we remove the r largest observations.

If $L_r <$ critical value then we declare that the r observations are outliers.

Consider the hypothetical data set :

$2.1, 2.6, 2.4, 2.5, 2.3, 2.1, 2.3, 2.6, 8.2, 8.3$

$$\begin{aligned} \overline{x} &= \frac{\sum_{i=1}^{10} x_i}{10} \\ &= \frac{35.4}{10} \\ &= 3.54 \\ \overline{x}_2 &= \frac{\sum_{i=1}^{8} x_{(i)}}{8} \\ &= \frac{18.9}{8} \\ &= 2.36 \end{aligned}$$

i	ordered x_i	$x_{(i)}-\bar{x}_2$	$x_{(i)}-\bar{x}$	$(x_{(i)}-\bar{x}_2)^2$	$(x_{(i)}-\bar{x})^2$
1	2.1	-0.26	-1.44	0.0676	2.0736
2	2.1	-0.26	-1.44	0.0676	2.0736
3	2.3	-0.06	-1.24	0.0036	1.5376
4	2.3	-0.06	-1.24	0.0036	1.5736
5	2.4	0.04	-1.14	0.0016	1.2996
6	2.5	0.14	-1.04	0.0196	1.0816
7	2.6	0.24	-0.94	0.0576	0.8836
8	2.6	0.24	-0.94	0.0576	0.8836
9	8.2		4.66		21.7156
10	8.3		4.76		22.6576
total				0.2788	55.744

$$L_2 = \frac{\sum_{i=1}^{8} (x_{(i)-}\bar{x}_2)^2}{\sum_{i=1}^{10} (x_{(i)-}\bar{x})^2}$$

$$= \frac{0.2788}{55.744}$$

$$= 0.005$$

From **table- 5**, the critical value $\alpha = 0.05$ and $r = 2$ is 0.2297

$L_2 < 0.2297$

$\therefore$ 8.2 and 8.3 are outliers.

The L_r - test has the problem of swamping, which exists when the r suspected observations contain r^* as a true outlier and $r - r^*$ not outliers. The r^* true outlier can make L_r small so that all r suspected observation appear to be outliers.

To illustrate swamping, consider the hypothetical data set

4.1,4.2,4.3,4.6,4.5,4.7,4.8,4.3,5.1,10.3 which contain the outlier 10.3

Suppose that we mistakenly feel that the two largest observations are outliers, we decide to use the L_r test with r=2 ,

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{50.9}{10} = 5.09$$

$$\bar{x}_2 = \frac{\sum_{i=1}^{8} x_{(i)}}{8} = \frac{35.5}{8} = 4.44$$

i	ordered x_i	$x_{(i)} - \bar{x}_2$	$x_{(i)} - \bar{x}$	$(x_{(i)} - \bar{x}_2)^2$	$(x_{(i)} - \bar{x})^2$
1	4.1	-0.34	-0.99	0.1156	0.9801
2	4.2	-0.24	-0.89	0.0576	0.7921
3	4.3	-0.14	-0.79	0.0196	0.6241
4	4.3	-0.14	-0.79	0.0196	0.6241
5	4.5	0.06	-0.59	0.0036	0.3481
6	4.6	0.16	-0.49	0.0256	0.2401
7	4.7	0.26	-0.39	0.0676	0.1521
8	4.8	0.36	-0.29	0.1296	0.0841
9	5.1		0.01		0.0001
10	10.3		5.21		27.1441
total				0.4388	30.989

$$\begin{aligned} L_2 &= \frac{\sum_{i=1}^{8} (x_{(i)} - \bar{x}_2)^2}{\sum_{i=1}^{10} (x_{(i)} - \bar{x})^2} \\ &= \frac{0.4388}{30.989} \\ &= 0.014 \end{aligned}$$

From **table -5**, for $\alpha = 0.05$, $r = 2$ the critical value is 0.2297

$$0.014 < 0.2297$$

So we reject the null hypothesis and declare that the two observations 5.1 and 10.3 are outliers, even though 5.1 is not discordant , this criteria is called swamping. we have only one true outlier, by applying L_r test we find that 5.1 is an outlier, but it is not.

3.2.6 The E_k - statistic [1]:

E_k -statistics are meant to cope with outliers in either direction, but E_k doesn't necessarily pick the k most outlying observations .

E_k -statistics

$$E_k = \frac{\sum_{j=1}^{n-k} (r_{(j)} - \bar{r}_{n-k})^2}{\sum_{j=1}^{n} (r_{(j)} - \bar{r})^2}$$

where $r_j = |x_j - \bar{x}|$ the absolute deviation x_j from the sample mean.

$\{r_{(j)}\}$ are the values of r_j in ascending order $r_{(1)} < r_{(2)} < \cdots < r_{(n)}$

$\bar{r}$ is the sample mean of $r's$

and $\bar{r}_{n-k}$ is the mean of the $(n-k)$ lowest $r's$ that is

$$\bar{r}_{n-k} = \frac{r_{(1)} + r_{(2)} + \cdots + r_{(n-k)}}{n-k}$$

Example

If we consider the hypothetical data set $2.1, 2.6, 2.4, 2.5, 2.3, 2.1, 2.3, 2.6, 8.2, 8.3$

We want to test at $k = 2$ outliers, which are 8.2 and 8.3

$$\bar{x} = \frac{\sum_{j=1}^{10} x_j}{10} = 3.54$$

$$\bar{r} = \frac{\sum_{j=1}^{10} r_{(j)}}{10} = \frac{18.84}{10} = 1.884$$

$$\bar{r}_{n-k} = \frac{\sum_{j=1}^{8} r_{(j)}}{8} = \frac{9.42}{8} = 1.178$$

$$\bar{r}_8 = \frac{9.42}{8}$$

$= 1.178$

j	x_j data	$\lvert x_j - \bar{x}\rvert$	ordered $\lvert x_j - \bar{x}\rvert = (r_j)$	$r_{(j)} - \bar{r}_8$	$r_{(j)} - \bar{r}$
1	2.1	1.44	0.94	-0.238	-0.944
2	2.6	0.94	0.94	-0.238	-0.944
3	2.4	1.14	1.04	-0.138	-0.844
4	2.5	1.04	1.14	-0.0380	-0.744
5	2.3	1.24	1.24	0.062	-0.644
6	2.1	1.44	1.24	0.062	-0.644
7	2.3	1.24	1.44	0.262	-0.444
8	2.6	0.94	1.44	0.262	-0.444
9	8.2	4.66	4.66	3.482	2.776
10	8.3	4.76	4.76	3.582	2.876

$$E_k = \frac{\sum_{j=1}^{8} (r_{(j)} - \bar{r}_8)^2}{\sum_{j=1}^{10} (r_{(j)} - \bar{r})^2}$$
$$= \frac{0.279}{20.249} = 0.014$$

The critical value from **table-6**

for $\alpha = 0.05$ is 0.172

$0.014 < 0.172$

so we reject the null hypothesis and we declare that there exist two outliers 8.2 and 8.3

3.2.7 Test based on the sample Kurtosis[7]

The sample kurtosis

$$b_2 = \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^2}$$

We use this test for identifying outliers and to measure the departures from normality and is easy to apply. we find the value of b_2 and the critical value for α obtained from the table.

The 10%, 5% and 1% critical values for sample size $7 \leq n \leq 50$ are given in **table-7**

If $b_2 >$ critical value , then the value x_j that

maximize $|x_i - \bar{x}|$ is declared an outlier .

x_j will be removed from the sample, and the same procedure is repeated

Example, consider the hypothetical data set

2.1, 2.6, 2.4, 2.5, 2.3, 2.1, 2.3, 2.6, 8.2, 8.3

Data	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^4$
2.1	−1.44	2.07	4.30
2.6	−0.94	0.88	0.78
2.4	−1.14	1.30	1.69
2.5	−1.04	1.08	1.17
2.3	−1.24	1.54	2.36
2.1	−1.44	2.07	4.30
2.3	−1.24	1.54	2.36
2.6	−0.94	0.88	0.78
8.2	4.66	21.72	471.57
8.3	4.76	22.66	513.37
Total		55.74	1002.68

$$\bar{x} = \frac{35.4}{10} = 3.54$$

$$b_2 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

$$= \frac{10(1002.68)}{(55.74)^2}$$

$$= \frac{10026.8}{3106.95}$$

$$= 3.23$$

From **table -7 ,** the critical value for $n = 10$, $\alpha = 0.05$ is 3.95 , $b_2 < 3.95$

We conclude that there are no outliers and the procedure stopped.

It is noticed that the b_2 test is not identifying 8.3 and 8.2 as outliers, the sample kurtosis is a good procedure and can be used in practice however, the procedure is effected by masking when neighboring outliers are present.

In our example the discordant observation 8.2 cancel the effect of more extreme observation 8.3 and the test declare no outliers.

3.2.8 The shapiro-Wilk W test (W-Statistics)[7]

W-statsiticcs is a procedure for testing outliers for normal distribution

let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be the order statistics of a sample from normal distribution with a mean μ and a variance σ^2 . The test is built on theoretical results and is basically how to combine the order statistics into an estimator σ which has the form

$$\sum_{i=1}^{n} c_i x_{(i)}$$

$$b = \sum_{i=1}^{h} a_{n,n+1-i} \left(x_{(n+1-i)} - x_{(i)}\right)$$

where h is the largest integer $\leq \frac{n}{2}$

If n is even $\quad h = \frac{n}{2}$

If n is odd $\quad h = \frac{(n-1)}{2}$

W Test procedure

W is defined the ratio of the square of some linear combinations of all ordered sample values to the sum of the squares of the deviations about the sample mean.

1- Find the constants $a_{n,n+1-i}$ for $i = 1, 2, ..., h$ from **table-8** (shapiro and Wilk's W-Statistics)

2- Find $\quad b = \sum_{i=1}^{h} a_{n,n+1-i} \left(x_{(n-i+1)} - x_{(i)}\right)$

and $\quad D = \sum_{i=1}^{n} (x_i - \bar{x})^2$

3- Find $\quad W = \frac{b^2}{D}$

4- Find the critical value C for the sample size n from **table -9**

If $W > C$ we declare that there are no outliers.

If $W < C$ we declare the existence of outlier which the most deviant observation from $\overline{x}$

5- Remove the outlier from the sample data and apply the same procedure for the reduced sample data.

For the hypothetical data set 2.1 , 2.6, 2.4 , 2.5, 2.3 , 2.1 , 2.3 , , 2.6 , 8.2, 8.3

1- Order the data

2.1 , 2.1 , 2.3 , 2.3 , 2.4 , 2.5 , 2.6 , 2.6 , 8.2, 8.3

$x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}, x_{(6)}, x_{(7)}, x_{(8)}, x_{(9)}, x_{(10)}$

$n = 10$ even $\Rightarrow h = \frac{n}{2} = 5$

$$b = \sum_{i=1}^{5} a_{10,10+1-i} \left(x_{(10+1-i)} - x_{(i)}\right)$$

$$= \sum_{i=1}^{5} a_{10,11-i} \left(x_{(11-i)} - x_{(i)}\right)$$

$$= a_{10,10} \left(x_{(10)} - x_{(1)}\right) + a_{10,9} \left(x_{(9)} - x_{(2)}\right) + a_{10,8} \left(x_{(8)} - x_{(3)}\right) + a_{10,7} \left(x_{(7)} - x_{(4)}\right)$$

$$+ a_{10,6} \left(x_{(6)} - x_{(5)}\right)$$

From **table-8**

$a_{10,10} = 0.5739$

$a_{10,9} = 0.3291$

$a_{10,8} = 0.2141$

$a_{10,7} = 0.1224$

$a_{10,6} = 0.0399$

$b = 0.5739\,(8.3 - 2.1) + 0.3291\,(8.2 - 2.1) + 0.2141\,(2.6 - 2.3) + 0.1224\,(2.6 - 2.3) +$

$0.0399\,(2.5 - 2.4)$

$b = 0.5739\,(6.2) + 0.3291\,(6.1) + 0.2141\,(0.3) + 0.1224\,(0.3) + 0.0399\,(0.1)$

$= 5.671$

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	2.1	-1.44	2.0736
2	2.6	-0.94	0.8836
3	2.4	-1.14	1.2996
4	2.5	-1.04	1.0816
5	2.3	-1.24	1.5376
6	2.1	-1.44	2.0736
7	2.3	-1.24	1.5376
8	2.6	-0.94	0.8836
9	8.2	4.66	21.7156
10	8.3	4.76	22.6576

$$D = \sum_{i=1}^{10} \left(x_i - \bar{x}\right)^2$$

$D = 55.744$

$$W = \frac{b^2}{D}$$

$$W = \frac{(5.671)^2}{55.744}$$

$$= \frac{32.16}{55.744}$$

$= 0.577$

the critical value from **table -9** at 0.05 for n=10 equal 0.842

$$0.577 < 0.842$$

We declare that the most deviant value from $\overline{x}$ is an outlier which is 8.3

Again we remove 8.3 from the sample data and we apply the test $n = 9$

1-order the sample data

2.1 , 2.1 , 2.3 , 2.3 , 2.4 , 2.5 , 2.6 , 2.6 , 8.2

$x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}, x_{(6)}, x_{(7)}, x_{(8)}, x_{(9)}$

$n = 9$ odd therefore,

$$h = \frac{n-1}{2}$$
$$= \frac{9-1}{2}$$
$$= \frac{8}{2}$$
$$= 4$$

$$b = \sum_{i=1}^{4} a_{9,9+1-i} \left(x_{(9+1-i)} - x_{(i)}\right)$$

$$= \sum_{i=1}^{4} a_{9,10-i} \left(x_{(10-i)} - x_{(i)}\right)$$

$$= a_{9,9}\left[x_{(9)} - x_{(1)}\right] + a_{9,8}\left[x_{(8)} - x_{(2)}\right] + a_{9,7}\left[x_{(7)} - x_{(3)}\right] + a_{9,6}\left[x_{(6)} - x_{(4)}\right]$$

from **table-8**

$a_{9,9} = 0.5888$

$a_{9,8} = 0.3244$

$a_{9,7} = 0.1976$

$a_{9,6} = 0.0947$

$$b = 0.5888\,(8.2 - 2.1) + 0.3244\,(2.6 - 2.1) + 0.1976\,(2.6 - 2.3) + 0.0947\,(2.5 - 2.3)$$

$$b = 0.5888\,(6.1) + \ 0.3244\,(0.5) + 0.1976\,(0.3) + 0.0947\,(0.2)$$

$$= 3.832$$

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	2.1	-0.91	0.8281
2	2.6	-0.41	0.1681
3	2.4	-0.61	0.3721
4	2.5	-0.51	0.2601
5	2.3	-0.71	0.5041
6	2.1	-0.91	0.8281
7	2.3	-0.71	0.5041
8	2.6	-0.41	0.1681
9	8.2	5.19	26.9361
total			30.569

$$\bar{x} = \frac{\sum_{i=1}^{9} x_i}{9}$$

$$= \frac{27.1}{9}$$

$$= 3.01$$

$$D = \sum_{i=1}^{10} (x_i - \bar{x})^2$$

$$= 30.569$$

$$W = \frac{b^2}{D} = \frac{(3.832)^2}{30.569} = 0.480$$

The critical value from **table-9** at 0.05 for n=9 equal 0.829

$$0.480 < 0.829$$

we declare that 8.2 is an outlier

we remove 8.2 from the sample data

Order the sample data

2.1 , 2.1 , 2.3 , 2.3 , 2.4 , 2.5 , 2.6 , 2.6

$x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}, x_{(6)}, x_{(7)}, x_{(8)}$

$n = 8$ even therefore,

$$h = \frac{8}{2}$$

$$= 4$$

$$b = \sum_{i=1}^{4} a_{8,8+1-i}\left(x_{(8+1-i)} - x_{(i)}\right)$$

$$= \sum_{i=1}^{4} a_{8,9-i}\left(x_{(9-i)} - x_{(i)}\right)$$

$$= a_{8,8}\left[x_{(8)} - x_{(1)}\right] + a_{8,7}\left[x_{(7)} - x_{(2)}\right] + a_{8,6}\left[x_{(6)} - x_{(3)}\right] + a_{8,5}\left[x_{(5)} - x_{(4)}\right]$$

from **table-8**

$a_{8,8} = 0.6052$, $a_{8,7} = 0.3164$, $a_{8,6} = 0.1743$, $a_{8,5} = 0.0561$

$$b = 0.6052\,(2.6 - 2.1) + 0.3164\,(2.6 - 2.1) + 0.1743\,(2.5 - 2.3) + 0.0561\,(2.4 - 2.3)$$

$$b = 0.6052\,(0.5) + \ 0.3164\,(0.5) + 0.1743\,(0.2) + 0.0561\,(0.1)$$

$$= 0.50127$$

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	2.1	-0.26	0.0676
2	2.6	0.24	0.0576
3	2.4	0.04	0.0016
4	2.5	0.14	0.0196
5	2.3	-0.06	0.0036
6	2.1	-0.26	0.0676
7	2.3	-0.06	0.0036
8	2.6	0.24	0.0576
total			0.2788

$$\bar{x} = \frac{\sum_{i=1}^{8} x_i}{8}$$

$$= \frac{18.9}{8}$$

$= 2.36$

$$D = \sum_{i=1}^{10} (x_i - \overline{x})^2 = 0.2788$$

$$W = \frac{b^2}{D} = \frac{0.25127}{0.2788} = 0.901$$

The critical value from **table-9** at $0.05 = 0.818$

$W > C \Rightarrow$ we accept the null hypothesis and we declare that there is no outlier.

So we have two outliers 8.2 and 8.3

The W test is clumsier because one needs h different constants for each value of n .

Chapter 4

Accommodation and Transformation of Outliers

What should we do about outliers?

The presence of outliers in numerical data will give difficult decisions neither ignoring nor deleting them, will give good solutions. If we do nothing, we will have a model that describes essentially none of the data , outliers can cause potential inference problem. Some possible approaches can work with outliers are listed below:

1- Accommodation

2- Transformation

3- Deletion

4.1 Accommodation of outliers

Outliers can sometimes be accommodated in the data by the use of Robust methods of inference which cover all the data and minimize the influence of any outlier and will prevent the outliers from biasing the estimated population parameters.

Some ways of outliers accommodation are the trimmed means, the scale estimator and the winsorized mean.

Resistant statistic[9]

If the computed value of the statistic is not affected by outliers we say that the statistic is resistant to outliers.

Sensitive statistic[9]

we say that a statistic is sensitive to outliers if the computed value of the statistic is significantly dependent on the outlier.

Robust estimator[1]

Real data sets usually contain outliers which both effect the training set and test in negative direction and thus we will have an incorrect estimators and a hidden outliers.

To solve this problem, robust statistical techniques have been developed, to illustrate these techniques we start by defining the Robust estimator as a statistical procedure that is insensitive to extreme values and widely used in population distribution.

The estimators usually give less weight to the extreme values and is not much affected by pad patterns, some robust estimators are:

1- Estimators of location : Trimmed mean, Winsorized mean and median.

2-Scale estimators

4.1.1 Estimation of Location

Some standard methods that protect the population against the outliers like trimming and winsorizing are used for estimating the location parameter μ .

4.1.2 Trimmed mean[9]

Trimmed mean is a robust estimator , which eliminates the extreme values by the effect of trimming, and can be chosen based on:

1-The shape of underlying parent distribution, where the parent underlying distribution is defined as the distribution of the measurements in the original population.

2-The properties of the estimator that is unbiased, reasonably robust and has a small standard deviation.

Trimmed mean is different from the arithmetic mean which is simply the sum of observations divided by the number of observations and thus the extreme values will have a great effect, on the other hand the trimmed mean minimizes the effect of the extreme values .

The $p\%$ trimmed sample mean is calculated as follows:

1-order the data

2-find the $p\%$ of n, trimming that many observations off both ends of the ordered data

3-find the arithmetic mean of the remaining data, if $p\%$ of n is not a whole number round up to the next whole number.

In a normal distribution, the best estimator of the population mean is the sample

mean, however this mean is affected by the outliers and hence the use of trimmed mean as an estimator of the population mean was introduced. The trimmed means accommodate outliers and are easily computed.

Applications

Trimmed means are often used in Olympic scoring to minimize the effects of extreme ratings caused by biased judge.

4.1.3 Winsorized mean[1]

Winsorized mean is a simple yet effective procedure which is used to estimate the location parameter μ which is effected by the existence of outliers.

Consider n observations, we begin by ordering the data from lowest to highest.

Let $x_{(1)}, x_{(2)}, \ldots, x_{(r)}$ the r lowest sample values

and $x_{(n-s+1)}, \ldots, x_{(n)}$ the s highest sample values

If we trimmed $r+s$ observations, the number of observations will be $n-(r+s)$

The winsorized sample is calculated by replacing the trimmed values of the sample data by the values that are in the next order of trimming we mean that the rth and sth highest observation will be replaced by the value which is next to the rth and s highest value respectively.

Example[9]

Researchers have investigated lead absorption in children of parents who worked in a factory where lead is used to make batteries.

The levels of lead in children blood in ug/dl of whole blood as follows

38	23	41	18	37	36	23	62
14	21	17	16	20	15	10	45
49	48	44	35	43	39	34	13
31	39	73	34	22	25	24	35
27							

Solution

1-Order the data before trimming

10	13	14	15	16	17	18	20	21	22	23
23	24	25	27	31	34	34	35	35	36	37
38	39	39	41	43	44	45	48	49	62	73

$n = 33$

$10\%\ of\ 33$ observations is $0.1 \times 33 = 3.3$

We trimmed 4 values off the lower end $10, 13, 14$and 15 and 4 values off the upper end $73, 62, 49$ and 48.

The average of the remaining 25 observations is

Trimmed mean= $T = \dfrac{767}{25} = 30.68$

where

$\bar{x}= \dfrac{1051}{33} = 31.85$

T is 1 unit less than $\bar{x}$ may be a better measure of the centre of the distribution.

The two observations 75 and 62 are rather distant from the rest of the data ,these two extreme scores

cause the sample mean to a large value ,clearly trimming minimizes the effect of outliers.

Winsorized mean

1- Replace the trimmed values with the values that were next for trimming

in the low end we replace 16 instead of $10, 13, 14, 15$

2- In the high end we replace 45 instead of 48,49, 62, 73

Winsorized mean $= \frac{1011}{33} = 30.64$

4.1.4 Median

Another resistant measure of location that is not affected by outliers is the meadian or ,more precisely , the sample median.

Suppose there are n observations in a sample ,

order the observations from smallest to largest .

The sample median is

1- The $\left(\frac{n+1}{2}\right) th$ largest observation if n is odd

2- The average of the $\left(\frac{n}{2}\right) th$ and $\left(\frac{n}{2}+1\right) th$ largest observation if n is even

Example

Consider the hypothetical data set

2.1, 2.6, 2.4, 2.5, 2.3, 2.1, 2.3, 2.6, 8.2, 8.3

1- Order the data

2.1, 2.1, 2.3, 2.3, 2.4, 2.5, 2.6, 2.6, 8.2, 8.3

$n = 10$ even

$\frac{n}{2}, \frac{n}{2}+1$

5, 6

The median $= \frac{2.4+2.5}{2}$

$= 2.45$

If we remove the two outliers 8.2 and 8.3

We have the ordered data

$2.1, 2.1, 2.3, 2.3, 2.4, 2.5, 2.6, 2.6$

$$\text{Median} = \frac{2.3 + 2.4}{2}$$

$$= 2.35$$

We can see from the above example the median is nearly the same with and without the outliers .

The sample median is not affected by the outliers, we say it is resistant to the influence of outliers.

4.1.5 Scale estimators[7]

The measure of variability is as important as the measure of location. The most common measure of variability are the range and the sample standard deviation, where:

The range $R = x_{(n)} - x_{(1)}$, $x_{(n)}$ is the largest observation and $x_{(1)}$ is the smallest observation.

The range is the simplest measure of variability and it is useful for small data sets.

For large data sets, a more sensitive measure of a variability is the measure of dispersion (the standard deviation).

The standard deviation $S = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

R and S are not resistant estimators, because they don't accommodate outliers .

To get resistant estimators we use the MAD estimator

Where $MAD = median_i \left\{ \left| x_i - \tilde{x} \right| \right\}$, $\tilde{x}$ is the sample median

MAD estimator of scale has a break point about 50% so it accommodates outliers.

Another measure of variability that is resistant to the influence of outliers is the

scale estimator

$$Q - spread \;\; R_F = Q_3 - Q_1$$

An estimator that has a high efficiency under varied distribution S_{bi} is defined as

$$S_{bi} = \frac{\sqrt{n \sum_{i=1}^{n} \left(x_i - \tilde{x}\right)^2 \left(1 - u_i^2\right)^4}}{\left| \sum_{i=1}^{n} \left(1 - u_i^2\right) \left(1 - 5u_i^2\right) \right|}$$

The estimator S_{bi} based on biweight estimator of location $W(u)$ where

$$W(u) = \begin{cases} \left(1 - u^2\right) & |u| \leq 1 \\ 0 & |u| > 1 \end{cases}$$

where $u_i = \dfrac{x_i - \tilde{x}}{9MAD}$

The biweight estimator of scale S_{bi} has a good efficiency when the distribution is normal and when outliers exist.

Example

Consider the hypothetical data set

2.1,2.6,2.4,2.5,2.3,2.1,2.3,2.6,8.2,8.3

$\tilde{x} = 2.45$

$MAD = median_i \left\{ \left| x_i - \tilde{x} \right| \right\}$

$MAD = 0.15$

$9MAD = 9(0.15) = 1.35$

i	x_i	ordered x_i	$x_i - \tilde{x}$	u_i	$1-u_i^2$	$(x_i-\tilde{x})^2$	$1-5u_i^2$
1	2.1	2.1	-0.35	-0.2593	0.9328	0.1225	0.6638
2	2.6	2.1	0.15	0.1111	0.9877	0.0225	0.9383
3	2.4	2.3	-0.05	-0.0370	0.9986	2.5×10^{-3}	0.9932
4	2.5	2.3	0.05	0.0370	0.9986	2.5×10^{-3}	0.9932
5	2.3	2.4	-0.15	-0.1111	0.9877	0.0225	0.9383
6	2.1	2.5	-0.35	-0.2593	0.9328	0.1225	0.6638
7	2.3	2.6	-0.15	-0.1111	0.9877	0.0225	0.9383
8	2.6	2.6	0.15	0.1111	0.9877	0.0225	0.9383
9	8.2	8.2	5.75	4.2593	-17.1416	33.0625	-89.7082
10	8.3	8.3	5.85	4.3333	-17.7775	34.2225	-92.8874

$$n\sum_{i=1}^{10}(x_i-\tilde{x})^2(1-u_i^2)^4$$

$$= 10\left[\begin{array}{c} 0.1225(0.9328)^4+0.0225(0.9877)^4+2.5\times10^{-3}(0.9986)^4 \\ +2.5\times10^{-3}(0.9986)^4+0.0225(0.9877)^4+0.1225(0.9328)^4 \\ +0.0225(0.9877)^4+0.0225(0.9877)^4+33.0625(-17.1416)^4 \\ +34.2225(-17.7775)^4 \end{array}\right]$$

$$= 10\,[6272749.506]$$

$$= 62727495.06$$

$$\left|\sum_{i=1}^{10}(1-u_i^2)(1-5u_i^2)\right| = 3195.9769$$

$$S_{bi} = \frac{\sqrt{62727495.06}}{3195.9769}$$

$$= \frac{7920.0691}{3195.9769}$$

$$= 2.4781$$

For the same data we can observe the difference between the different estimators.

$$\begin{aligned} Range &= x_{(n)} - x_{(1)} \\ &= 8.2 - 2.1 \\ &= 6.2 \end{aligned}$$

$$MAD = 0.15$$

$$S_{bi} = 2.4781$$

We have

$$\begin{aligned} Q_1 &= x_{(f)},\ the \text{ fth ordered observation} \\ f &= \frac{\left[\frac{n+1}{2}\right]+1}{2} \end{aligned}$$

where $[x]$ represents the greatest integer $\leq x$. If f involves a fraction,

Q_1 is the average of $x_{[f]}$and $x_{[f+1]}$

Q_3 equals $[x_{n+1-f}]$

$$\begin{aligned} f &= \frac{\left[\frac{10+1}{2}\right]+1}{2} \\ &= \frac{5+1}{2} \\ &= 3 \end{aligned}$$

$$\begin{aligned} Q_1 &= x_{(3)} \\ &= 2.3 \end{aligned}$$

$$\begin{aligned} Q_3 &= x_{(n+1-f)} \\ &= x_{(10+1-3)} \\ &= x_{(8)} \\ &= 2.6 \end{aligned}$$

$$\begin{aligned} Q - spread &= Q_3 - Q_1 \\ &= 2.6 - 2.3 \\ &= 0.3 \end{aligned}$$

4.1.6 Confidence intervals[7]

Confidence intervals are affected by the existence of outliers in the data, and thus to solve this problem, the confidence interval can be adjusted by the use of trimmed means and the winsorized variance, as winsorizing and trimming will minimize the effect of outliers.

The customary $100(1-\alpha)\%$ confidence interval is given by:

$$\bar{x} \pm \frac{t_{n-1}(1-\frac{\alpha}{2})S}{\sqrt{n}}$$

Where

$t_{n-1}(1-\frac{\alpha}{2})$ is the $100(1-\frac{\alpha}{2})$ percentage point of the t distribution with $n-1$ degrees of freedom

The alternative $100(1-\alpha)\%$ confidence interval is given by

$$T \pm t(1-\frac{\alpha}{2})S_w$$

by replacing $\bar{x}$ by T and $\frac{S}{\sqrt{n}}$ by S_w and the degrees of freedom is equal to $n-2r-1$

where T is the trimmed mean.

$t(1-\frac{\alpha}{2})$ is obtained from the t distribution with $n-2r-1$ degrees of freedom, and $2r$ is the number of trimmed observations.

S_w is the standard error obtained from the winsorized variance

$$\text{Winsorized variance} = \frac{1}{n-1}\left[\sum_{i=r+1}^{n-r}(x_{(i)}-T_w)^2 + r\left[(x_{(r+1)}-T_w)^2 + (x_{(n-r)}-T_w)^2\right]\right]$$

where T_w is the winsorized mean is defined as

$$T_w = \frac{1}{n}\left\{\sum_{i=r+1}^{n-r}(x_{(i)} + r\left[x_{(r+1)} + x_{(n-r)}\right]\right\}$$

$$S_w^2 = \frac{\sum_{i=r+1}^{n-r}(x_{(i)}-T_w)^2 + r\left[(x_{(r+1)}-T_w)^2 + (x_{(n-r)}-T_w)^2\right]}{(n-2r)(n-2r-1)}$$

Example

Consider the hypothetical data

$$2.1, 2.6, 2.4, 2.5, 2.3, 2.1, 2.3, 2.6, 8.2, 8.3$$

1-Ordered data

$$2.1, 2.1, 2.3, 2.3, 2.4, 2.5, 2.6, 2.6, 8.2, 8.3$$

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{10} x_i}{10} \\ &= \frac{35.4}{10} \\ &= 3.54 \end{aligned}$$

$$\begin{aligned} S^2 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \\ &= \frac{55.744}{9} \\ &= 6.1938 \\ S &= 2.489 \end{aligned}$$

Standard confidence interval for the population mean

$$= \bar{x} \pm \frac{t_{n-1}(1-\frac{\alpha}{2})S}{\sqrt{n}} \text{ where } \alpha = 0.05$$

$$= 3.54 \pm \frac{t_9(0.975)(2.489)}{\sqrt{10}}$$

we find the value of $t_9(0.975)$ from **table -10**

$$= 3.54 \pm \frac{(2.262)(2.489}{\sqrt{10}})$$

$$= 3.54 \pm 1.78$$

$\therefore$ Standard confidence interval $= (1.76\ ,\ 5.32)$

The 95% confidence interval for the population mean is $(1.76, 5.32)$

Now we find the confidence interval based on the trimmed mean and the winsorized variance.

The 15% trimmed mean, $p\% = 0.15, n = 10$

Fifteen percent of ten observations is 1.5

The next whole integer *is* 2 , so we trim two values of each end of the ordered data.

i.e we trim two values of the high end 8.2, 8.3 and from the low end we trim 2.1, 2.1

The trimmed mean T

$$\begin{aligned} T &= \frac{x_{(3)} + x_{(4)} + \cdots + x_{(8)}}{6} \\ &= \frac{14.7}{6} \\ &= 2.45 \end{aligned}$$

Winsorized mean T_w

$$\begin{aligned} T_w &= \frac{2.3 + 2.3 + 2.3 + 2.3 + 2.4 + 2.5 + 2.6 + 2.6 + 2.6 + 2.6}{10} \\ &= \frac{24.5}{10} \\ &= 2.45 \end{aligned}$$

Number of degrees of freedom is

$$n - 2r - 1 = 10 - 4 - 1$$

$$= \quad 5$$

Where $2r$ is the number of the trimmed observations.

From **table- 10**, we find $\quad t_5(0.975) = 2.571$

$$\begin{aligned}
S_w^2 &= \frac{\sum_{i=2+1}^{10-2} (x_{(i)} - 2.45)^2 + 2\left[(x_{(3)} - 2.45)^2 + (x_{(8)} - 2.45)^2\right]}{(10-4)(10-4-1)} \\
&= \frac{\sum_{i=3}^{8} (x_{(i)} - 2.45)^2 + 2\left[(2.3 - 2.45)^2 + (2.6 - 2.45)^2\right]}{6 \times 5} \\
&= \frac{0.095 + 0.09}{30} \\
&= \frac{0.185}{30} \\
&= 0.00617 \\
S_w &= 0.079
\end{aligned}$$

The alternative confidence interval is

$$\begin{aligned}
& T \pm t(1 - \frac{\alpha}{2}) S_w \text{ with } n - 2r - 1 \text{ degrees of freedom} \\
&= 2.45 \pm 2.571(0.079) \\
&= 2.45 \pm 0.203
\end{aligned}$$

The alternative confidence interval is $\quad (2.247, 2.653)$

The width of this interval is 0.406

which is shorter than the customary interval has a width of 3.56

A simple computer programme can easily find the confidence interval.

It shall be noted here that the use of confidence interval based on a suitable trimmed means is better whenever the outliers exist .

Robust confidence intervals remove outliers and calculate the confidence interval for the remaining data.

4.2 Transformation

Transformation is a mathematical operation such as taking square roots, logarithm or exponentials. In one way transformation reduces the impact of outliers and at the same time it makes the population more consistent to the normal assumption and transform the data to pull in outliers and other values as well.

Transformation may be applied to any sample size and may be used whenever one is concerned about the existence of outliers, simply because mathematical operations such as multiplication or addition will not change the shape of the population.

Multiplication by a small number will shrink all the values including the outliers proportionally , however to give and formulate a prior opinion of the transformed data we have to asses values to the original scale.

The required transformation is the one that shrink the population differentially (larger observations to be shrunk more than the smaller ones).

One way of transformation is the square root which can't be applied for negative values, the most used transformation is the logarithm which is used only for positive numbers and counts.

Another useful way for transformation is ranking .

The main disadvantage of transforming data is that it is difficult to think about the population and accordingly to formulate a prior opinion easily.

To formulate opinions of transformed data, we have to evaluate them in the original scale.

When should we consider a log transformation?[17]

We can use the log transformation in any of the following situations:

1-When the data is bounded by zero from below, the problem of skewness exists and the outliers are prevented on the low end and the left tail of the population will be tightly packed.

Groups with means close to zero are more constrained than those with means far away from zero.

2-When the data are defined as a ratio scale of measurements.

3-When the largest value in the data is three times larger than the smallest value.

Squeezing and stretching will effect the data if the range is wide (i.e. the majority of the data is not at least three times as big as the minimum data, then the logarithm can't squeeze and stretch the data to have any impact.)

4.2.1 Transforming data by taking logarithm [17]

The logarithm method is one of the transformation methods, which can be applied only for positive data and the most common use of it is for counting data such as a number of heart beats, number of electrons and so on.

In applying this method , we need to take the logarithm of each data and analyze the results.

The log transformation change a multiplicative scale into an additive scale and the division scale into subtraction.

Large values will give large change while the small values give small changes, in another words the logarithm shrinks large values (values greater than 1) and stretches small values (values less than 1).

The squeezing (shrinking) and stretching of data can help us in solving one or more of the following problems which might appear in the data:

1- Skewed data

2-Outliers

3-Unequal variation.

Skewness

If the data is skewed to the right, then small values which are very near to each other exist in the left tail , in the right tail the larger values spread apart.

If the population has a heavy tail at the right end, then taking the logarithm of the data will cause the values in the right tail to shrink , those in the left tail to be stretched and it may change the shape of the population similar to the normal distribution, in an other words taking the logarithms of the population, will result in a more nearly normal ones.

If the data is symmetric or skewed to the left , then a log transformation may worsen the things.

Log transformation will be not effective in case the data has a small range (that is if the largest value of the data is not more than three times larger than the smallest value).

If outliers exist on the high end of the data, then the log transformation will squeeze all large values and will pull the outliers back to the rest of the data.

Unequal variations

Some statistical procedures require that all subject groups must have a comparable variation, if there is unequal variation between the data, then some tests and the confidence interval can't be used.

A log transformation may help in sorting out some types of unequal variation, among those types is the groups of large mean and large standard deviation.

A common pattern of unequal variation (groups with large mean also tend to have large standard deviation)

, the variation can be equalized by taking the log transformation , that is because the log transformation will squeeze the groups that have larger standard deviation more than those of smaller standard deviations.

The log transformation will be more effective when the size of the groups standard deviation is directly proportional to the size of it's mean.

Example[2]

A $3 \times 3 \times 3$ factorial experiment related the number of cycles of loading to failure of worsted yarn. The number of cycles of loading to failure of worsted yarn to length of test specimen(Lng, in mm) amplitude of loading cycle (Amp in mm) and load (Ld in g). The scheme and results are shown in the following table

Lng	Amp	Ld	Cycles	Lng	Amp	Ld	Cycles	Lng	Amp	Ld	Cycles
250	8	40	674	300	8	40	1414	350	8	40	3636
250	8	45	370	300	8	45	1198	350	8	45	3184
250	8	50	292	300	8	45	634	350	8	50	2000
250	9	40	338	300	9	40	1022	350	9	40	1568
250	9	45	266	300	9	45	620	350	9	45	1070
250	9	50	210	300	9	50	438	350	9	50	566
250	10	40	170	300	10	40	442	350	10	40	1140
250	10	45	118	300	10	45	332	350	10	45	884
250	10	50	90	300	10	50	220	350	10	50	360

Let's consider length 250 mm VS 300 mm and use the other two factors to pair the observations. This pair is shown in the following table

Length(mm)	Cycles								
300	1414	1198	634	1022	620	438	442	332	220
250	674	370	292	338	266	210	170	118	90
Difference	740	828	342	684	354	228	272	214	130

The second table considers logs of number of cycles

Length (mm)	Log Cycles								
300	7.25	7.09	6.45	6.93	6.43	6.08	6.09	5.81	5.39
250	6.51	5.91	5.68	5.82	5.58	5.35	5.14	4.77	4.50
Difference	0.74	1.18	0.77	1.11	0.85	0.73	0.95	1.04	0.89

Figure- 2 shows the diffences in cycles: **Figure- 3** shows the differences in log cycles

When we compare the two dot plots our conclusions are:

1-The differences in log cycles are close together in copmarison with zero.

2-The sample is more consistent with a normal population

Example

Some people with heart disease, as well as some with hearts that are normal otherwise, experience frequent premature heart beats. A premature beat is not efficient, because it occurs before the heart is filled with blood and so the heart pumps less blood. The problem may be mechanical or electrical.

An experimental drug for treating this condition was evaluated in 12 patients in a clinic. Investigators counted these patients premature beats during a 1- minute period, administered the drug, waited a short time and then counted premature beats another 1-minute period. The results are shown in the following table:

Patient Number	1	2	3	4	5	6	7	8	9	10	11	12
Before	6	9	17	22	7	5	5	14	9	7	9	51
After	5	2	0	0	2	1	0	0	0	0	13	0
Difference	1	7	17	22	5	4	5	14	9	7	-4	51

Figure - 4 shows the dot plot of the difference

From the dot plot we can conclude that the patient no 12 is very far from the other patients

$$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{138}{12} = 11.5$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} (xi - \bar{x})^2}{n-1}} = 13.68$$

If we remove patient 12 from the data and calculate the sample mean and the standard deviation we find that

$$\bar{x} = 7.91 \quad , \quad S = 7.03$$

4.2.2 Combining data[2]

Combining similar observation will certainly reduce the impact of outliers, however this implies that some important results say of an experiment may not appear. Although it is widely agreed that if we combine the data then we have to give the full data as collected

in the experimental , still some reports or publications combine the data without informing the readers about it, the fact which might lead to a wrong conclusions about the experiment.

the following example shows how the combining of the data reduces the impact of outliers.

Example:[2]

Ten patients took digoxin for several weeks and the experimental drug for part of the time. Periodically (approximately every 4 days) investigators measured the patient's digoxin level in the plasma. The results are shown in the following table. On days A and B and again on days E and F the patients had been taking only digoxin, but on days C and D they had been taking also the experimental drug.

Digoxin levels over time

Patient number	Day A	Day B	Day C	Day D	Day E	Day F
1	34	40	56	45	121	34
2	74	15	44	49	38	37
3	31	49	52	58	41	23
4	40	22	38	25	29	32
5	143	33	57	51	49	46
6	54	58	59	43	41	35
7	31	48	65	53	44	30
8	52	61	70	62	66	38
9	51	63	76	50	32	16
10	21	36	56	40	39	38

In this example the outlying observations are the number 143 for patient 5 in Day A, and the number 121 for patient 1 in Day E. After checking, The investigator concluded that these outlying observations do not come from recording errors and he did not find any explanation.

The outlying observations made the means of days A,B,E and F for patients 1 and 5 larger than that for the other patients.

the following table shows that if we take the average for days A,B,E and F and the average for days C and D it can be seen that the average lessen the impact of the existing outliers.

Patient number	Average days C,D	Average days A,B,E,F
1	50.5	57.25
2	46.5	41
3	55	36
4	31.5	30.75
5	54	67.75
6	51	47
7	59	38.25
8	66	54.25
9	63	40.5
10	48	33.5

4.2.3 Ranking Data[2]

Ranking data is a type of transformation, which leads to a robust analysis, in this method we rank the observations from smallest to largest by assigning the smallest observation rank 1 and the next smaller rank 2 and so on.

In general ranking data are not affected by outliers

Example [2]

Consider the numbers

$40, 76, 89, 106, 120, 130, 150, 155, 382$

We rank these numbers as

$1, 2, 3, 4, 5, 6, 7, 8, 9$

In these data, we transform the number 155 to 8 and 382 to 9 , and thus we can see that ranking shrinks the outliers.

If we have two or more tied observations , then we assign the average rank among these observations for example:

Consider the numbers

$40, 40, 89, 106, 130, 130, 130, 155, 382$

The rank of these observations are

$1.5, 1.5, 3, 4, 6, 6, 6, 8, 9$

For two samples we have several possibilities of ranking depending on the set of interest.

Consider male and female kittens .

Ranking within samples changes

Males	40	76	89	106	120	130	150	155	382
Females	66	69	94	103	117	391			

into

Males	1	2	3	4	5	6	7	8	9
Females	1	2	3	4	5	6			

Ranking within samples and comparing the populations makes no sense on the basis of the ranks.

Another possibility for ranking two samples data is to rank the whole data set as

Males	1	4	5	8	10	11	12	13	14
Females	2	3	6	7	9	15			

If the ranks of the combined groups were as follows

Males	7	8	9	10	11	12	13	14	15
Females	1	2	3	4	5	6			

If the sample size increase then the ranks increase and the sample mean, we have to consider ranks when we test the null hypothesis and when the data or other evidence suggest the existence of outliers in the population.

Example

consider an experiment relating the number of cycles of loading to failure of worsted yarn to length of test specimen, amplitude of loading cycle and load.

Consider lengths of 250 mm and 300 mm where the other factors are used to form the various pairs:

length(mm)	cycles								
300	1414	1198	634	1022	620	438	442	332	220
250	674	370	292	338	266	210	170	118	90

Now consider the rank of the cycles and their differences

Length(mm)	Rank of Cycles								
300	18	17	13	16	14	11	12	8	5
250	15	10	7	9	6	4	3	2	1
Difference	3	7	6	7	8	7	9	6	4

Figure -5 shows the dot plot of the difference in ranks of numbers of cycles

We can see from the figure that the dots are closely packed, especially when it is compared with the distance from zero.

4.2.4 Square root

Square root is another way of transforming the data but cannot be used for negative values, the advantage of using square root is that shrinking larger values more than smaller ones.

The disadvantage of transforming data we can't give prior opinion about the population for example:

If we consider the following data

$64, 45, 45, 45, 121$

The sample mean is 64, and when we take the square root for each data as

$8, 6.708, 6.708, 6.708, 11$

The mean of these values is 7.825 and the square of this value is 61.23 as opposed

to 64 and thus it can be noticed that the square root squeezes the values .

To give prior opinions of transformed values, first we have to assess them in the original data.

(**Figure -6**) and (**Figure- 7**) show dot plots of a hypothetical data set 1,2,4,5,8,9 and the dot plot for the square roots of the same data

From (**figure -6**) and (**figure -7**) dot plots it is noted that the square root squeezes the values for m values > 1

with the squeezing greater for larger m's.

4.2.5 Deletion of outliers

The last resort we use is to delete the outliers. We use deletion only if we found errors which can't be corrected or when the data lie very far from the range of the remainder of the data that they distort statistical inferences, when we are in doubt, we find the results both with and without outliers and we see the change in the results.

Chapter 5

Regression

5.1 Scatter plots

Scatter plot is a plot of the points$(x's, y's)$ where independent variable x is represented on the horizontal axis and the dependent variable y in the vertical axis.

We can observe from the scatter plot shown in **(figure 8)** , that if the points fall around a straight line or there is a definite curve in the pattern also we can see from the scatter plot if there is any outlier exist and fall far from the remaining data.

Consider the linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, the scatter plot of y versus x can help to identify the existence of outliers.

The outliers affect the linear fit to the data by pulling the fitted line towards the outliers, this is shown in table and(**figure 8**) where the outlier is the circled data point.

Consider the following data:

x	10	13	16	18	20	22	24	27	30
y	31	35	42	45	51	53	59	31	70

We can see from the scatter plot the observation $(27, 31)$ is a high influence outlier

The regression equation distorted by including the point $(27, 31)$

If the point (27,31) is included then the regression equation is

$\hat{y} = 19.94 + 1.32x$

But if we exclude the point $(27, 31)$, then the regression equation is

$\hat{y} = 9.93 + 2x$

5.2 Jackknife method

The Jackknife method is a method used for detecting outliers, by excluding one data point each time and calculating the regression coefficients and so on, when outliers are excluded the coefficients of regression will change considerably.

This method can be extended to the case of multiple outliers, and here two or three points can be excluded at a time.

Example

consider the following example

x	10	13	16	18	20	22	24	27	30
y	31	35	42	45	51	53	59	31	70

1- first we exclude the data point $(10, 31)$

$$\bar{x} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{170}{8} = 21.25$$

$$\bar{y} = \frac{\sum_{i=1}^{8} y_i}{8} = \frac{386}{8} = 48.25$$

x	13	16	18	20	22	24	27	30
y	35	42	45	51	53	59	31	70
$(x_i - \bar{x})$	-8.25	-5.25	-3.25	-1.25	0.75	2.75	5.75	8.75
$(x_i - \bar{x})^2$	68.06	27.56	10.56	1.56	0.56	7.56	33.06	76.56
$(y_i - \bar{y})$	-13.25	-6.25	-3.25	2.75	4.75	10.75	-17.25	21.75
$(x_i - \bar{x})(y_i - \bar{y})$	109.31	32.81	10.56	-3.44	3.56	29.56	-99.19	190.31

$$\sum_{i=1}^{8} (x_i - \bar{x})(y_i - \bar{y}) = 273.48$$

$$\sum_{i=1}^{8} (x_i - \bar{x})^2 = 225.48$$

$$\begin{aligned} \text{Slope} &= \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \\ &= \frac{\sum_{i=1}^{8} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{8} (x_i - \bar{x})^2} \\ &= \frac{273.48}{225.48} \\ &= 1.212879 \\ &\approx 1.21288 \end{aligned}$$

$$\begin{aligned} \text{Intercept } \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= 48.25 - 1.21288(21.25) \\ &= 22.476317 \end{aligned}$$

Each time we exclude one data point and calculate the slope and the intercept we get the result as shown in

the following table

Data point excluded	slope	intercept
(10,31)	1.21288	22.4763
(13,35)	1.26116	21.423
(16,42)	1.33281	19.552
(18,45)	1.32834	19.601
(20,51)	1.31953	19.359
(22,53)	1.29235	19.1976
(24,59)	1.21563	21.045
*(27,31)	2.00354	9.93239
*(30,70)	0.79712	28.42905

The two points (27,31) and (30,70) appear to be outliers because when they were excluded they showed a considerable change in the slope as well as the intercept.

5.3 High leverage and high influence point

The estimation of the regression slope can be potentially affected by high leverage points; those points have great weight when the slpoe is estimated.

If we omit a point which has a high leverage from a set of data, it will change the regression line substantially.

High influence point will change the slope and will twist the line badly.

If we have high leverage and one point fall outside the pattern of the remaining data, then we have a high influence point

To illustrate the above idea we consider a hypothetical set of ten observations as follows:

X	7.4	7.7	7.6	7.9	7.1	7.3	10	7.8	13	7.2
Y	7.4	7.7	7.6	7.9	7.1	7.3	5	7.8	13.3	7.2

	x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
	7.4	7.4	-0.9	-0.43	0.387	0.81
	7.7	7.7	-0.6	-0.13	0.078	0.36
	7.6	7.6	-0.7	-0.23	0.161	0.49
	7.9	7.9	-0.4	0.07	0.028	0.16
	7.1	7.1	-1.2	-0.73	0.876	1.44
	7.3	7.3	-1	-0.53	0.53	1
	10	5	1.7	-2.83	-4.811	2.89
	7.8	7.8	-0.5	-0.03	0.015	0.25
	13	13.3	4.7	5.47	25.709	22.09
	7.2	7.2	-1.1	-0.63	0.693	1.21
Total	83	78.3			23.666	30.7

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{83}{10} \\ &= 8.3\end{aligned}$$

$$\begin{aligned}\bar{y} &= \frac{\sum_{i=1}^{n} y_i}{n} \\ &= \frac{78.3}{10} \\ &= 7.83\end{aligned}$$

$$\begin{aligned}S_{xy} &= \sum_{i=1}^{10} (x-\bar{x})(y-\bar{y}) \\ &= 23.666 \\ S_{xx} &= \sum_{i=1}^{10} (x-\bar{x})^2 \\ &= 30.7\end{aligned}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$= \frac{23.666}{30.7}$$

$$= 0.771$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= 7.83 - 0.771(8.3)$$

$$= 1.43$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{y} = 1.43 + 0.771x$$

$$x = 10$$

$$\hat{y} = 1.43 + 0.771(10)$$

$$= 9.14$$

$$x = 12$$

$$\hat{y} = 1.43 + 0.771(12)$$

$$\hat{y} = 10.68$$

Removing the point (13,13.3) from the hypothetical data set and drawing the robust line (dached)and the least -squares line (solid) we get the graph as shown in **figure-9c** and the following table .

	x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
	7.4	7.4	-0.38	0.18	-0.068	0.144
	7.7	7.7	-0.08	0.48	-0.038	6.4×10^{-3}
	7.6	7.6	-0.18	0.38	-0.068	0.032
	7.9	7.9	0.12	0.68	0.082	0.014
	7.1	7.1	-0.68	-0.12	0.082	0.462
	7.3	7.3	-0.48	0.08	-0.038	0.230
	10	5	2.22	-2.22	-4.928	4.928
	7.8	7.8	0.02	0.58	0.012	4×10^{-4}
	7.2	7.2	-0.58	-0.02	0.012	0.336
Total	70	65			-4.952	6.153

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{70}{9}$$

$$= 7.778 \approx 7.78$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{65}{9}$$

$$= 7.222 \approx 7.22$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-4.952}{6.153}$$

$$= -0.8$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= 7.22 - (-0.8)(7.78)$$

$$= 7.22 + 6.22$$

$$= 13.44$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{y} = 13.44 - 0.8\, x$$

$$x = 10 \therefore y = 5.44$$

$$x = 8 \therefore y = 7.04$$

Consider the scatter plot **figure-9a** the bulk of the data (8observations at the left) lie on a straight line, the point in the upper right corner $(13, 13.3)$ is far to the right of other points.

Also it has a very high x value, so it has high leverage.

By including this point , the slope of the least - square line would change considerably as can be seen from

figure-9b . The point is not located on the dashed line and thus it is an outlier in the y direction, therefore this point

has a high influence.

Including or excluding high influence points depends on its influence (if it falls near the line or far from the line

through the remaining points).

Mathematically, the effect of high leverage points can be seen from the value of S_{xy} , which is used for finding

the slope, the value of S_{xy} can be written as $S_{xy} = \sum(x_i - \bar{x})y_i$

Most of the computer programs used for regression analysis will calculate one or more diagnostic measures

of leverage and its influence.

The simple regression model

Given the indepent varible X and the dependent variable Y , respectively ,with observed values (x_i, y_i) of X and Y for$i = 1, 2, 3, ..., n$,the simple regression model that relates the two variabels is

$y_i = \beta_0 + \beta_1 x_i + e_i \qquad i = 1, 2, ..., n$

Where β_0, β_1 are unknown parameters

e_i is the statistical error for the ith case , $i = 1, 2, ..., n$

The error term

Real data will almost never fall exactly on a straight line ,the differences between the values of the response obtained and the values given by the model are called statistical errors.

The error is the quantity that causes the points to fall a way from the line .

The $e_i's$ are unobservable quantities introduced into the model to account for the failure of the observed values to fall exactly on a straight line $x_i's$ and $y_i's$ are observed values.

Statistical errors e_i is defined as the following :

$e_i = y_i - (\beta_0 + \beta_1 x_i)$

Assumptions about the error term in the regression model

The error term e_i are random variable that satisfiies these conditions :

1- The e_i has mean value , 0 $E(e_i) = 0$, $i = 1, 2, ..., n$

2-The e_i has a normal distribution

3-Any two observed values of e_i are independent of each other

4-The errors are mutually uncorrelated

$cov(e_i, e_j) = 0$ for all $i \neq j$

$$corr(e_i, e_j) = \frac{cov(e_i, e_j)}{\sqrt{var(e_i)\, var(e_j)}}$$

$$= 0$$

Since $corr(e_i, e_j) = 0$ therfore the variables e_i *and* e_j are unncorrelated

5- $var(e_i) = \sigma^2 \quad , i = 1, 2, ..., n$ where σ^2 is unknown

Least squares estimation :

It is a method for obtaining estimates of parameters in a model in which parameters estimates are chosen to minimize a quantity called the residual sum of squares,we denote it by Rss

$$Rss = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

we differentiate Rss with respect to $\hat{\beta}_0$ *and* $\hat{\beta}_1$, set the derivatives to zero we get

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \tag{1}$$

$$\sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \tag{2}$$

Rearranging terms we get

$$\hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{4}$$

Equations 3 , 4 are called the normal equations .

Solving the equations we get

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}$$

$\hat{\beta}_1$ is the estimator of β_1

$\hat{\beta}_0$ is the estimator of β_0

$\hat{e}_i$ is the residual for the ith case

$$\hat{e}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) \qquad i = 1, 2, ..., n$$

Residual d.f =number of cases– number of parameters in model

For simple regression residual $d.f = n - 2$

$$\hat{\sigma}^2 = \frac{Rss}{n-2}$$

$$= \frac{\sum \hat{e}_i^2}{n-2}$$

Any sum of squares divided by its degrees of freedom is called mean square.

$\hat{\sigma}^2$ is called residual mean square

$\hat{\sigma}$ is called the stanadard error of regression.

5.4 The residual

The ith residual is the difference between the observed value y_i and the corresponding fitted value $\hat{y}_i$.

The residual is denoted by $\hat{e}_i$ it is written as

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Residuals are used for studying whether a given regression model is appropriate for the data , provide information about assumptions about error terms.

Residual outliers can be identified from residual plots.

Properties of residules

1- The sum of the residuals is zero

$$\begin{aligned}\sum_{i=1}^{n}\hat{e_i} &= \sum_{i=1}^{n}(y_i-\hat{\beta_0}-\hat{\beta_1}x_i)\\ &= \sum_{i=1}^{n}y_i-n\hat{\beta_0}-\hat{\beta_1}\sum_{i=1}^{n}x_i\end{aligned}$$

By the normal equation

$$\sum_{i=1}^{n}y_i=n\hat{\beta_0}+\hat{\beta_1}\sum_{i=1}^{n}x_i$$

therefore

$$\sum_{i=1}^{n}\hat{e_i}=n\hat{\beta_0}+\hat{\beta_1}\sum_{i=1}^{n}x_i-n\hat{\beta_0}-\hat{\beta_1}\sum_{i=1}^{n}x_i$$

$$=0$$

2- The sum of the squared residuals is minimum

$\sum_{i=1}^{n}\hat{e_i}^2$ is a minimum

3-Mean :-

The mean of the n residuals $\hat{e_i}$ for the simple linear regression is zero and is denoted by $\bar{e}$

$$\bar{e}=\frac{\sum_{i=1}^{n}\hat{e_i}}{n}=0$$

since $\sum_{i=1}^{n}\hat{e_i}=0$

4-Variance:-

The variance of the n residuals $\hat{e_i}$ for simple regression model

$$\frac{\sum_{i=1}^{n}(\hat{e}_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^{n}\hat{e}_i^{\,2}}{n-2} = MSE$$

5-Nonindependence

The residuals $\hat{e}_i$ are not independent random variables because they involve the fitted value $\hat{y}_i$it is based on the estimated $\hat{\beta}_0$ *and* $\hat{\beta}_1$.

$$var(\hat{e}) = \sigma^2(I - H)$$

Example

Consider the hypothetical data set

x	7.4	7.7	7.6	7.9	7.1	7.3	10	7.8	13	7.2
y	7.4	7.7	7.6	7.9	7.1	7.3	5	7.8	13.3	7.2

$$\begin{aligned}
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\
&= 1.43 + 0.771\ x_i \\
\hat{y}_1 &= 1.43 + 0.771(7.4) \\
&= 7.1354 \\
\hat{e}_i &= y_i - \hat{y}_i \\
\hat{e}_1 &= 7.4 - 7.1406 \\
&= 0.2646 \\
&\approx 0.26
\end{aligned}$$

Residuals in the hypothetical data set

i	x_i	y_i	$\hat{e}_i$	*standarised residual*
1	7.4	7.4	0.26	0.16
2	7.7	7.7	0.33	0.2
3	7.6	7.6	0.31	0.19
4	7.9	7.9	0.38	0.23
5	7.1	7.1	0.19	0.12
6	7.3	7.3	0.24	0.15
7	10	5	-4.14	-2.54
8	7.8	7.8	0.35	0.21
9	13	13.3	1.86	1.14
10	7.2	7.2	0.22	0.13

from **figure -10** we can identify two outliers $(10, 5)$, $(13, 13.3)$

5.4.1 Standardized residuals [12]

It is defined as follows

$$\text{standardized residuals} = \frac{\hat{e}_i}{\sqrt{MSE}}$$

$$\text{where } MSE = \frac{\sum_{i=1}^{n} \hat{e}_i^{\,2}}{n-2}$$

standardized residuals can help in identifying the outlying values. Outliers are points that lie far beyond the scatter of the remaining residuals, it may be four or more standrad deviations from mean.

From the above example $\sum_{i=1}^{10} \hat{e}_i^{\,2} = 21.2808$

$$MSE = \frac{21.2808}{8}$$

$$= 2.6601$$

$$\sqrt{MSE} = 1.63$$

$$\text{Standardized residuals} = \frac{\hat{e}_i}{1.63}$$

the values of the standardized residuals are in the above table.

The standardized residual plot in **figure -11** contains two outliers which are circled.

Regression in matrix notation

The multiple regression equation can be written in matrix terms as

$Y = X\beta + e$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \mathrm{e} = \begin{matrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{matrix}$$

Where

X is a matrix of constants of order $n \times (p+1)$

β is a vector of parameters of order $(p+1) \times 1$, including the intercept β_0

Y is a vector of responses of order $n \times 1$

e is a vector of independent normal random variablesof order $n \times 1$

with expectations $E(e) = 0$

variance- covariance matrix $\sigma^2(e) = \sigma^2 I_n$ of order $n \times 1$

I_n is the $n \times n$ identity matrix

0 is the $n \times 1$ vector of zeroes

$e \sim N\left(0, \sigma^2 I_n\right)$

As a matter of notation ,we define $\acute{p}$ as the number of columns in the matrix X

We define

$\acute{p} = p + 1$ if there is the intercept term

$\acute{p} = p$ without an intercept term

5.4.2 The residuals in matrix notation

It is convenient to use matrix notation when we are studying residuals .

Consider the basic model

$$Y = X\beta + e$$

Let X is a matrix of full rank of n rows and $\acute{p}$ columns

Where

$\acute{p} = p + 1$

The estimated value of β is $\hat{\beta}$ where

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

The fitted values $\hat{Y}$ is given as

$$\begin{aligned} \hat{Y} &= X\hat{\beta} \\ &= X\left[(X^T X)^{-1} X^T Y\right] \\ &= X(X^T X)^{-1} X^T Y \end{aligned}$$

where

X^T is the transpose matrix.

$\hat{e}$ is the vector of residuals defined as

$$\begin{aligned}\hat{e} &= Y - \hat{Y} \\ &= Y - X(X^TX)^{-1}X^TY \\ &= [I - X(X^TX)^{-1}X^T]Y\end{aligned}$$

The matrix $X(X^TX)^{-1}X^T$ is very important in our study and we shall denote it by H

H is called the Hat matrix

$$\begin{aligned}\hat{e} &= (I-H)Y \\ Y &= IY \\ &= (H+I-H)Y \\ &= HY + (I-H)Y \\ &= \hat{Y} + \hat{e} \\ Y &= X\hat{\beta} + \hat{e}\end{aligned}$$

The vector of residuals $\hat{e}$ is not the same as the vector of errors

$$E(\hat{e}) = 0$$

The variances and covariances of the residuals involving the Hat matrix

$$var(\hat{e}) = \sigma^2(I-H)$$

Proof:

Since $\hat{e} = (I - H)Y$

applying the theorem if $W = AY$

then $\sigma^2(W) = A\ var\{Y\}\ \acute{A}$

we get $var(\hat{e}) = (I - H)\ var(Y)(I - H)\acute{}$

$var\{Y\} = \sigma^2 I$

for the normal model error also we have

$(I - H)\acute{} = I - H$ because of the symmetry of the matrix.

$var(\hat{e}) = \sigma^2(I - H)I(I - H)$

$= \sigma^2(I - H)(I - H)$

since $I - H$ is idempotent matrix then $(I - H)(I - H) = I - H$

therefore

$var(\hat{e}) = \sigma^2(I - H)$

The variance of residual $\hat{e}_i$ is denoted by $var(\hat{e}_i)$

Since $var(\hat{e}) = \sigma^2(I - H)$

we have

$$var(\hat{e}_i) = \sigma^2(1 - h_{ii})$$

where h_{ii} is the ith element on the main diagonal of the hat matrix

Each residual has a mean equals to zero and has different variances which are not correlated.

Whenever , $e_i \sim N(0, \sigma^2)$ the residuals are normally distributed.

The element of H is denoted by h_{ij} where

$$h_{ij} = x_i^T (X^T X)^{-1} x_j$$

the diagonal elements

$$h_{ii} = x_i^T (X^T X)^{-1} x_i$$

where x_i^T and x_j^T are the ith and the jth row of the data matrix X

$$cov(\hat{e_i}, \hat{e_j}) = -\sigma^2 h_{ij}$$

$$\begin{aligned} corr(\hat{e_i}, \hat{e_j}) &= \frac{cov(\hat{e_i}, \hat{e_j})}{\sqrt{var(\hat{e_i}) \times var(\hat{e_j})}} \\ &= \frac{-\sigma^2 h_{ij}}{\sqrt{\sigma^2(1-h_{ii}) \times \sigma^2(1-h_{jj})}} \\ &= \frac{-h_{ij}}{\sqrt{(1-h_{ii}) \times (1-h_{jj})}} \end{aligned}$$

H is a symmetric matrix and has a special property called idempotency

$$H^2 = H \times H = H$$

$$rankH = rank(X) = p$$

where p is the number of regression parameters in the rgeression function including the intercept term

Summation of the diagonal element is called the trace

$$trace = \sum_{i=1}^{n} h_{ii} = rank(X) = p$$

$$h_{ii} = \sum_{j} h_{ij^2}$$

let c denote the number of rows in X that are exactly the same as x_i, including x_i

then for the model with the intercept model, we have

$$\frac{1}{n} \leq h_{ii} \leq \frac{1}{c}$$

5.5 Simple regression in matrix term

For simple regression X and Y are given by

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad , \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$(X^T X) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

$$= \frac{1}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \begin{bmatrix} \frac{\sum_{i=1}^{n} x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

$$= \frac{1}{S_{xx}} \begin{bmatrix} \frac{\sum_{i=1}^{n} x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$= \frac{1}{s_{xx}} \begin{bmatrix} \frac{\sum_{i=1}^{n} x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\, y_i \end{bmatrix}$$

$$= \begin{bmatrix} \bar{y} - \hat{\beta}_1 \bar{x} \\ \frac{s_{xy}}{s_{xx}} \end{bmatrix}$$

$$h_{ij} = x_i^T (X^T X)^{-1} x_j$$

$$= (1 \quad x_i) \begin{bmatrix} \frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}} & \frac{-\bar{x}}{S_{xx}} \\ \frac{-\bar{x}}{S_{xx}} & \frac{1}{S_{xx}} \end{bmatrix} \begin{bmatrix} 1 \\ x_j \end{bmatrix}$$

$$= \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}$$

The diagonal elements h_{ii} is defined as

$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$, $\sum_{i=1}^{n} h_{ii} =$ = number of regerisson parameters in regression function including the intercept term

$$= 2$$

5.6 Leverage and the Hat Matrix

The degree of leverage is important to measure the effect of extreme values, and the most common measure of it is the diagonal element of the hat matrix (h_{ii}) which is considered as a good way to identify the observations with high leverage, and often the effect of these observations on the model is easily seen.

h_{ii} is a usefull indicator of whether or not the ith case is outlying with respect to its x values .

h_{ii} is called the leverage (in terms of x values)of the ith case.

Consider the estimated regression model, which gives the regression of Y on X as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

And the residuals $\hat{e}_i = y_i - \hat{y}_i$

To describe the impact of an observation ,we consider the standard formula for the slope of the least-squares line .

The intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and the slope of the least square line is $\hat{\beta}_1$ where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

let

$$w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

then

$$\hat{\beta}_1 = \sum_{i=1}^{n} w_i (y_i - \bar{y})$$

$$\sum_{i=1}^{n} w_i = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \times \sum_{i=1}^{n} (x_i - \bar{x})$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{n} w_i = 0$$

$$\begin{aligned}
\hat{\beta}_1 &= \sum_{i=1}^{n} w_i (y_i - \bar{y}) \\
&= \sum_{i=1}^{n} w_i y_i - \bar{y} \sum_{i=1}^{n} w_i \\
&= \sum_{i=1}^{n} w_i y_i - \bar{y} (0) \\
&= \sum_{i=1}^{n} w_i y_i \\
&= \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\end{aligned}$$

If we substitute the value of $\hat{\beta}_0, \hat{\beta}_1$ in the estimated regression model , then

$$\begin{aligned}\hat{y}_i &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i \\ &= \bar{y} + \hat{\beta}_1 (x_i - \bar{x})\end{aligned}$$

$$\hat{y}_i = \bar{y} + \frac{(x_i - \bar{x}) \sum_{j=1}^{n} (x_j - \bar{x}) y_j}{\sum_{k=1}^{n} (x_k - \bar{x})^2}$$

If we substitute the value of $\bar{y}$ we get,

$$\begin{aligned}\hat{y}_i &= \frac{1}{n} \sum_{j=1}^{n} y_j + \frac{(x_i - \bar{x}) \sum_{j=1}^{n} (x_j - \bar{x}) y_j}{\sum_{k=1}^{n} (x_k - \bar{x})^2} \\ &= \sum_{j=1}^{n} \left\{ \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n} (x_k - \bar{x})^2} \right\} y_j\end{aligned}$$

let

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n} (x_k - \bar{x})^2}$$

then $\hat{y}_i$can be written as

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

h_{ij} shows how $\hat{y}_i$ is affected by the change of y_j

Iglewicz described the hat matrix as $H = (h_{ij})$, it is symmetric and of order n

The leverage of an observation i is defined as h_{ii} which is the diagonal element of the hat matrix.

h_{ii} measures the potential effect of y_i on $\hat{y}_i$, where $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n} (x_k - \bar{x})^2}$ for the simple regression line.

Some properties of the hat matrix and the diagonal element were mentioned by Iglewicz

1-$\sum_{i=1}^{n} h_{ii} = 2$ for the simple linear regression , unless all the x_i are equal.

2-The average size of $h_{ii} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{2}{n}$ (simple linear regression)

3-For any linear regression $0 \leq h_{ii} \leq 1$

4-For any simple linear regression $\frac{1}{n} \leq h_{ii} \leq 1$ (including intercept)

5-If $h_{ii} = 0$ or 1, then $h_{ij} = 0$ for all $j \neq i$

Graphically when there is an observation with a high leverage, then the points with high leverage x_i are relatively far from $\bar{x}$.

The notion of h_{ii} give a measure of how far the ith case is far from the center of the data .

Weisberg mentioned the following cases of h_{ii} :

1-.h_{ii} will be minimum ($h_{ii} = \frac{1}{n}$) if $x_i = \bar{x}$ may be arbitrarily close to one

2- h_{ii} will increase for unusual x_i,that is, for x_i far from $\bar{x}$.

3- If x_i is sufficiently far from $\bar{x}$ then the value of $\frac{(x_i - \bar{x})^2}{s_{xx}}$ becomes large and the values of h_{ii} will be close to one, h_{ii} measures how far the ith case from the centre of the data.

If the value of h_{ii} is large we call this case high leverage (Weisberg)

We have the $var(\hat{e}_i) = \sigma^2(1 - h_{ii})$, when the value of h_{ii} is large then $var(\hat{e}_i)$ will be small , when the value of h_{ii} approaches one , the variance will approach zero.

Graphically an observation has high leverage if the formula for h_{ii} shows that ,in the plot of y versus x,the points with high leverage have x_i relatively far away from $\bar{x}$.

Numerically (Iglewicz) mention convenient criterion when an observation has high leverage ,

if $h_{ii} \leq 0.2$, the values will be safe but if it lies between 0.2 and 0.5 then they are likely to be risky and values above 0.5 indicate very high leverage.

(Neter) mentioned another rule that a leverage value h_{ii} is usually considered to be large value if it is more than twice as larage as the mean leverage value ,denoted by $\bar{h}$ where $\bar{h} = \dfrac{\sum_{i=1}^{n} h_{ii}}{n}$

Hence ,leverage values greater than $2\,\bar{h}$ are considered by this rule to indicate outlying cases with reggard to their x values .

Additional evidence of an outlying case is the existence of a gap between the leverage values for most of the cases and the unusually large leverage values .

Example

Consider the hypothetical data set

x	7.4	7.7	7.6	7.9	7.1	7.3	10	7.8	13	7.2
y	7.4	7.7	7.6	7.9	7.1	7.3	5	7.8	13.3	7.2

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{83}{10} = 8.3$$

i	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})y_i$	h_{ii}	$\hat{e}_i$
1	7.4	7.4	-0.9	0.81	-6.66	0.126	0.26
2	7.7	7.7	-0.6	0.36	-4.62	0.112	0.33
3	7.6	7.6	-0.7	0.49	-5.32	0.116	0.31
4	7.9	7.9	-0.4	0.16	-3.16	0.105	0.38
5	7.1	7.1	-1.2	1.44	-8.52	0.147	0.19
6	7.3	7.3	-1	1	-7.3	0.133	0.24
7	10	5	1.7	2.89	8.5	0.194	-4.14
8	7.8	7.8	-0.5	0.25	-3.9	0.108	0.35
9	13	13.3	4.7	22.09	62.51	0.820	1.86
10	7.2	7.2	-1.1	1.21	-7.92	0.139	0.22
total	83	78.3		30.7	23.61	2	

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$= \frac{78.3}{10}$$

$$= 7.83$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{23.61}{30.7}$$

$$= 0.769$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= 7.83 - 0.769(8.3)$$

$$= 1.45$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{y}_1 = 1.45 + 0.769(7.4)$$

$$= 7.1406$$

$$\hat{e}_i = y_i - \hat{y}_i$$

$$\hat{e_1} = y_1 - \hat{y_1}$$

$$\hat{e_1} = 7.4 - 7.1406$$

$$= 0.26$$

The average size of $h_{ii} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{2}{10} = 0.2$

From the table we can see the leverage of observation 9 has $h_{ii} = 0.820$ which is higher than other observations. If we take the nine observations (all the observations except the observation 9) as a one group, we can see that observation 7 is not much larger than the others.

According to the formula of h_{ii} ,an observation in simple linear regression has high leverage when it constitutes a large fraction of $\sum_{i=1}^{n} (x_i - \bar{x})^2$ or when n is a small enough.

In our example

$\bar{h} = 0.2$

$2\bar{h} = 2(0.2) = 0.4$

$h_{99} = 0.820$ exceed the criterion of twice the mean leverage , h_{99} is seperated by a substantial gap from the next largest leverage value $h_{77} = 0.194$

we identify case9 as outlying in terms of x values .

Iglewicz and Neter criterion help us to identify observations that have high leverage so we can investigate them further.

5.7 Multiple Regression[15]

In multiple regression h_{ii} will measure the distance between the point $x_i$and the centre of the data. The value of h_{ii}will be large for cases with unusual values for the independent variables,will tend to have larage values of h_{ii}.

There are many ways to define the distance for models with an intercept, we use the one that is defined in terms of deviations from means forms of the model.

Let $\bar{x}=(\bar{x_1}, \bar{x_2}, \ldots, \bar{x_p})^T$ be the $p \times 1$ vector of sample averages of the X's of the p independent variables, and let $\varkappa$ be the $n \times p$ matrix of the original data with the averages subtracted off, x_i^T is the ith row of X without the one with intercept the $(i,j)th$ element of $\varkappa$ is $x_{ij} - \bar{x_i}$

$$\varkappa^T\varkappa = \begin{bmatrix} \sum(x_{i1}-\bar{x}_1)^2 & \sum(x_{i1}-\bar{x_1})(x_{ip}-\bar{x_p}) \ldots\ldots \\ \sum(x_{i1}-\bar{x}_1)(x_{ip}-\bar{x_p}) \ldots & \sum(x_{ip}-\bar{x_p})^2 \end{bmatrix}$$

$$h_{ii} = \frac{1}{n} + (x_i - \bar{x})^T(\varkappa^T\varkappa)^{-1}(x_i - \bar{x})$$

5.8 Studentized residuals[15]

Since $var(\hat{e}_i) = \sigma^2(1 - h_{ii})$, if the value of h_{ii} is large then the value of $var(\hat{e}_i)$ will be small and if h_{ii} approaches to one, then the value $var(\hat{e}_i)$ will approach zero.

To help in determining the outliers we express residuals on a standard scale (A standard scale is the one with $\mu = 0$ and $\sigma = 1$).

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}}$$

If x_i is far from $\bar{x}$ then the model will fit good, and if x_i is near $\bar{x}$ then the model is bad fit.

when h_{ii} is large then the residual are small, we can improve the set of residuals by scaling so that the cases of with large h_{ii} will give large residuals and cases with smaller h_{ii} will give small residuals.

The studentized residuals $r_i$is defined as dividing the residuals by their standard error, where the standard error of the residual $= \sqrt{\hat{\sigma}^2 (1 - h_{ii})}$

Studentized residualsr_i have two properties :

1- They have mean zero and standard deviation one.

2-Studentized the residuals will show the outliers regardless of the leverage of the observations.

Most of the time we don't know the value of the variance σ^2 and thus we estimate σ^2 by the mean square error, as the standard error uses the estimated variance (MSE).

MSE is always a sum of squares divided by the appropriate degrees of freedom.

The studentized residuals $= \dfrac{\hat{e_i}}{\sqrt{MSE(1 - h_{ii})}}$

where

$MSE = \dfrac{\sum_{i=1}^{n} \hat{e_i}^2}{n - 2}$

5.9 The influence function[15]

Residuals and studentized residuals are used to measure the success or failure of fitting a model, however in some data sets, one of the cases may have sufficient impact upon the regression, such that if we delete this case then different results will be obtained .

Measuring the influence is made by deleting the *ith* case from the data and computing the parameters for the $n-1$ observations.

We find the difference between estimates of the parameters with the *ith* case included and with the *ith* case deleted.

Cases that give different estimates are called influential.

5.10 The empirical influence function [15]

Usually we are interested to know the change in the estimate of β when a case i is deleted from the data, to achieve this we first compute the value of $\hat{\beta}$ from the full data, then let $\hat{\beta}_{-i}$ denote the value of $\hat{\beta}$ when case i is deleted from the data. We find the difference between $\hat{\beta}_{-i} - \hat{\beta}$.

Acase i is called influential if when deleting this case there is a substantial change in the estimate of the parameter vector. where estimate of β is

$$\hat{\beta}_{-i} = \hat{\beta} - \frac{(X^T X)^{-1} x_i \hat{e}_i}{1 - h_{ii}}$$

$$y_i - \tilde{y}_i = \frac{\hat{e}_i}{1 - h_{ii}}$$

Estimate of σ^2

$$\hat{\sigma}^2_{-i} = \frac{1}{n - \acute{p} - 1} \hat{\sigma}^2 (n - \acute{p} - r_i^2)$$

5.11 Testing for outliers[15]

Suppose that the *ith* case is suspected to be an outlier, then the following procedure may be used to test that case :

1-Delete the *ith* case from the data set.

2-Let $\hat{\beta}_{-i}$ and $\hat{\sigma}^2_{-i}$ be the estimates of β and σ^2using the $n-1$ data , $\hat{\sigma}^2_{-i}$ has $n-\acute{p}-1$ degrees of freedom.

3-we find the value of $\tilde{y}_i = x_i^T \hat{\beta}_{-i}$ for the deleted case

$$var(\tilde{y}_i) = \sigma^2 x_i^T (X_{-i}^T X_{-i}^T)^{-1} x_i$$

where X_{-i} is the matrix X with the *ith* row deleted .

we estimate $var(\tilde{y}_i)$ by replacing the value of σ^2 by $\hat{\sigma}^2_{-i}$

so $$var(\tilde{y}_i) = \hat{\sigma}^2_{-i}\, x_i^T (X_{-i}^T X_{-i}^T)^{-1} x_i$$

4-If $E(y_i - \tilde{y}_i) \neq 0$ then y_i is an outlier.

and if $E(y_i - \tilde{y}_i) = 0$ then y_i is not an outlier.

$var(y_i - \tilde{y}_i) = \sigma^2(1 + x_i^T (X_{-i}^T X_{-i})^{-1} x_i)$

If we assume normal errors then we test the hypothesis

$E(y_i - \tilde{y}_i) = 0$ by the student t test

$$t_i = \frac{y_i - \tilde{y}_i}{\hat{\sigma}_{-i}\sqrt{1 + x_i^T (X_{-i}^T X_{-i})^{-1} x_i}}$$ with $n-\acute{p}-1$ degrees of freedom.

t_i can be computed as

$$t_i = r_i \sqrt{\frac{n-\acute{p}-1}{n-\acute{p}-r_i^2}}$$

If X is a matrix of order $n \times \acute{p}$ then X_{-i} is of order $(n-1) \times \acute{p}$

we use the identification

$$(X_{-i}^T X_{-i})^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - h_{ii}}$$

we can apply the above formula for the regression without case i , the distance between x_i and the centre of the remaining $n-1$ case is defined by $x_i^T (X_{-i}^T X_{-i})^{-1} x_i$, if we apply the identity then

$$\begin{aligned}
x_i^T (X_{-i}^T X_{-i})^{-1} x_i &= x_i^T (X^T X)^{-1}\, x_i + \frac{x_i^T (X^T X)^{-1} x_i x_i^T (X^T X)^{-1} x_i}{1 - h_{ii}} \\
&= h_{ii} + \frac{h_{ii}^2}{1 - h_{ii}} \\
&= \frac{h_{ii} - h_{ii}^2 + h_{ii}^2}{1 - h_{ii}} \\
&= \frac{h_{ii}}{1 - h_{ii}} \\
t_i &= \frac{\dfrac{\hat{e}_i}{1 - h_{ii}}}{\hat{\sigma}_{-i}\sqrt{1 + \dfrac{h_{ii}}{1 - h_{ii}}}} \\
&= \frac{\hat{e}_i}{\hat{\sigma}_{-i}(1 - h_{ii})\sqrt{\dfrac{1}{1 - h_{ii}}}} \\
&= \frac{\hat{e}_i \sqrt{n - \acute{p} - 1}}{\hat{\sigma}\sqrt{n - \acute{p} - r_i^2}\sqrt{\dfrac{(1 - h_{ii})^2}{1 - h_{ii}}}} \\
&= \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \times \frac{\sqrt{n - \acute{p} - 1}}{\sqrt{n - \acute{p} - r_i^2}}
\end{aligned}$$

$$= r_i \sqrt{\frac{n-\acute{p}-1}{n-\acute{p}-r_i^2}}$$

If the ith case is suspected to be an outlier then t_i must be compared to the central t-distribution with an appropriate number of degrees of freedom.

We use a special table of $t-$distribution to find the critical values, we choose α level (0.01 *or* 0.05)

Example[15]

Forbe's data, giving boiling point (F^0) and barometric pressure (inches of mercury) for 17 locations in the Alps and in Scotland.

case number	boiling point (F^0)	pressure (in Hg)	log (pressure)	100× log(pressure)
1	194.5	20.79	1.3179	131.79
2	194.3	20.79	1.3179	131.79
3	197.9	22.4	1.3502	135.02
4	198.4	22.67	1.3555	135.55
5	199.4	23.15	1.3646	136.46
6	199.9	23.35	1.3683	136.83
7	200.9	23.89	1.3782	137.82
8	201.1	23.99	1.3800	138.00
9	201.4	24.02	1.3806	138.06
10	201.3	24.01	1.3805	138.05
11	203.6	25.14	1.4004	140.04
12	204.6	26.57	1.4244	142.44
13	209.5	28.49	1.4547	145.47
14	208.6	27.76	1.4434	144.34
15	210.7	29.04	1.4630	146.30
16	211.9	29.88	1.4754	147.54
17	212.2	30.06	1.4780	147.80

number of cases	x_i	y_i	$\hat{y}_i$	$\hat{e}_i$
1	194.50	131.79	132.04	-0.25
2	194.30	131.79	131.86	-0.07
3	197.90	135.02	135.08	-0.06
4	198.4	135.55	135.53	0.02
5	199.40	136.46	136.42	0.04
6	199.90	136.83	136.87	-0.04
7	200.90	137.82	137.77	0.05
8	201.10	138	137.95	0.05
9	201.4	138.06	138.22	-0.16
10	201.30	138.05	138.13	-0.08
11	203.60	140.04	140.19	-0.15
12	204.60	142.44	141.08	1.36
13	209.50	145.47	145.47	0.00
14	208.60	144.34	144.66	-0.32
15	210.70	146.30	146.54	-0.24
16	211.90	147.54	147.62	-0.08
17	212.20	147.80	147.89	-0.09

$\bar{x}$= 202.95294118 $S_{xx} = 530.78235294$

$\bar{y}$= 139.60588235 $S_{yy} = 427.76281177$

$S_{xy} = 475.29570589$

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n-2} = \frac{2.1582}{15} = 0.14388$$

$$\hat{\sigma} = 0.379$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0.895$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -42.131$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$= -42.131 + 0.895x_i$$

$$h_{12,12} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

$$= \frac{1}{17} + \frac{(204.6 - 202.95294118)^2}{530.78235294}$$

$$= \frac{1}{17} + \frac{2.71280302}{530.78235294}$$

$$= 0.063934481$$

$$\approx 0.0639$$

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

$$r_{12} = \frac{1.36}{0.379\sqrt{1 - 0.0639}}$$

$$= \frac{1.36}{0.379\sqrt{0.9361}}$$

$$= \frac{1.36}{0.36669}$$

$$= 3.7078$$

To test the outlier we find t_{12}

$$t_i = r_i\sqrt{\frac{n-\acute{p}-1}{n-\acute{p}-r_i^2}}$$

$$t_{12} = 3.7078\sqrt{\frac{17-2-1}{17-2-(3.7078)^2}}$$

$$= 12.40$$

number of degrees of freedom $= 14$

$$n = 17$$

$$\acute{p} = 2$$

The critical value for $\alpha = 0.05$ from **table -11** $= 3.59$

this value exceeds the value of t_{12}, so we reject the hypothesis therefore the number 204.6 is an outlier

Chapter 6

Conclusion

Outliers are those values lying very far from the rest of the data, it may exist from measurement errors , mistake in entering values or simply a typing error when inserting the data. Some times outliers are not erroneous but they are different from the rest.

A small proportion of outliers can distort the results based on classical procedures and thus checking for outliers is very important for any data analysis, exploring why outliers exist can give many hints to the development of better models .

Once outliers are found, one have to decide what to do with them, neither ignoring nor deleting them are good solutions.

If we do nothing about them, we will not get a good model describing the data or even the bulk of the data. Accordingly data should be analysed carefuly to investigate the sources of outliers before deciding to eliminate or dicard or retain them.

If we decide to retain outliers, then we have to be careful and we have to use robust statistical procedures which minimize the effect of the outliers on the sample data.

Traditional nonrobust measures will not be helpful in this case as they are affected by the presence of outliers.

There are several procedures for outlier identification, some are highly effected by swamping or masking such as the L_r *test* and the Dixon type test, others require specific knowledge of the number and location of outliers and thus such procedures are not recommended, instead the ESD perform specially well under each of the above conditions and is relatively easy to compute and to interpret and thus this procedoure is much recommended in several studies.

Even one or more outliers have a great effect on classical estimates of location and scale and might cause considerable damage to it.

A mong the few estimators of location that accommodate outliers, a special attention was paid to the trimmed mean whcih is easy to compute and has a high effeciency relative to the mean. Moreover, trimmed mean leads to confidence intervals, which are superior in the presence of outliers despite they are slightly more difficult to compute than the usual t-interval.

Several scale estimators were discussed, however the effeciency of those outliers in the presence of outliers varied considerably, for example standard deviation S is not effecient , while the biweight estimator of scale S_{bi} is more effecient.

Outliers can distort the fitted linear regression lines and may shift them as to lead to a systematic pattern of deviations from the fitted line especially when the data set with a small number of observations is being considered, this is evident in the case of scatter plots where all the observations except the outlier fall a round a straight line statistical

relationship and the residual plot may illustrate the lack of fittness in the regression model and will clearly flag the outlier.Moreover the outlying cases may involve large residuals and often have a dramatic effects on the fitted least square regression.

Bibliography

[1] Barnett, V., and Lewis, T., (1994) 'Outlier in Statistical Data' 3rd edition, John Wiley and Sons, Chichester.

[2] Berry, D., (1996) 'Statistics A Bayesian Perspective',Duxbury Press, Belmont.

[3] Clarke, G., and Cooke, D. (1978) 'A Basic Course in Statistics' Edward Arnold, London.

[4] David, A. H., (1981) 'Order Statistics' second edition, John Wiley and Sons, New York.

[5] Hair, J., Anderson, R and Black, W. (1995) 'Multivariate Data Analysis' fifth edition, DuxBury, Australia.

[6] Hawkins, D, M., (1980) 'Identification of Outliers' Chapman and Hall, London.

[7] Iglewicz, B., and Hooglin, D. C., (1993) 'How to Detect and Handle Outliers', John Wiley and Sons, New York.

[8] Kendella, M., and Stuart, A., (1979) 'The advanced theoty of Statistics' fourth edition, Charles Griffen & Co. Ltd, London.

[9] Kitchens, J., L,(1998) 'Exploring Statistics' 2nd edition, Brooks/Cole, New York.

[10] Lyman L. R., (2001) 'Statistical Methods and Data Analysis' fifth edition, DuxBury, Australia.

[11] Mann, P., (1998) 'Introductory Statistics' third edition, John Wiley and Sons, New York.

[12] Neter, J., Wasserman, W and Kutner, M (1989) 'Applied Linear Regression Models', second edition, Irwin, Burr Ridge.

[13] Rosner, B., (2000) 'Fundamentals of Biostatistics' fifth edition, Duxbury Resource Center, Australia.

[14] Wayne, D., (1999) 'Biostatistic: Afoundation for analysis in the health sciences' seven edition, John Wiley and Sons, New York.

[15] Weisberg, S., (1980) 'Applied Linear Regression', John Wiley and Sons, New York.

[16] High R. "Dealing with Outliers."

2000, http://cc.uoregon.edu/cnews/spring2000/outliers.html

[17] Mean "Log Transformations"

1999, http://www.cmh.edu/stats/graph/grlog20.html

[18] Wessel P. "Robust Estmation"

2000, http://www.soest.hawaii.edu/wessel/courses/gg313DA_book/node37.html

[19] Wuder M. "Data Transformation"

2001, http://www.pcf.foresty.ca/notice_e.html

[20] "Outliers"

http://www.gseis.ucla.edu/courses/ed2319/lecture/course5N.doc

Appendix A

Figures

Figure 1 Box plot for a hypothetical data set

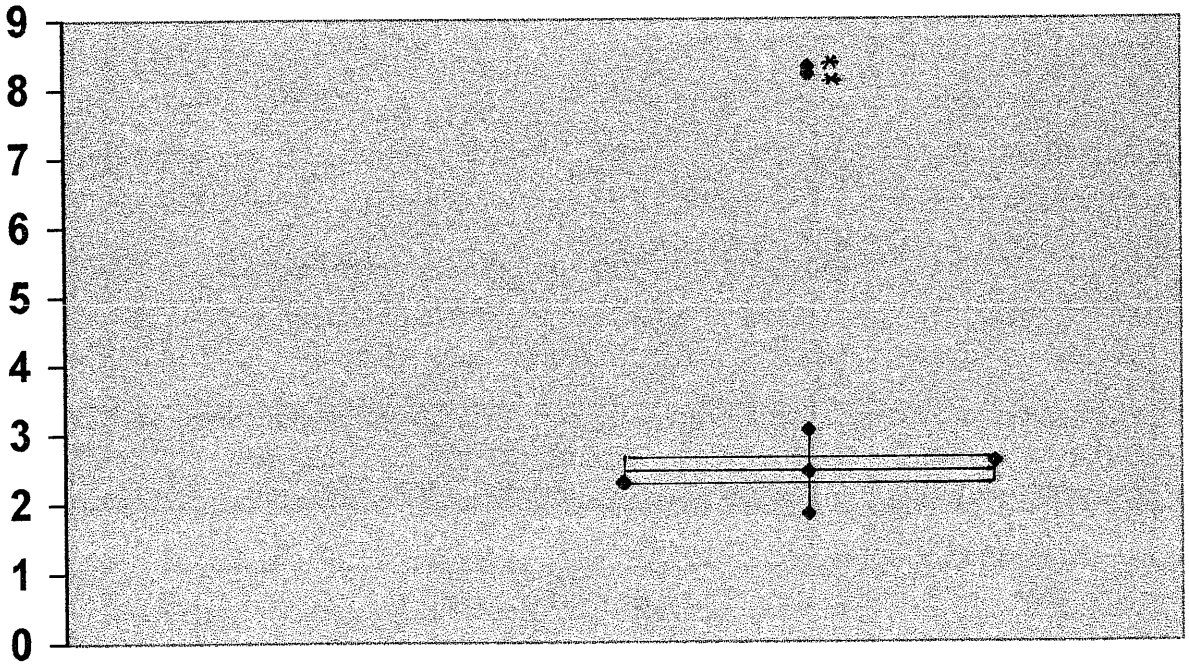


Figure 2 Dot plot for example worsted yarn
for lengths 300mm vs 250mm

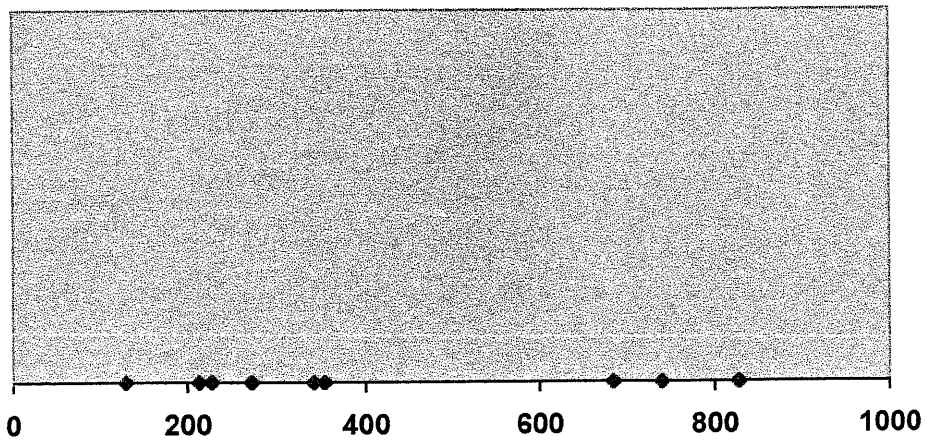


Figure 3 Dot plot for example worsted yarn
Log cycles for lengths of 300mm vs 250mm

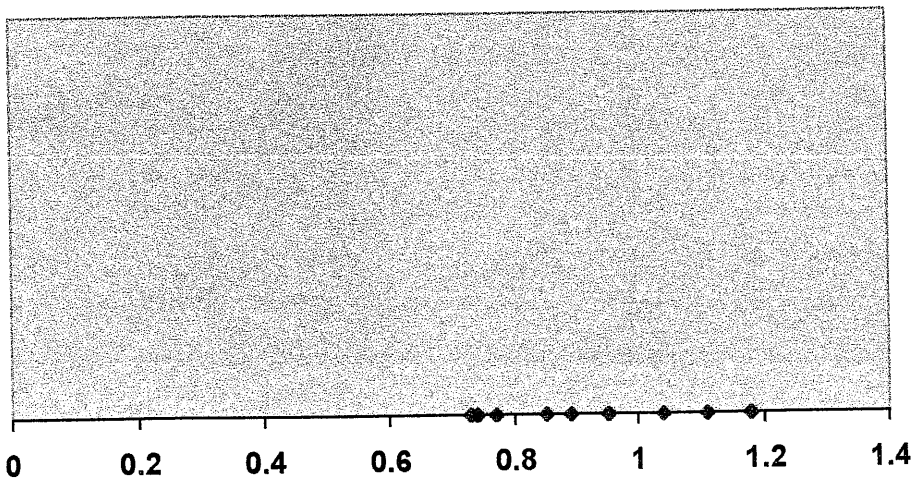


Figure 4 Dot plot categorizing drug effect for example heart beats

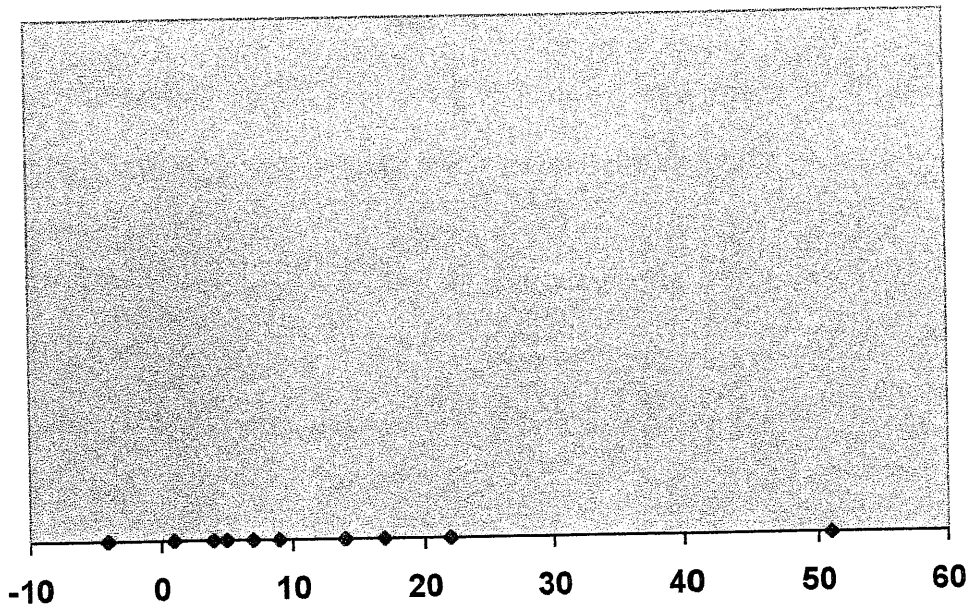


Figure 5 Dot plot for example worsted yarn" comparing yarn lengths of 300 mm vs 250 mm"

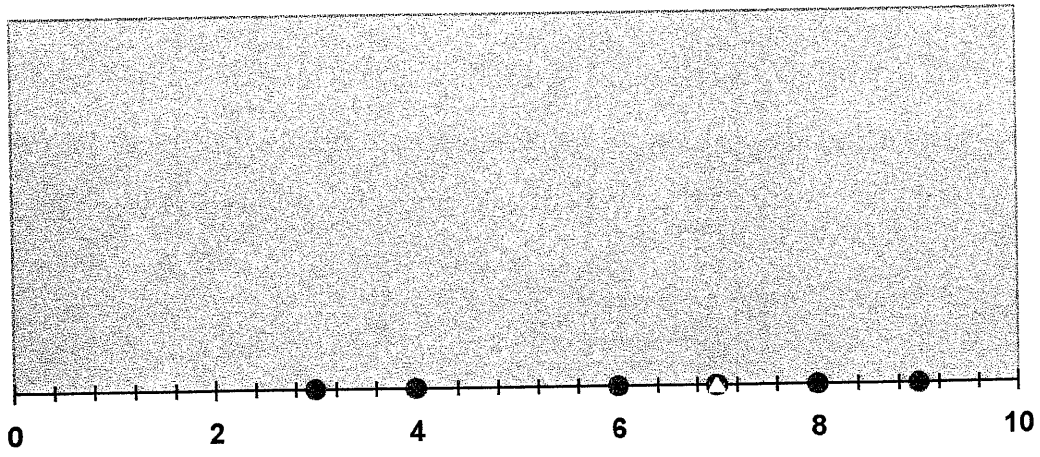


Figure 6 Dot plot for a hypothetical data set

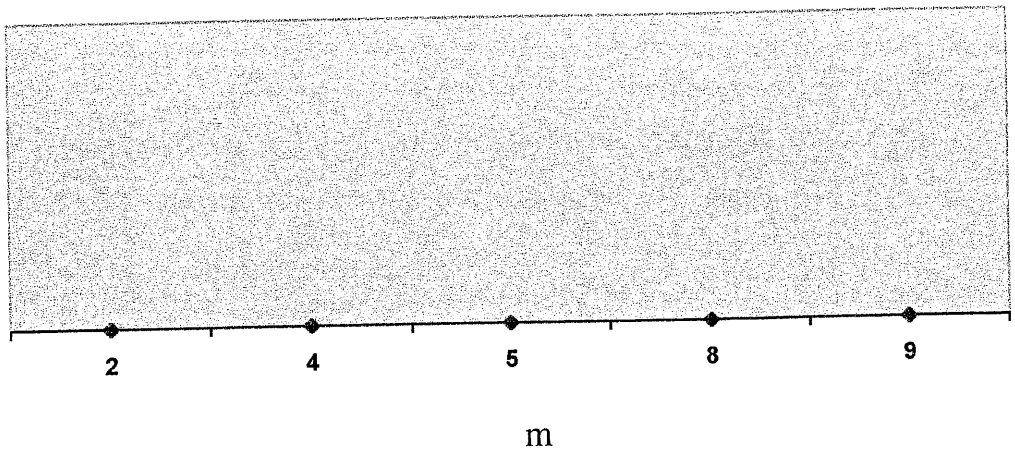


Figure 7 Dot plot for the square root of the hypothetical data set (shown in figure 6)

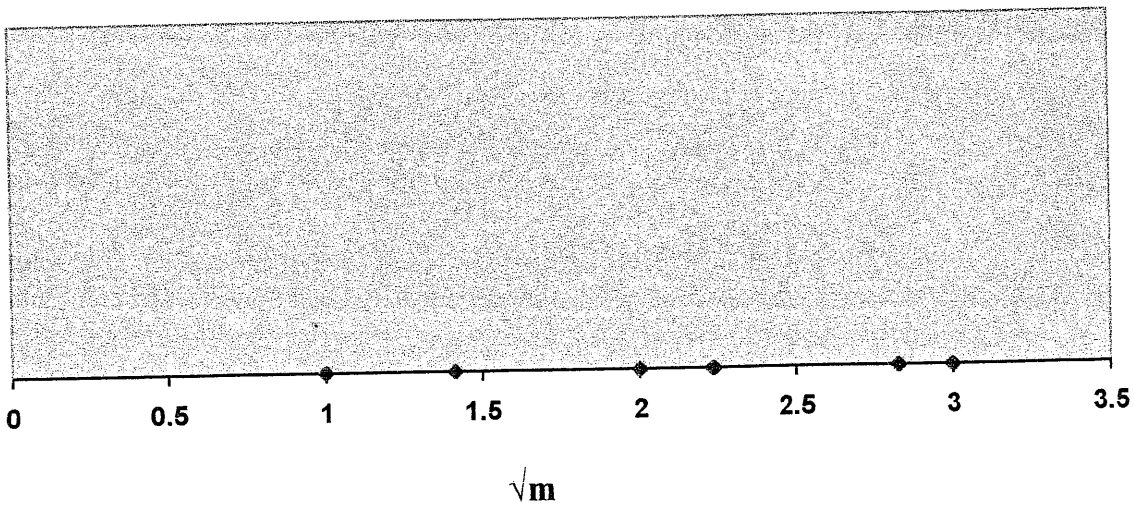


Figure 8 Scatter plot example

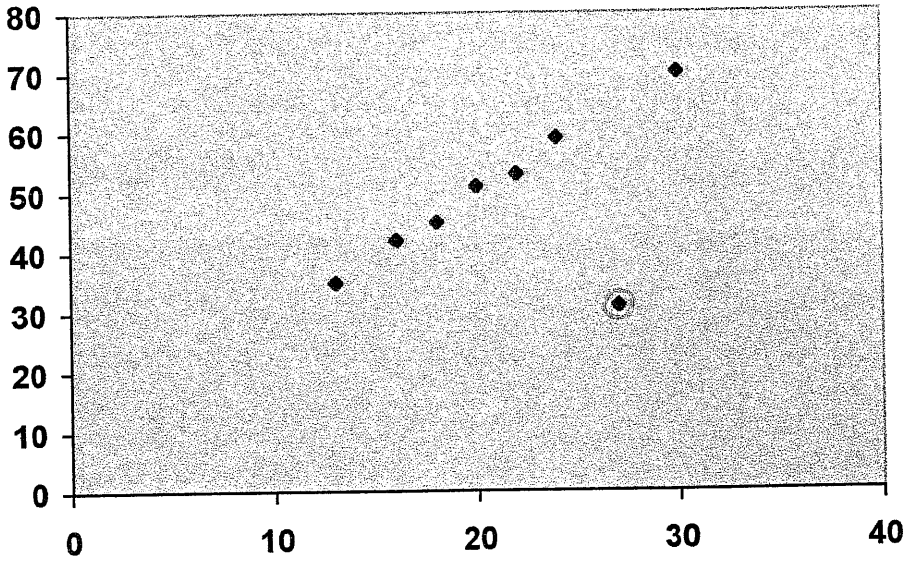


Figure 9-a Scatter of a hypothetical data

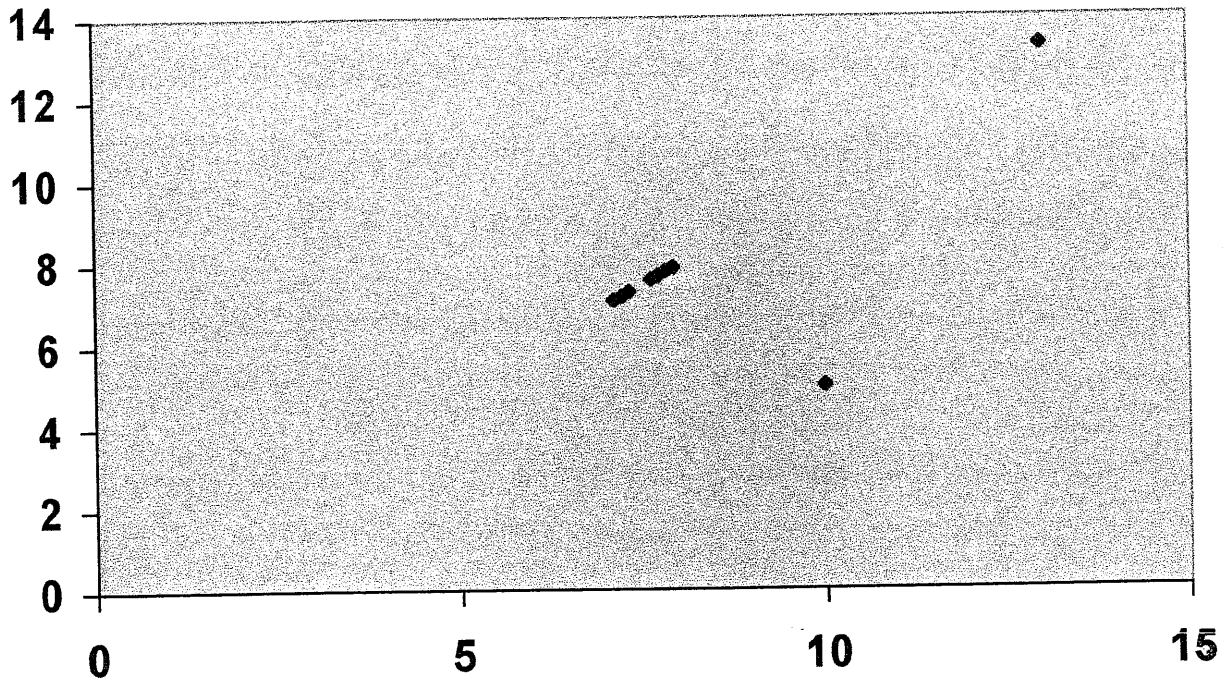


Figure 9-b The data with least- squares line (solid) and the robust line dashed.

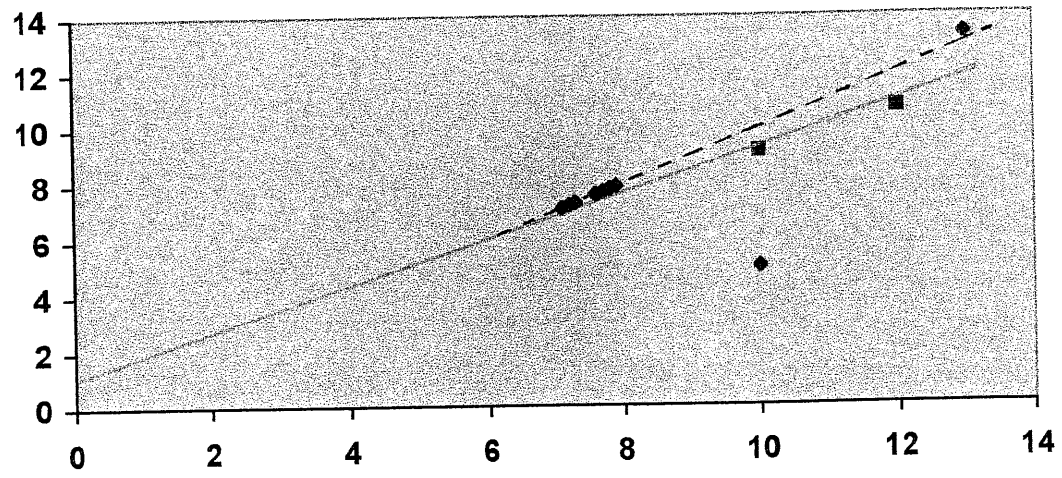


Figure 9-c The data (in figure 9-b) and the two fitted lines after removing the point (13,13.3)

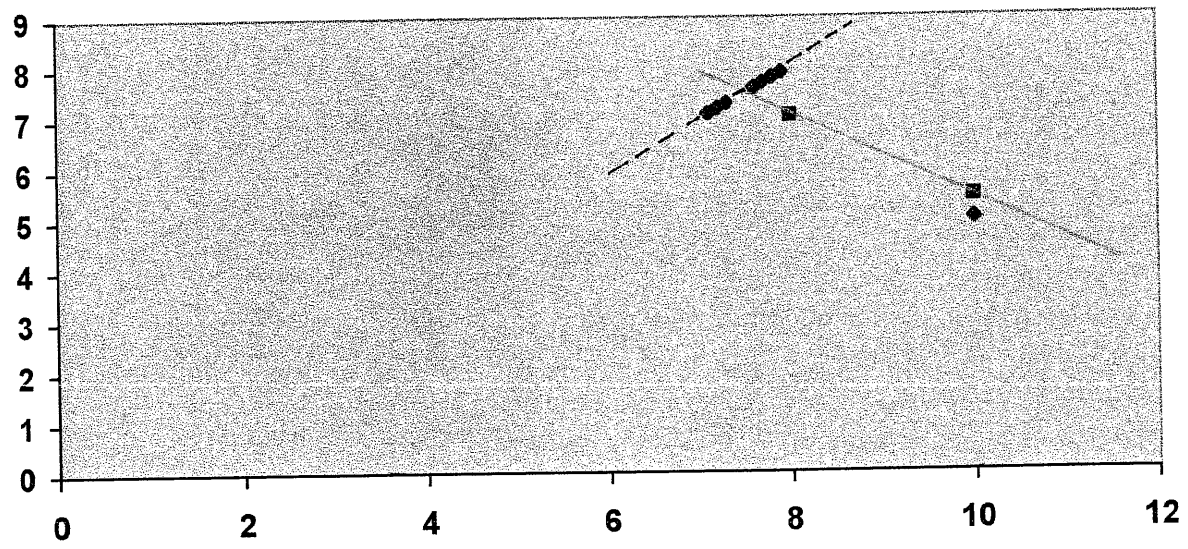
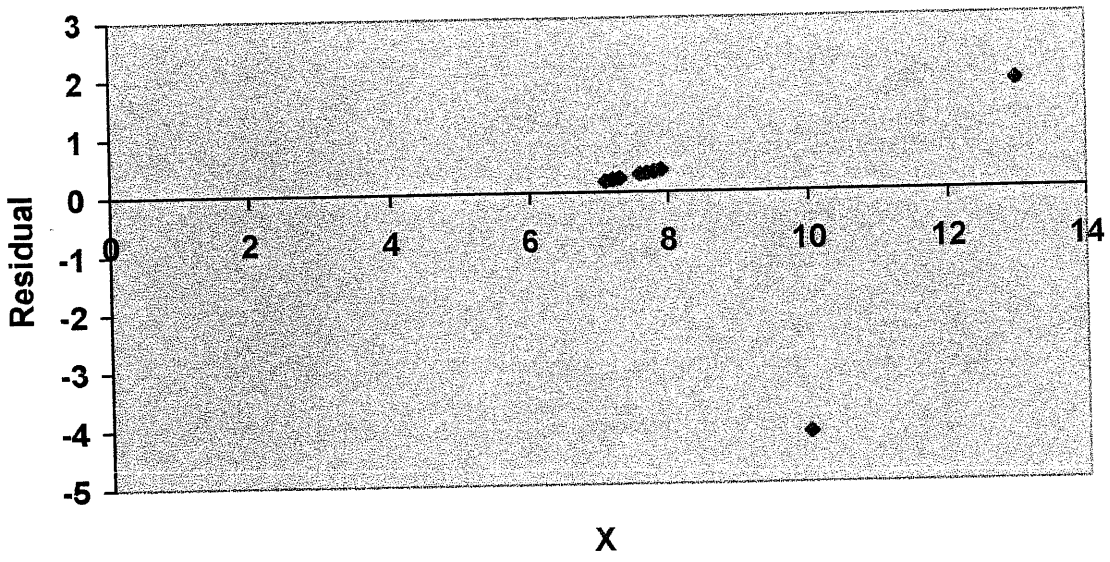


Figure 10 Residual plot



follow linear regression

Figure 11 Standardized residuals plot

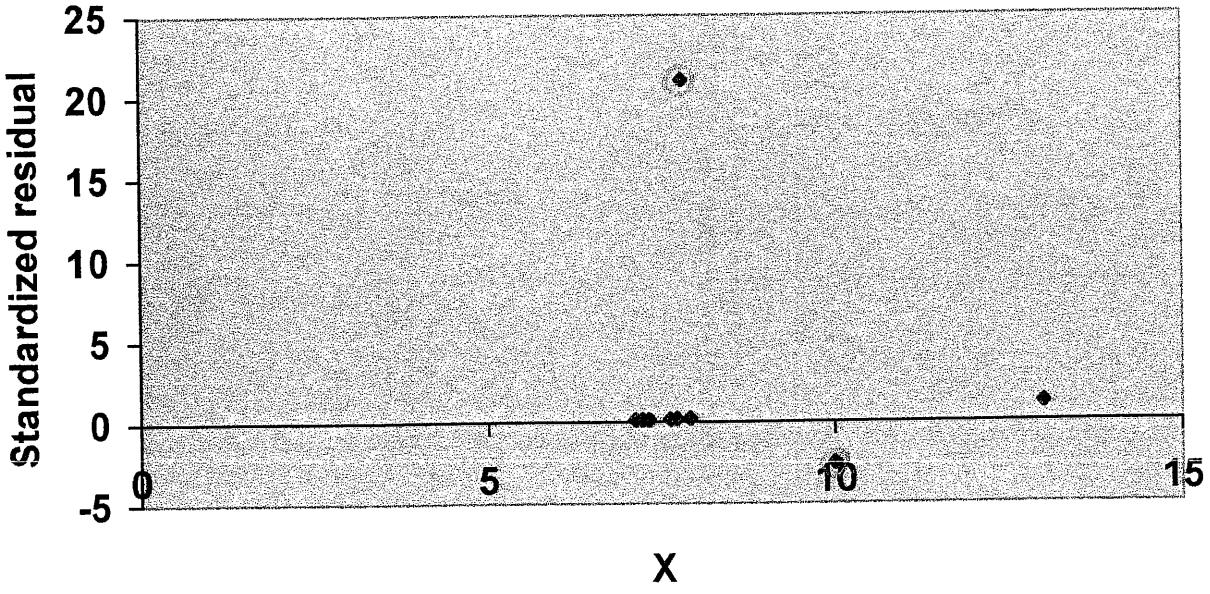


Table - 1

n	10%	5%	0.025
10	3.41	4.19	5.08
20	3.26	3.78	4.25
40	3.22	3.60	3.92

- The values of D that cause the stated proportion of normal random observations.
- This table is taken from the book "How to Detect and Handle Outliers"

Table - 2

Critical values for the ESD (Extreme Studentized Deviate)

Outlier statistic $(ESD_{n,1-\alpha}, \alpha = 0.05, 0.01)$

	1-α			1-α	
n	0.95	0.99	n	0.95	0.99
5	1.72	1.76	25	2.82	3.14
6	1.89	1.97	26	2.84	3.16
7	2.02	2.14	27	2.86	3.18
8	2.13	2.28	28	2.88	3.20
9	2.21	2.39	29	2.89	3.22
10	2.29	2.48	30	2.91	3.24
11	2.36	2.56	35	2.98	3.32
12	2.41	2.64	40	3.04	3.38
13	2.46	2.70	45	3.09	3.44
14	2.51	2.75	50	3.13	3.48
15	2.55	2.81	60	3.20	3.56
16	2.59	2.85	70	3.26	3.62
17	2.62	2.90	80	3.31	3.67
18	2.65	2.93	90	3.35	3.72
19	2.68	2.97	100	3.38	3.75
20	2.71	3.00	150	3.52	3.89
21	2.73	3.03	200	3.61	3.98
22	2.76	3.06	300	3.72	4.09
23	2.78	3.08	400	3.80	4.17
24	2.80	3.11	500	3.86	4.23

This table is taken from the book "Fundementals of Biostatistics"

Table 3

1% and 5% crictical values for three versions of Dixon tests

	λ_{11}		λ_{21}		λ_{10}	
n	1%	5%	1%	5%	1%	5%
4	.991	.995			.992	.831
5	.916	.807	.995	.976	.831	.717
6	.805	.689	.951	.872	.737	.621
7	.740	.610	.885	.780	.694	.570
8	.683	.554	.829	.710	.638	.524
9	.635	.512	.776	.657	.594	.492
10	.597	.477	.726	.612	.564	.464
11	.566	.450	.679	.576		
12	.541	.428	.642	.546	.520	.429
13	.520	.410	.615	.521		
14	.502	.395	.593	.501	.485	.397
15	.486	.381	.574	.483		
16	.472	.369	.557	.467	.461	.376
17	.460	.359	.542	.453		
18	.449	.349	.529	.440	.438	.354
19	.439	.341	.517	.428		
20	.430	.334	.506	.419	.417	.340
21	.421	.327	.496	.410		
22	.414	.320	.487	.402		
23	.407	.314	.479	.395		
24	.400	.309	.471	.388		
25	.394	.304	.464	.382	.386	.316
26	.389	.299	.457	.376		
27	.383	.295	.450	.370		
28	.378	.291	.444	.365		
29	.374	.287	.438	.360		
30	.369	.283	.433	.355	.368	.300

This table is taken from the book "How to Detect and Handle Outliers'

Table 4

Values of K for boxplot rule with specified probability of at least one outside observation per random normal sample.

n	0.05	0.1
5	5.6	3.7
6	2.9	2.0
7	3.0	2.3
8	2.2	1.8
9	3.3	2.7
10	2.4	2.0
11	2.7	2.2
12	2.2	1.8
13	2.8	2.3
14	2.5	2.0
15	2.5	2.1
16	2.3	1.9
17	2.6	2.2
18	2.4	2.0
19	2.6	2.2
20	2.2	1.9
30	2.2	2.0
40	2.2	2.0
50	2.2	2.0
75	2.3	2.1
100	2.2	2.1
200	2.4	2.2
300	2.4	2.2

This table is taken from the book "How to Detect and Handle Outliers"

Table 5

L_r for normal samples $\alpha = 0.05$

n	r 1	2	3	4	5	6	7	8	9	10
5	0.1239	0.0187								
6	0.1971	0.0561	0.0096							
7	0.2656	0.1025	0.0341							
8	0.3195	0.1446	0.0635	0.0218						
9	0.3772	0.1922	0.0985	0.0460						
10	0.4187	0.2297	0.1311	0.0706	0.0331					
15	0.5613	0.3850	0.2742	0.1964	0.1398	0.0975	0.0658			
20	0.6381	0.4812	0.3761	0.2971	0.2366	0.1889	0.1483	0.1153	0.0886	0.0666
25	0.6916	0.5492	0.4500	0.3746	0.3129	0.2627	0.2206	0.1852	0.1545	0.1270
30	0.7321	0.6013	0.5098	0.4381	0.3782	0.3265	0.2835	0.2457	0.2132	0.1848
35	0.7615	0.6428	0.5544	0.4848	0.4277	0.3790	0.3363	0.2990	0.2666	0.2362
40	0.7855	0.6721	0.5901	0.5256	0.4701	0.4224	0.3813	0.3441	0.3119	0.2821
45	0.8035	0.7000	0.6220	0.5584	0.5050	0.4594	0.4195	0.3834	0.3513	0.3223
50	0.8181	0.7209	0.6479	0.5878	0.5366	0.4911	0.4516	0.4160	0.3850	0.3565
60	0.8428	0.7553	0.6884	0.6335	0.5861	0.5448	0.5074	0.4743	0.4434	0.4154
70	0.8610	0.7833	0.7222	0.6715	0.6276	0.5892	0.5536	0.5215	0.4918	0.4642
80	0.8746	0.8021	0.7453	0.6980	0.6576	0.6209	0.5873	0.5569	0.5291	0.5033
90	0.8863	0.8195	0.7668	0.7227	0.6834	0.6486	0.6163	0.5880	0.5611	0.5365
100	0.8957	0.8333	0.7844	0.7419	0.7057	0.6729	0.6432	0.6150	0.5898	0.5656

n = number of observations

r = number of outliers

This table is taken from the book "Identification of Outliers"

Table 6

Critical values for 5% and 1% tests of discordancy for k outliers in a normal sample, using Tietjen and M($E_{k\text{- Statistic}}$

n	k=2		k=3		k=4	
	5%	1%	5%	1%	5%	1%
5	0.010	0.002				
6	0.034	0.012				
7	0.065	0.028	0.016	0.006		
8	0.099	0.050	0.034	0.014		
9	0.137	0.078	0.057	0.026	0.021	0.009
10	0.172	0.101	0.083	0.044	0.037	0.018
12	0.234	0.159	0.133	0.083	0.073	0.042
14	0.293	0.207	0.179	0.123	0.112	0.072
16	0.340	0.263	0.227	0.166	0.153	0.107
18	0.382	0.306	0.267	0.206	0.187	0.141
20	0.416	0.339	0.302	0.236	0.221	0.170
30	0.549	0.482	0.443	0.386	0.364	0.308
40	0.629	0.574	0.534	0.480	0.458	0.408
50	0.684	0.636	0.599	0.550	0.529	0.482

n= number of observation
k= number of outliers

This table is taken from the book "Outliers in Statistical Data "

Table -7

Approximate 10%, 5% and 1% critical values of b_2 the sample kurtosis in samples from a normal distribution

n	α 0.1	0.05	0.01
7	3.20	3.55	4.23
8	3.31	3.70	4.53
9	3.43	3.86	4.82
10	3.53	3.95	5.00
12	3.55	4.05	5.20
15	3.62	4.13	5.30
20	3.68	4.17	5.36
25	3.68	4.16	5.30
30	3.68	4.11	5.21
35	3.68	4.10	5.13
40	3.67	4.06	5.04
45	3.65	4.00	4.94
50	3.62	3.99	4.88

This table is taken from the book
" How to Detect and Handle Outliers"

Table 8

Values of the constants $a_{n,n-i+1}$ required for calculating Shaprio and Wilk's W-statistic

n \ i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
3	0.7010																
4	0.6872	0.1677															
5	0.6646	0.2413															
6	0.6431	0.2806	0.0875														
7	0.6233	0.3031	0.1401														
8	0.6052	0.3164	0.1743	0.0561													
9	0.5888	0.3244	0.1976	0.0947													
10	0.5739	0.3291	0.2141	0.1224	0.0399												
12	0.5475	0.3325	0.2347	0.1586	0.0922	0.0303											
14	0.5251	0.3318	0.2460	0.1802	0.1240	0.0727	0.0240										
16	0.5056	0.3290	0.2521	0.1939	0.1447	0.1005	0.0593	0.0196									
18	0.4886	0.3253	0.2553	0.2027	0.1587	0.1197	0.0837	0.0496	0.0163								
20	0.4734	0.3211	0.2565	0.2085	0.1686	0.1334	0.1013	0.0711	0.0422	0.0140							
25	0.4450	0.3069	0.2543	0.2148	0.1822	0.1539	0.1283	0.1046	0.0823	0.0610	0.0403	0.0200					
30	0.4254	0.2944	0.2487	0.2148	0.1870	0.1630	0.1415	0.1219	0.1036	0.0862	0.0697	0.0537	0.0381	0.0227	0.0076		
35	0.4096	0.2834	0.2427	0.2127	0.1883	0.1673	0.1487	0.1317	0.1160	0.1013	0.0873	0.0739	0.0610	0.0484	0.0361	0.0239	0.0119
40	0.3964	0.2737	0.2368	0.2098	0.1878	0.1691	0.1526	0.1376	0.1237	0.1108	0.0986	0.0870	0.0759	0.0651	0.0546	0.0444	0.0343
45	0.3850	0.2651	0.2313	0.2065	0.1865	0.1695	0.1545	0.1410	0.1286	0.1170	0.1062	0.0959	0.0860	0.0765	0.0673	0.0584	0.0497
50	0.3751	0.2574	0.2260	0.2032	0.1847	0.1691	0.1554	0.1430	0.1317	0.1212	0.1113	0.1020	0.0932	0.0846	0.0764	0.0685	0.0608

n \ i	18	19	20	21	22	23	24	25
40	0.0244	0.0146	0.0049					
45	0.0412	0.0328	0.0245	0.0163	0.0082			
50	0.0532	0.0459	0.0386	0.0314	0.0244	0.0174	0.0104	0.0035

n= number of observations

This table is taken from the book " Outliers in Statistical Data"

Table 9

Critical values for 5% and 1% tests for the presence of an undefined number of contminants in a normal sample using Shapiro and Wilk's W- statistic

n	5%	1%
3	0.767	0.753
4	0.748	0.687
5	0.762	0.686
6	0.788	0.713
7	0.803	0.730
8	0.818	0.749
9	0.829	0.764
10	0.842	0.781
12	0.859	0.805
14	0.874	0.825
16	0.887	0.844
18	0.897	0.858
20	0.905	0.868
25	0.918	0.888
30	0.927	0.900
35	0.934	0.910
40	0.940	0.919
45	0.945	0.926
50	0.947	0.930

n= number of observations

This table is taken from the book "Outliers in Statical Data"

Table 10

Percentage points of the t distribution $t_{(d,u)}^{a}$

Degrees of freedom, d	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995	0.9995
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

This table is taken from the book "Fundamentals of Biostatistics"

Table 11

Critical values for the outliers test

The values in the table in row n and column p ´are t (α/n;n-p ´) for the chioce of α= 0.05

Critical values for outlier test,α = 0.05

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30
6	4.85	6.23	10.89	76.39														
7	4.38	5.07	6.58	11.77	89.12													
8	4.12	4.53	5.26	6.90	12.59	101.90												
9	3.95	4.22	4.66	5.44	7.18	13.36	114.6											
10	3.83	4.03	4.32	4.77	5.60	7.45	14.09	127.30										
11	3.75	3.90	4.10	4.40	4.88	5.75	7.70	14.78	140.10									
12	3.69	3.81	3.96	4.17	4.49	4.98	5.89	7.94	15.44	152.80								
13	3.65	3.74	3.86	4.02	4.24	4.56	5.08	6.02	8.16	16.08	165.50							
14	3.61	3.69	3.79	3.91	4.07	4.30	4.63	5.16	6.14	8.37	16.69	178.20						
15	3.58	3.65	3.73	3.83	3.95	4.12	4.36	4.70	5.25	6.25	8.58	17.28	191.00					
16	3.56	3.62	3.68	3.77	3.87	4.00	4.17	4.41	4.76	5.33	6.36	8.77	17.85	203.70				
17	3.54	3.59	3.65	3.72	3.80	3.90	4.04	4.21	4.46	4.82	5.40	6.47	8.95	18.40	216.40			
18	3.53	3.57	3.62	3.68	3.75	3.83	3.94	4.08	4.26	4.51	4.88	5.47	6.57	9.13	18.93			
19	3.52	3.56	3.60	3.65	3.71	3.78	3.86	3.97	4.11	4.30	4.55	4.93	5.54	6.67	9.30			
20	3.51	3.54	3.58	3.62	3.67	3.73	3.81	3.89	4.00	4.15	4.33	4.59	4.98	5.60	6.76			
21	3.50	3.53	3.57	3.60	3.65	3.70	3.76	3.83	3.92	4.03	4.18	4.37	4.64	5.03	5.67			
22	3.50	3.52	3.55	3.59	3.63	3.67	3.72	3.78	3.86	3.95	4.06	4.21	4.40	4.68	5.08	280.10		
23	3.49	3.52	3.54	3.57	3.61	3.65	3.69	3.75	3.81	3.88	3.98	4.09	4.24	4.44	4.71	21.41		
24	3.49	3.51	3.53	3.56	3.59	3.63	3.67	3.71	3.77	3.83	3.91	4.00	4.12	4.27	4.47	10.07		
25	3.48	3.50	3.53	3.55	3.58	3.61	3.65	3.69	3.73	3.79	3.85	3.93	4.02	4.14	4.30	7.17		
26	3.48	3.50	3.52	3.54	3.57	3.60	3.63	3.66	3.70	3.75	3.81	3.87	3.95	4.05	4.17	5.95		

This tasble is taken from the book"Applied Linear Regression"

في الفصل الرابع :
تمت مناقشة طرق احتواء الشواذ في البيانات ذات التوزيع الطبيعي مثل الوسط (Trimmed) و الوسط (Wonsorized) و مقاييس التوقع و فترات الثقة.
و لم يقتصر الأمر على ذلك بل تعداه إلى دراسة البيانات ذات التوزيع غير الطبيعي عن طريق تحويل البيانات لتماثل التوزيع الطبيعي و ذلك باستخدام اللوغاريتمات و الجذور التربيعي .
في الفصل الخامس :
تم مناقشة تأثير الشواذ على نماذج الإنحدار الخطي كما تم توضيح طرق تحديد الشواذ مخطط الانتشار (Scater Plots) وطريقة البواقي (Residual Plots)
ولقد تم تطبيق جميع الطرائق السابقة على بيانات افتراضية وتم توضيح تأثير الشواذ في كل حالة .
وفي الخاتمة فإننا نستنتج أن دراسة الشواذ واستكشاف وجودها مهم جدا" لتحليل النتائج في جميع نواحي البحث العلمي ولذلك فإن على جميع الإحصائيين والعلماء والباحثين في مختلف المجالات على معرفة تامة بالقيم الشواذ ويجب عليهم إعطاء اهتمام خاص بها كونها تؤدي إلى تشويه وتحريف الاستنتاجات الممكن استنباطها من تلك التجارب .

الخلاصة

تحتوي البيانات الإحصائية على قيم تختلف بشكل ملحوظ عن معظم القيم الأخرى و يطلق عليها الشواذ (القيم المتطرفة) Outliers وهذا موضوع بحثنا .
إذا تمكنا من تحديد المصدر الذي نتجت عنه الشواذ فسوف يكون بإمكاننا من تفسيرها عندئذ . فمن الممكن أن تكون تلك القيم ناتجة من خطأ في القراءة أو التدوين أو إجراء العمليات الحسابية أو في الأدوات المستخدمة .
إذا ما تأكد لدينا أن خطأ ما قد وقع أثناء جمع أو تدوين البيانات فإن الحل ببساطة يتمثل في إهمال الشواذ ولكن في كثير من الأحيان تحتوي الشواذ على معلومات هامة و بالتالي فمن الواجب فحص جميع القيم بدقة قبل إهمالها أو إبقائها.
تدرس هذه الإطروحة مصادر الشواذ ،طبيعتها ، تأثيرها ، المشاكل المترتبة على وجودها ، طرق اكتشافها و احتوائها وكذالك تأثيرها في عمليات الإنحدار الخطي .
وبشكل عام فإن أهداف هذه الإطروحة تتمثل في الإجابة عن التساؤلات الرئيسة الآتية:

١- هل ثمة وجود لمثل هذه القيم ؟
٢- إذا كانت الإجابة نعم فهل يجب إهمالها أم الإبقاء عليها ؟
٣- إذا كان يجب الإبقاء عليها فكيف نتمكن من معالجتها ؟ و كيف نتمكن من تقليل تأثيرها علي نتائج البحث وعلي الاستنتاجات المبنية عليه إلى أقل قدر ممكن .

في الفصل الأول :
إلى عرض تاريخي للقيم الشواذ مع التعريفات الخاصة بها و تعدد مصادرها وطرق اكتشافها وتحديدها وكذلك توضيح لماذا تشكل القيم الشواذ مشكلة لأي دراسة .
في الفصل الثاني :
عرض لجميع التعاريف و الرموز المستخدمة في هذه الإطروحة .
في الفصل الثالث :
عرض لطرق تصنيف الشواذ مثل طريقة ال (Z-Scores)،الطريقة المعدلة (Modified Z- Scores) و طريقة ال مخطط السبلة (Box Plots)
وكذالك طرق التعرف عليها مثل (ESD- Single Outlier) ،
(ESD- Many Outlier) و طرق أخرى تم عرضها في الإطروحة .