Multivariate Regression Compared with Moving Average Smoothing

Emmanuel Asuming

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Applied Mathematics and Computer Science

Eastern Mediterranean University
August 2021
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science.

Prof. Dr. Nazim Mahmudov
Chair, Department of Mathematics

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science.

Asst. Prof. Dr. Yücel Tandoğdu
Supervisor

Examining Committee

1. Prof. Dr. Aghamirza Bashirov

2. Assoc. Prof. Dr. Tolgay Karanfiller

3. Asst. Prof. Dr. Yücel Tandoğdu

ABSTARCT

Regression analysis is a statistical method having application in all fields of scientific and technological studies. Theoretical concepts lead to the development of regression theory are examined in some detail to lay down the foundation for the application of the theory. In statistics regression is mainly used to establish the kind of relationship between dependent and independent variables, i.e. linear or any other type.

Moving average is a statistical method widely used for smoothing out raw data trajectories to obtain trends by filtering out the noise from the random fluctuations. The trend is an estimation of the functional behavior of the variable under study.

This thesis is first centered on the theoretical characteristics of linear regression in chapter 3, examining the abstract concepts behind the regression theory, and the least squares method for establishing the model to be fitted from available data. Chapter 4 is allocated for the moving average technique used as a smoother of the trajectory for a variable. That smooth trend can be generated for every variable.

The fitted regression model itself can be considered a smooth functional representation of the response variable in relation to the predictor/s. In Chapter 5 a case study of a data set is undertaken, where moving average technique was implemented for smoothing with 2 different orders, using $m = 3$ and $m = 6$ values for averaging of a real life data. It became evident that the smoother the data, the lower the error measures will be in a regression analysis. However, too much smoothing of a variable will runs the risk of obtaining close to a perfect regression fit, which will not be realistic.

Based on the results obtained in the case study, it was then recommended that where large data sets are used for regression study, some smoothing can be beneficial as it will result in reduced estimation errors.

Some software programs like Excel, Minitab, and S.P.S.S were all used to help in data processing to find the needed outputs.

Keywords: matrix algebra, regression analysis, estimation, predictors, response, regression coefficient, moving average, stretched interpolated moving average.

ÖZ

İstatistiksel bir metod olan regresyon analizi bilim ve teknolojinin her alanında kullanılabiliyor. Regresyon teorisinin geliştirilmesinde kullanılan kavramların uygulanabilmesi için gerekli altyapıyı oluşturmak açısından detaylı bir şekilde incelenmiştir. İstatistikte regresyon bağımlı ve bağımsız değişkenler arasındki ilişkiyi tayin etmede kullanılıyor.

Hareketli ortalama yöntemi bir değişkene ait ham verilerin grafiğindeki aşırı dalgalamaları azaltma veya düzgünleştirme amaçlı kullanılıyor. Bir bakıma şansa bağlı aşırı dalgalanmaları filtre ediyor. Elde edilen düzgünleştirilmiş grafik değişkenin fonksiyonel hareketinin bir tahminisi olarak da düşünülebilir.

Bu tezde lineer regresyonun teorik karakteristikleri Kısım 3 de ele alınmıştır. Regresyon teorisinin temelini oluşturan bazı soyut kavramlar, ve veriden oluşturulacak regresyon modelinin belirlenmesinde elzem olan en küçük kareler metodu incelenmiştir. Kısım 4 hareketli ortalamar metodunun incelenmesine ayrılmıştır. Her değişken için düzgünleştirilmiş grafiğin nasıl üretilebileceği anlatılmıştır.

Veriden elde edilen regresyon grafiği, bağımlı değişkenin bağımsız değişken/lere olan ilişkisinin fonksiyonel bir temsiliyetidir. Kısım 5de yapılan uygulamada hareketkli ortalama metodu ile $m = 3$ ve $m = 6$ değerleri kullanılarak gerçek hayattan alınmış verilerin düzgünleştirilmesi yapılmıştır. Ham ve düzgünleştirilmiş veriler kullanılarak yapılan regresyon analizlerinden de görülmüştür ki düzgünleştirme arttıkca, regresyonda ortaya çıkan hata payları azalmıştır. Ancak aşırı düzgünleştirmenin regresyon hatalarını sıfıra doğru indireeği düşünülürse, greçekci olmadığı ortadadır.

Uygulamadan elde edilen sonuçlara bakarak büyük verilerin elde olduğu durumlarda bir miktar düzgünleştirmenin hata payların azaltmak açısından faydalı olacağı ortadadır.

Bu çalışmada Excel, Minitab, ve SPSS gibi istatistik paket yazılımlardan fadalanılmıştır.

Anahtar kelimeler: matrisler cebiri, regresyon analizi, tahmin, öngörü, yanıt, regresyon katsayısı, hareketli ortalama, esnetilmiş enterpolasyonlu hareketli ortalama.

DEDICATION

I dedicate this thesis to God, my mother Janet Osei, my siblings Rebecca Mensah Mills Fourdjour , Priscilla Asuming and Queenzy Laura Asuming, my father Kofi Asuming, Also to my supervisor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

DF	Degree of Freedom
e	Residual or Error
MA	Moving Average
MS	Mean Square
MSE	Mean Square Error
MST	Mean Square Total
R	Sample Correlation Matrix
R	Sample Correlation
RMSD	Root Mean Square Deviation
RRMSD	Relative Root Mean Square Deviation
S	Sample Covariance Matrix
S^2	Sample Variance
SIMA	Stretched Interpolated Moving Average
SSCP	Sum of Square Cross Product
SSCT	Sum of Square Cross Total
SSE	Sum of Square Error
SSR	Sum of Square Regression
SST	Sum of Square Total
X	Data Matrix
$\boldsymbol{x}$	Data Vector
$\boldsymbol{X}^T$	The Transpose of Matrix **X**
$\boldsymbol{X}^{-1}$	The Inverse of Matrix **X**
$\bar{\boldsymbol{X}}$	Mean Data Matrix

$\bar{\boldsymbol{x}}$	Mean Data Vector
β_1	Regression Coefficient
$\boldsymbol{\mu}$	Population Mean
$\boldsymbol{\mu}$	Population Mean Vector
μ_l	l^{th} Population Mean
$\boldsymbol{\rho}$	Population Correlation Matrix
$\boldsymbol{\Sigma}$	Population Covariance Matrix
τ_L	l^{th} Population Mean (treatment effects)

Chapter 1

INTRODUCTION

Regression as an estimation and projection technique is a very important method, especially when there is a high dependence between one or more dependent variables (response variable/s) onto another one or more independent variables (predictor variable/s. On the other hand regression itself is also a kind of smoothing technique. To clarify this point thinking about the simple linear regression, the fitted line is in fact giving an average value for the dependent variable given a certain value of the independent variable. Hence the true data points scatter around a single line, meaning the scatter of points are smoothed to obtain a line. Position of the line is determined through the least squares method

In this thesis chapter 2 is allocate for a brief literature review in regression and moving average concepts. In chapter 3 starting with simple linear regression up to the multivariate multiple regression concepts together with some related theorems with proofs are explained in detail. At the end of each sub section following the theoretical explanations, a numeric example is included to highlight the application of the theory.

Chapter 4 is devoted to the moving average concept which is widely used for trend generation, its relation to time series, and a summary of various moving average techniques. For the purpose of this thesis the simplified version of stretched

interpolated moving average technique (SIMA) is briefly explained and applied in the analysis of data.

Chapter 5 is a case study on a multivariate data to show how to apply the theory covered in this thesis, and discuss its results by means of error measuring parameters such as MSE, RMSE, and RRMSE, giving appropriate comments, where necessary. Meaning of error measuring parameters are also briefly explained for clarity. Data used relates to forest fire of an area in northern Portugal. Out of 4 dependent and 5 independent variables in the original data, 2 dependent and 2 independent ones were considered with a data subset of 50 observations out of 517 records in the original data set. Obtained error measuring parameter results are tabulated for clarity. Use of smoothed data clearly stood out in terms of lower error measures.

Chapter 2

LITERATURE REVIEW

This chapter discusses some of the literature available on multivariate regression, moving average and models that have been beneficial during the course of this study. Various articles and publications were also examined with regard to the model being used and the general working title.

The idea of regression analysis is traced back to the nineteenth century when Galton decided to collect information on the height of individuals as well as the height of their parents and upon gathering the information he decided to draw a frequency table and using them to classify the individuals by their height and the average height of their parents. Based on his studies he concluded that tall parent are more likely to have tall children while short children are likely to have short parents (Galton, 1989).

As Galton work focused on biological meaning, later research undertaken by Kurl and Yule was aimed at detailing the theoretical and inferential statistics which brought about multiple regression model. (Karl Pearson, G. U. Yule,Blanchard, Norman; Lee,Alice, 1903)

Least square method was introduced by Adrien-Marie Legendre a French mathematician in the early 1800. An American mathematician Robert Adrain also proposed the similar theory in 1808 (Stigler, 1980). Gauss made a publication on the

theory of least square in the 1821 and he further went on to add Guass-Markov theorem to the least square method. (Guass, 1821).

York introduced a method of weighing (York, 1966) which was known to be the cubic in the sense that the cubic equation was used to determine the regression equation in a case of a slope of the regression. Ricker worked on the geometric mean regression, and on the confidence limits for the slope of the geometric mean regression, (Ricker, 1984).

Since the introduction of the regression analysis concept research has continued to date with countless contributors. A very recent study undertaken by (Tandogdu Y. & Esager M., 2018) on the sensitivity of the regression parameters explaining in detail the simple and multivariate regression techniques and principal component analysis and showing the validity of the concept via a real life data.

Moving average (MA) technique was initially used for trend generation mostly in time dependent variables. Its effect is smoothing the trajectory of a raw data resulting in a trend line depicting an estimate of the expected trend for the variable under consideration. Different approaches are taken in the generation of the moving average trend. Some approaches in use are simple, cumulative, weighted, and exponential MA techniques.

Cumulative MA can be used when the total average starting from some point up to another point is needed. When different weights are to be assigned for different observations the weighted MA is suitable for use. Methods of determining the weights may depend on the case under study, hence necessitating different techniques for their

computation. Some such techniques used are linearly or exponentially decreasing or increasing weights, depending on the need. (Durbin, 1959). Weighted MA technique is widely used in finance, economy, or in some medical applications.

On the other hand moving average can be considered to be a simple smoothing method. It is possible to develop advanced moving average techniques to cater for special situations. One such method is the moving average as a trend generator on a trajectory where a smooth trend is generated, such that a smooth value can correspond to every raw data value with the same time or space coordinate. Hence the name Stretched Interpolated Moving Average (SIMA).

Use of computers in statistical computations has significantly changed the nature of statistical science. Ability to process huge data sets on one hand eliminated the burden and limitations of manual data processing, on the other hand it enabled the validation and applicability of certain statistical theory (Yates, 1966). Joiner in 1972 introduced a light version of OMNITAB 80, a statistical analysis program by NIST which was conceived by Joseph Hilsenrath in years 1962-1964 as OMNITAB program (minitab, 2011). Statistical Package for the Social Sciences (SPSS) founded in 1975 and acquired by IBM in 2009. Another software package called Statistical Analysis System (SAS) was founded by two professors in 1976 from North Carolina State University Dr. James Goodnight and John Sall. SAS is one of the fore runner statistical packages that enables both the development of new ideas and theories under its modules, as well as providing advice in decision making based on customers data analysis.

Chapter 3

REGRESSION ANALYSIS

3.1 Regression Analysis

Regression is a statistical method that enables the prediction of a dependent variable *Y* based on one or more than one independent variables X_i, i=1,2,…,*k.* It has found application in almost every field of study from finance to education, medical studies, engineering, just to mention a few. Its foundation rests on probability theory expressed as the expected value of a random variable, conditioned on one or more random variables. This concept is developed in statistics, utilizing available data on each variable. There are many types of regression analysis, such as *simple linear regression, multiple linear regression, multivariate multiple regression* just to mention few. (Keenan Pituch & James P. Stevens, 2016)

Multivariate regression analysis enables the establishment of the relationships between a dependent variable and independent variables. In application it uses multivariate data sets to establish the linear relationship between the dependent variable *Y* and a set of *k* independent variables X_i. In multivariate regression analysis to implement the methodology, matrix algebra is utilized. Hence, a preview of some basic statistical parameters using matrix algebra is given under section 3.2.

3.2 Descriptive Multivariate Statistics using Matrix Algebra

In descriptive statistics the measure of central tendency of a certain variable such as sample mean ($\bar{x}$), median ($\tilde{x}$), mode ($\hat{x}$) with *n* data values can be determined.

Measures of co-variation, and correlation between two or more variables are also determined.

3.2.1 Sample Mean

Let vector **x** represent the *n* data values that belong to a variable. That is

$\mathbf{x}_{n\times 1} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ has *n* rows and 1 column.

The transpose of $\mathbf{x}_{n\times 1}$ is denoted as $\mathbf{x}'_{1\times n}$ or $\mathbf{x}^T_{1\times n}$ and $\mathbf{x}'_{1\times n} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$.

Then a multivariate data with *p* variables and *n* observations in each variable is represented by the data matrix $\mathbf{X}_{n\times p}$ as (Neil, 2002).

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{21} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & x_{n1} & \ldots & x_{np} \end{bmatrix}$$

Transpose of the data matrix **X** is denoted as $\mathbf{X}'_{p\times n}$ or $\mathbf{X}^T{}_{p\times n}$. Hence,

$$\boldsymbol{X}^T = \begin{bmatrix} x_{11} & x_{21} & \ldots & x_{n1} \\ x_{12} & x_{22} & \ldots & x_{n2} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{1p} & x_{np} & \ldots & x_{np} \end{bmatrix}$$

It is also worth remembering that in probability theory as the expected value or the first moment of the random variable *X* around the origin is given as

$$E(X) = \mu = \begin{cases} \int_{\Re} xf(x)dx & X \text{ continuous} \\ \sum_{\forall x} xf(x) & X \text{ discrete} \end{cases}$$

Then the sample mean of a data set of *n* observations obtained from the domain of a single random variable *X* can be computed as the simple arithmetic mean of the data given by

$$\bar{x} = \frac{\sum_{i=}^{n} x_i}{n}$$ (KAREN A. RANDOLPH & LAURA L. MYERS, 2013)

In the case of multivariate data with *p* variables we have a vector of sample averages denoted as $\bar{\mathbf{x}}$ where $\bar{\mathbf{x}}' = \left[\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p\right]$. Using matrix algebra on the multivariate data $\bar{\mathbf{x}}$ is computed as

$$\bar{\mathbf{x}} = \frac{1}{n}\left(\mathbf{X^T} \times \mathbf{1}\right)$$

$$\bar{x} = \frac{1}{n}\begin{bmatrix} x_{11} & x_{21} & \ldots & x_{i1} & \ldots & x_{n1} \\ x_{12} & x_{22} & \ldots & x_{i2} & \ldots & x_{n2} \\ . & . & \ldots & . & \ldots & . \\ . & . & \ldots & . & \ldots & . \\ . & . & \ldots & . & \ldots & . \\ x_{1i} & x_{2i} & \ldots & x_{ij} & \ldots & x_{nj} \\ . & . & \ldots & . & \ldots & . \\ . & . & \ldots & . & \ldots & . \\ . & . & \ldots & . & \ldots & .. \\ x_{1p} & x_{2p} & \ldots & x_{ip} & \ldots & x_{np} \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \\ . \\ . \\ . \\ 1 \end{bmatrix} = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i1} \\ . \\ . \\ . \\ \sum_{i=1}^{n} x_{ij} \\ . \\ . \\ . \\ \sum_{i=1}^{n} x_{ip} \end{bmatrix}$$

Where the k^{th} ($k=1,2,\ldots, p$) variable's mean is

$$\bar{x}_k = \frac{\sum_{i=1}^{n} x_{ik}}{n}, \quad k = 1, 2, \ldots, p$$

3.2.2 Sample Variance and Covariance

Covariance is used to measure the joint variability of two variables. It determines the direction of the relationship between the two variables whether variables turn to move

together (positive covariance) or inversely (negative covariance). When their covariance is 0 indicates no relationship among the variables.

When a single variable is in question, the variation and deviation of values around the sample mean is of prime concern. As the expected value of difference between the population mean and the values of the random variable is always zero $E(X-\mu)=0$, then the square of these differences is defined as the population variance

$$\text{var}(X)=E(X-\mu)^2=E\big((X-\mu)(X-\mu)\big)=\sigma^2$$

The square root of the variance gives a measure of deviation from the mean value μ, that is the population standard deviation $\sigma=\sqrt{\sigma^2}$. This is a very useful measure and its equivalent in statistics is defined as

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}.$$

When dealing with multivariate case where *p* variables with *n* data values in each variable are given, in addition to the standard deviation of each variable, the co-variation between the variables is of prime concern. In this respect, the covariance and correlation between the variables is defined in probability as

$$\text{Cov}(X_j X_k)=\sigma_{\boldsymbol{X}_j \boldsymbol{X}_k}=E\big((X_j-\boldsymbol{\mu}_j)(X_k-\mu_k)\big)=\sigma_{jk} \qquad k, j=1,2,\ldots,\text{p}$$

and

$$\text{Cor}(X_j, X_k)=\frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}}=\rho \text{ Respectively.}$$

Then the covariance matrix $\boldsymbol{\Sigma}$ for *p* random variables $X_1, X_2, \ldots, X_p$ can be expressed as a $p\times p$ matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1j} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2j} & \ldots & \sigma_{2p} \\ . & . & \ldots & . & \ldots & . \\ . & . & \ldots & . & \ldots & . \\ . & . & \ldots & . & \ldots & . \\ \sigma_{i1} & \sigma_{i2} & \ldots & \sigma_{ij} & \ldots & \sigma_{ip} \\ . & . & \ldots & . & \ldots & . \\ . & . & \ldots & . & \ldots & . \\ . & . & \ldots & . & \ldots & . \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pj} & \ldots & \sigma_{pp} \end{bmatrix}$$

The diagonal elements of the covariance matrix represents the variance of each variable $\sigma_{ij} = \sigma_k^2;\ i = j$ and $i, j, k = 1, 2, \ldots, p$. The off diagonal elements are the covariance values σ_{X_i, X_j}; i,j=1, 2,… p, and $i \neq j$ between the variables X_i, X_j .

For a given $n \times p$ multivariate data matrix the covariance is defined as

$$\boldsymbol{s_j^2} = \boldsymbol{s_{jj}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}_j\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}_j\right)\left(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}_j\right) = \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{x}_{ij}^{*}\boldsymbol{x}_{ij}^{*}$$
$$i, j = 1, 2, \ldots, p$$

Where $\left(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}_j\right) = \boldsymbol{x_{ij}^{*}}$

Therefore

$$\boldsymbol{s}_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}_j\right)\left(\boldsymbol{x_{ik}} - \bar{\boldsymbol{x}}_k\right) = \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{x}_{ij}^{*}\boldsymbol{x}_{ik}^{*}$$
$$\mathrm{k, j} = 1, 2, \ldots, \mathrm{p} \ \text{ where } \mathrm{k} \neq \mathrm{j}$$

Where $\left(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}_j\right) = \boldsymbol{x_{ij}^{*}}$ and $\left(\boldsymbol{x_{ik}} - \bar{\boldsymbol{x}}_k\right) = \boldsymbol{x_{ik}^{*}}$, the matrix multiplication of both equations becomes

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x_{ij}} - \bar{\mathbf{x}}_j\right)^T\left(\mathbf{x_{ij}} - \bar{\mathbf{x}}_j\right) = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{x}_{ij}^{*T}\mathbf{x}_{ij}^{*} \qquad i, j = 1, 2, \ldots, p$$

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x_{ij}} - \bar{\mathbf{x}}_j\right)^T\left(\mathbf{x_{ik}} - \bar{\mathbf{x}}_k\right) = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{x}_{ij}^{*T}\mathbf{x}_{ij}^{*} \quad k, j = 1, 2, \ldots, p \ \text{where} \ k \neq j$$

Expressing these as matrix multiplication and dividing by *n-1*

$$\begin{bmatrix} x_{11}^* & x_{12}^* & \dots & x_{1j}^* & \dots & x_{1p}^* \\ x_{21}^* & x_{22}^* & \dots & x_{2j}^* & \dots & x_{2p}^* \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & . \\ x_{i1}^* & x_{i2}^* & \dots & x_{ij}^* & \dots & x_{ip}^* \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & .. \\ x_{n1}^* & x_{n2}^* & \dots & x_{nj}^* & \dots & x_{np}^* \end{bmatrix} \times \begin{bmatrix} x_{11}^* & x_{21}^* & \dots & x_{i1}^* & \dots & x_{n1}^* \\ x_{12}^* & x_{22}^* & \dots & x_{i2}^* & \dots & x_{n2}^* \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & . \\ x_{1i}^* & x_{2i}^* & \dots & x_{ij}^* & \dots & x_{nj}^* \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & .. \\ x_{1p}^* & x_{2p}^* & \dots & x_{ip}^* & \dots & x_{np}^* \end{bmatrix} =$$

the covariance matrix of the data matrix $\mathbf{X}_{n\times p}$ can be written as

$$\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1j} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2j} & \dots & s_{2p} \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & . \\ s_{i1} & s_{i2} & \dots & s_{ij} & \dots & s_{ip} \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & . \\ s_{p1} & s_{p2} & \dots & s_{pj} & \dots & s_{pp} \end{bmatrix}$$

Elements of the covariance matrix in the *j*, *k* position are known as the covariance between the jth and kth variables.

From practical point of view, standard deviation *s* is a measure of deviation from the sample mean $\bar{x}$. The greater the *s* value the greater the deviation of data values from $\bar{x}$.

Covariance between any two variables X_i and X_j, $-\infty < s_{ij} < \infty$ is an indication to the relation between the two variables.

$s_{ij} < 0$: As one variable increase the other tend to decrease

$s_{ij} > 0$: As one variable increase the other also tend to increase

$s_{ij} = 0$: Is a critical value that may indicate the independence of the two variables. In fact if X_i and X_j are independent, then $s_{ij} = 0$, but the vice versa case is not always true. However, the magnitude of s_{ij} is not an indication of the strength of the relation between the variables X_i and X_j. Therefore, another measure that can indicate the strength of the relation between the variables X_i and X_j is needed. That is the correlation coefficient.

3.2.3 Linear Correlation Coefficient

Correlation is a measure that gives the strength and direction of the relation between two variables X_i and X_j. Correlation coefficient for the random variables X_i and X_j, is denoted as ρ and $-1 \le \rho \le 1$. The following can be said about ρ depending on the values it takes.

$\rho > 0$: Positive correlation means that the variables X_i and X_j increases at the same time.

$\rho = 1$: Perfect positive correlation. Any increase in one variable corresponds to the same amount of increase in the other.

$\rho < 0$: Negative correlation. It means as one variable increase the other decrease.

$\rho = -1$: Perfect negative correlation. Any increase in one variable corresponds to the same amount of decrease in the other.

$\rho = 0$: No correlation between the two variables, or variables X_i and X_j are independent.

Linear correlation coefficient between two populations is defined by

$$\mathrm{Cor}\left(X_j, X_k\right) = \rho = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}}; \quad -1 \le \rho \le 1$$

Here

σ_{ik} : Covariance between the variables X_i and X_k

σ_{ii} : Variance of the variable X_i

σ_{kk} : Variance of the variable X_k

Consequently the correlation coefficient of a variable by itself can be expressed as

$$\mathrm{Cor}\left(X_j\right) = \frac{\sigma_{jj}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{jj}}} = 1$$

Expressing the pairwise linear correlation between the *p* variables of a process can be written as in the matrix below.

$$\rho = \begin{bmatrix} 1 & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{21}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & 1 & \cdots & \frac{\sigma_{2j}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{22}}} \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \frac{\sigma_{i1}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{ii}}} & \frac{\sigma_{i2}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{22}}} & \cdots & 1 & \cdots & \frac{\sigma_{ip}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{pp}}} \\ \cdot & \cdot & \cdots & \cdot & \cdots & \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ \frac{\sigma_{p1}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \frac{\sigma_{p2}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{pj}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{jj}}} & \cdots & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1j} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2j} & \cdots & \rho_{2p} \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ \rho_{i1} & \rho_{i2} & \cdots & \rho_{ij} & \cdots & \rho_{ip} \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pj} & \cdots & \rho_{pp} \end{bmatrix}$$

When a random sample of *p* variables and *n* observations are given, the pairwise linear correlation coefficients between the variables X_i and X_k ($i, k = 1,2, \ldots, n$) can be computed as

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$$

where

s_{ik} : Covariance between the variables X_i and X_k

s_{ii} : Variance of the variable X_i

s_{kk} : Variance of the variable X_k

Note that $-1 \le r \le 1.$

$$r_{ii} = \frac{s_{ii}}{\sqrt{s_{ii}}\sqrt{s_{ii}}} = 1 \qquad \text{i} = 1, 2, \ldots, \text{p}$$

Then the linear correlation matrix for *p* variables becomes

$$R = \begin{bmatrix} 1 & \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \cdots & \frac{s_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} & \cdots & \frac{s_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{s_{21}}{\sqrt{s_{11}}\sqrt{s_{22}}} & 1 & \cdots & \frac{s_{2j}}{\sqrt{s_{jj}}\sqrt{s_{22}}} & \cdots & \frac{s_{2p}}{\sqrt{s_{pp}}\sqrt{s_{22}}} \\ . & . & \cdots & . & \cdots & . \\ . & . & \cdots & . & \cdots & . \\ . & . & & . & & . \\ \frac{s_{i1}}{\sqrt{s_{11}}\sqrt{s_{ii}}} & \frac{s_{i2}}{\sqrt{s_{ii}}\sqrt{s_{22}}} & \cdots & 1 & \cdots & \frac{s_{ip}}{\sqrt{s_{ii}}\sqrt{s_{pp}}} \\ . & . & \cdots & . & \cdots & . \\ . & . & \cdots & . & \cdots & . \\ . & . & \cdots & . & \cdots & . \\ \frac{s_{p1}}{\sqrt{s_{11}}\sqrt{s_{pp}}} & \frac{s_{p2}}{\sqrt{s_{pp}}\sqrt{s_{22}}} & \cdots & \frac{s_{pj}}{\sqrt{s_{pp}}\sqrt{s_{jj}}} & \cdots & 1 \end{bmatrix}$$

$$= \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1j} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2j} & \cdots & r_{2p} \\ . & . & \cdots & . & \cdots & . \\ . & . & \cdots & . & \cdots & . \\ . & . & \cdots & . & \cdots & . \\ r_{i1} & r_{i2} & \cdots & r_{ij} & \cdots & r_{ip} \\ . & . & \cdots & . & \cdots & . \\ . & . & \cdots & . & \cdots & . \\ . & . & \cdots & . & \cdots & . \\ r_{p1} & r_{p2} & \cdots & r_{pj} & \cdots & r_{pp} \end{bmatrix}$$

Example 1*:* A data containing 3 observations associated with 2 variables is been considered for computation and explanation:

$\mathbf{x}_1$ represent the number of bikes rented

$\mathbf{x}_2$ represent the temperature under which the bikes were rented

$$\boldsymbol{X} = \begin{bmatrix} 4 & 4 \\ 3 & 5 \\ 5 & 9 \end{bmatrix}$$

$\underline{x_1}$	$\underline{x_2}$
4	4
3	5
5	9

and the mean of the vectors was calculated using the sample mean formula by

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

$$\bar{x}_1 = \frac{4+3+5}{3} = 4 \qquad \bar{x}_2 = \frac{4+5+9}{3} = 6$$

The matrix deviations represented by $\mathbf{X}_d = \mathbf{x}_i - \bar{x}_i;\ i = 1,2$

$$X_d = \underset{X}{\begin{bmatrix} 4 & 4 \\ 3 & 5 \\ 5 & 9 \end{bmatrix}} - \underset{\bar{X}}{\begin{bmatrix} 4 & 6 \\ 4 & 6 \\ 4 & 6 \end{bmatrix}} = \begin{bmatrix} 0 & -2 \\ -1 & -1 \\ 1 & 3 \end{bmatrix}$$

Taking the sum of squares and cross products (SSCP) as is the product of $\boldsymbol{X}_d^T \cdot \boldsymbol{X}_d$

where $\boldsymbol{X}_d^{'}$ is just the transpose of $\boldsymbol{X}_d$

$$\boldsymbol{X}_d^T \cdot \boldsymbol{X}_d = \begin{bmatrix} 0 & -1 & 1 \\ -2 & -1 & 3 \end{bmatrix} \begin{bmatrix} 0 & -2 \\ -1 & -1 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 4 & 14 \end{bmatrix}$$

Obtained deviation sums of squares forms the numerator in the formula when computing the variances for each variable.

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

Covariance between the variables x_1 and x_2 is

$$\boldsymbol{S}_{12} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{n-1}$$

The deviation value for the i[th] observation for x_1 is $(x_{i1} - \bar{x}_1)$ *and* $(x_{i2} - \bar{x}_2)$ is the deviation value for i[th] observation for x_2.

The matrix of variances and covariances **S** is obtained as

$$\boldsymbol{S} = \frac{SSCP}{n-1}$$

$$\boldsymbol{S} = \frac{1}{2}\begin{bmatrix} 2 & 4 \\ 4 & 14 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 7 \end{bmatrix}$$

We therefore say that 1 and 7 are the variance and 2 is the covariance for the variables x_1 and x_2.

The linear correlation coefficient between the two variables is given by

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$$

$$r_{12} = \frac{4}{\sqrt{2}\sqrt{14}} = r_{21} = 0.76$$

$$R = \begin{bmatrix} 1 & 0.76 \\ 0.76 & 1 \end{bmatrix}$$

3.2.4 Estimating $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$

Sample mean vector $\bar{\mathbf{x}}$ obtained from a random sample represented by the $n \times p$ matrix **X**, is an unbiased and consistent estimator of $\boldsymbol{\mu}$ which is the mean vector of the p variables from where the sample is taken. Then the estimator of $\boldsymbol{\mu}$ can be written as $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$.

Estimating the population covariance matrix $\boldsymbol{\Sigma}$ for a multivariate distribution, the covariance matrix obtained from *n*×*p* data set. Theorem 1 given below has its proof in the reference (Tandogdu Y. & Esager M., 2018)

Theorem 1: A random sample $X_1, X_2, \ldots, X_n$ with joint distribution having mean vector $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ as its covariance matrix, will have a sample covariance matrix $\frac{n}{n-1}\mathbf{S}$ which is an unbiased estimator of $\boldsymbol{\Sigma}$.

3.3 The General Probabilistic Regression Model

In probability theory the regression of a random variable *Y*, on to another random variable *X* necessitates the two r.v.s have a joint probability distribution *f*(*x*,*y*). Then the regression of *Y* on *X* is defined as

$$E\left[f(Y|X=x)\right] = E\left[\frac{f(x,y)}{f(x)} \right] = \mu_{Y|x} \qquad 3.3.1$$

This is the general expression for the regression of *Y* on *X*. Similarly the regression of *X* on *Y* can be written as

$$E\left[f(X|Y=y)\right] = E\left[\frac{f(x,y)}{f(\mathrm{y})} \right] = \mu_{X|y}. \qquad 3.3.2$$

The concept can be extended to *k* random variables $X_1, X_2, \ldots, X_k$, with multivariate joint probability distribution $f(x_1, x_2, \ldots x_k)$, leading to multivariate regression expressions. To highlight this concept the following example will be used

Example 1: The joint probability density function *f(x,y)* is given below.

$$f(x,y) = \begin{cases} x + \frac{3}{2}y^2; & 0 \le \mathrm{x} \le 1, 0 \le \mathrm{y} \le 1 \\ 0 \quad ; & \text{elsewhere} \end{cases}$$

Regression equation of *Y* on *X*, and *X* on *Y*, can be determined following the methodology explained.

If *X* and *Y* can be shown to be independent by satisfying the *f(x,y) = f(x)f(y)* condition, then it is not possible to find the kind of regression relation between *X* and *Y*.

For checking the independence of the r.v.s we need the marginal densities. (Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers,Keying Ye, 2011)

$$f(x) = \int_{\Re} f(x,y)dy = \int_0^1 (x + \frac{3}{2}y^2)dy = xy + \frac{1}{2}y^2\Big|_0^1 = x + \frac{1}{2};\quad 0 \le x \le 1$$

$$f(y) = \int_{\Re} f(x,y)dx = \int_0^1 (x + \frac{3}{2}y^2)dx = \frac{x^2}{2} + \frac{3}{2}xy^2\Big|_0^1 = \frac{1}{2} + \frac{3}{2}y^2;\quad 0 \le y \le 1$$

Then applying the independence condition we have

$f(x,y) = f(x)f(y) \to x + \frac{3}{2}y^2 \ne \left(x + \frac{1}{2}\right)\left(\frac{1}{2} + \frac{3}{2}y^2\right)$. Hence *X* and *Y* not independent.

Regression of *X* on *Y*

$$E\left[f(X\,|Y = y)\right] = E\left[\frac{f(x,y)}{f(y)}\right] = \mu_{X|y} \to E\left(\frac{x + \frac{3}{2}y^2}{\frac{1}{2} + \frac{3}{2}y^2}\right) = E\left(\frac{2x + 3y^2}{1 + 3y^2}\right)$$

$$= \int_0^1 x\frac{2x + 3y^2}{1 + 3y^2}dx = \frac{4 + 9y^2}{6(1 + 3y^2)};\ 0 \le y \le 1$$

Regression of *Y* on *X*

$$E\left[f(Y\,|X = x)\right] = E\left[\frac{f(x,y)}{f(x)}\right] = \mu_{Y|x} \to E\left(\frac{x + \frac{3}{2}y^2}{x + \frac{1}{2}}\right) = E\left(\frac{2x + 3y^2}{2x + 1}\right)$$

$$= \int_0^1 y\frac{2x + 3y^2}{2x + 1}dx = \frac{3 + 4x}{4(1 + 2x)};\ 0 \le x \le 1$$

3.4 Statistical Linear Regression Model

Bivariate probabilistic regression model is based on the abstract definition given in equations 3.5.1 and 3.5.2. However, in application a bivariate or multivariate joint probability density function is not available. Hence, using available data from the process under study and adhering to the rules of probability, the statistical regression model is developed.

3.4.1 The Simple Linear Regression Model

This is the model that can be expressed with a linear equation $Y = \beta_0 + \beta_1 x + \varepsilon$. Here *y* is the dependent variable and *x* the independent variable, β_0 and β_1 are the *y* intercept and the slope of the regression line respectively, ε represents the random error. The random error is assumed to have mean zero and variance σ^2, i.e. $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$. Without the random error, the model would become a deterministic one, not allowing the use of probabilistic approach, and estimation of the dependent variable. While $Y = \beta_0 + \beta_1 x + \varepsilon$ is the ideal and unknown model, it can be estimated by using available data for the variables *X* and *Y*.

On the other hand while the response and predictor variables are continuous, in application the data is discrete in nature and is used to estimate the model regression equation by determining the parameters of the fitted model. Hence, the theory used in determining the parameters of the fitted model will be based on the discrete data analysis concepts.

We know $Y = \beta_0 + \beta_1 x + \varepsilon$ theoretical linear regression model. It has an estimator

$y = b_0 + b_1 x_i + e_i,$

$b_0,\ b_1$: Regression coefficients.

e_i: Residual or error.

However, for every different data set used from the same population a different fitted model $\hat{y} = \hat{b}_0 + \hat{b}_1 x$ will be obtained. Here the hat used on $\hat{y},\ \hat{b}_0,\ \hat{b}_1$ indicates that they are estimators of corresponding $y,\ b_0,\ b_1$ parameters. Since the error $e_i = y_i - \hat{y}_i, i = 1, 2, \ldots, n$ always has a mean zero does not appear in the fitted model.

The fitted model must be such that $\sum_{i=1}^{n} (x_i - \hat{x}_i)^2$ must be minimum. That is:

$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{b}_0 + \hat{b}_1 x))^2$ must be minimum. This is achieved by the least squares method. For the simple linear regression the method works as follows

$$y_i = b_0 + b_1 x_i + e_i$$
$$e_i = y_i - \hat{y}_i = y_i - \hat{b}_0 - \hat{b}_1 x_1$$

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$$
$$\frac{\partial(\text{SSE})}{\partial \hat{b}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{b}_0 - \hat{b}_1 x_i) = 0$$
$$\frac{\partial(\text{SSE})}{\partial \hat{b}_1} = -2 \sum_{i=1}^{n} \mathrm{x}_i (y_i - \hat{b}_0 - \hat{b}_1 x_i) = 0$$

Rearranging yields the normal equations

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

From the simultaneous solution of the normal equations the fitted linear regression equation parameters are obtained as

$$\hat{b}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{b}_0 = \frac{\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{b}_1\bar{x}$$

Pit hole to be avoided. When substituting *x* values in the regression equation, such values should not be far below the minimum or far above the maximum values in the data set. Such values will result in unreliable prediction of the dependent variable *Y*.

Definition 1: Determination of the smallest *x* value below the $x_{\min}$ and largest *x* value above the $x_{\max}$ can be done following the steps below:

i. Standardize the predictor data.

ii. Find the z_{min} and $z_{\max}$ value corresponding to the $x_{\min}$ and $x_{\max}$ respectively.

iii. Determine the z_{min} - 1 and $z_{\max}$+1 values

iv. Compute the limit *x* values corresponding to the *z* values found in step iii by assuming $\bar{x}$ and s as point estimators of the population parameters μ and σ respectively. Call these x_L and x_U.

Then the *x* values to be substituted in the regression equation should be taken from the (x_L and x_U) interval.

Example: The data given in Table 3.1 represents the age and Urea levels in people

Table 3. 1: Age of people and Ln Urea levels for 15 persons

Age (*X*)	Urea (*Y*)	StdScr *X*
39	1.526	-1.72282
44	1.686	-1.40996
45	1.548	-1.34739
55	1.131	-0.72167
58	1.988	-0.53395
60	1.099	-0.40881
67	1.386	0.0292
71	2.002	0.279489
72	2.617	0.342061
74	1.917	0.467206
76	1.723	0.59235
76	2.054	0.59235
81	2.054	0.905211
89	2.262	1.405789
91	2.701	1.530933

The fitted linear regression equation for the data from Table 3.1, where the response variable is Urea level (*Y*) and the predictor (*X*) is the age, is obtained as *y = 0.5428 + 0.0196x*.

Based on this data the Urea level of a person can be predicted for the ages between the min 39 and max 91 with no hesitation. However, any age value below 39 and above 91 has to have its limits according to the logic given Definition 1. The scatter plot of the z scores for the *X* values given in Figure 3.1 is clearly indicating a good symmetry in the distribution of the data. Therefore, according to definition 1, the lowest x_L and highest x_L values that can safely be substituted in the fitted model are computed as:

For x_L : -1.72282 - 1 = -2.72282.

Then $z = \frac{x-\mu}{\sigma} \rightarrow -2.723 = \frac{x-66.53}{15.98} \rightarrow x = 23.02 = x_L$

For x_L : 1.53 + 1 = 2.53Then $z = \frac{x-\mu}{\sigma} \rightarrow 2.53 = \frac{x-66.53}{15.98} \rightarrow x = 106.95 = x_U$

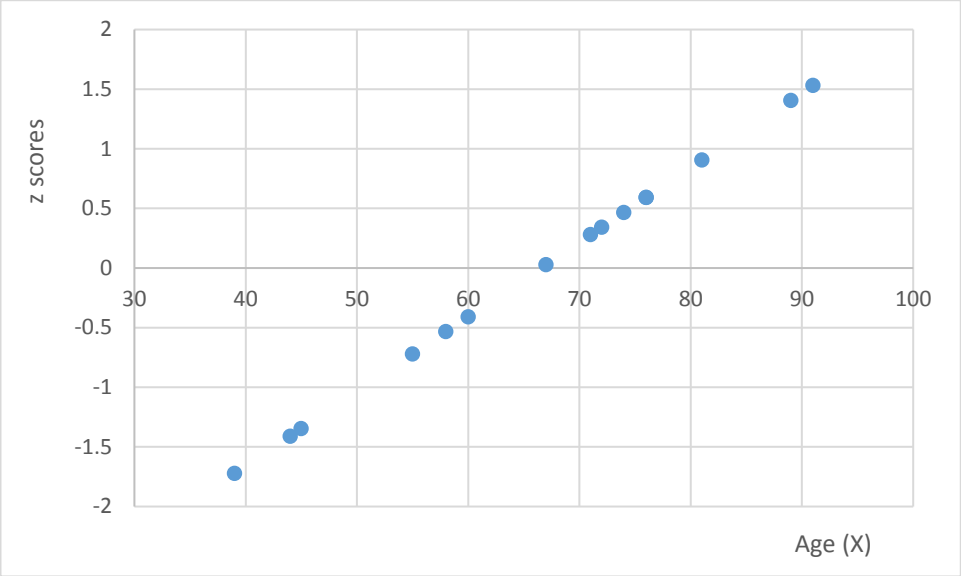

Figure 3. 1: The z scores of the predictor X clearly indicating the existence of symmetry in the data

3.4.1.1 Mean and Variance of Least Squares Estimators

The unknown regression model parameters β_0 and β_1 are estimated based on some randomly collected data. The computed regression parameters b_0 and b_1 based on the data are just realizations of the random variables B_0 and B_1. Then B_0 and B_1 are unbiased estimators of β_0 and β_1 respectively. To find the mean and variance for B_1 we proceed as follows.

The intercept parameter is computed as

$$B_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad .$$

Let

$$C_i = \frac{\sum_{i=1}^{n}(x_i-\bar{x})}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad i=1,2,3,\ldots,n$$

then

$$B_1 = \sum_{i=1}^{n} C_i Y_i$$

It is also a fact that independent random variable $X_1, X_2, ..., X_n$ have a normal distribution with mean $\mu_1, \mu_2, \mu_3, \ldots, \mu_n$ and variance $\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_n$ respectively. Then the random variable

$$Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

Will follow a normal distribution with

$$\mu_Y = a_1\mu_1 + a_2\mu_2 + a_3\mu_3 + \ldots + a_n\mu_n$$

and

$$\sigma_Y^2 = a_1\sigma_1^2 + a_2\sigma_2^2 + a_3\sigma_3^2 + \ldots + a_n\sigma_n^2 \ .$$

It follows that

$$\mu_{B_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 - \beta_1 x_i)}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1$$

and

$$\mu_{B_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma_y^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

It can also be shown that $\mu_{\beta_0} = \beta_0$ and variance of B_0 is

$$\sigma_{B_0}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2$$

3.4.1.2 Analysis of Variance in Simple Regression ANOVA

Every sum of square is divided by the required degree of freedom to obtain the mean of the sum of squares stemming from regression and mean sum of squares stemming from the error. The summarizing of the decomposition associated with *y* in terms of

an ANOVA the variation are converted into variance by dividing by its degree of freedom which helps in finding the goodness of fit associated to the regression line

In ANOVA the null hypothesis is set in a way that the regression is not significant and the alternative is written to cover the fact that the regression is significant at a certain significance level, for example 5%. Rejecting the null hypothesis will lead to concluding that the regression coefficient is significant on the basis that the calculated value of the statistic is found or fall in the critical region. Accepting the null hypothesis will lead to the conclusion that the regression coefficient is not significant on the basics that calculated value of the statistic which falls outside the critical region. F distribution is used for the hypothesis test. Based on the given significance level and the degrees of freedom the critical *f* value is read from the *F* table. The computed *F* value also called the test statistic is used in determining whether the null hypothesis is acceptable or not. The test statistic or *F* statistic is computed as $\mathrm{F} = \frac{MSR}{MSE}$.

The null hypothesis for testing any of the regression parameters β_i to decide whether it is zero or not,

$$\mathrm{H}_0 : \beta_i = 0$$

Alternative as

$$\mathrm{H}_1 : \beta_i \neq 0$$

Table 3. 2: ANOVA Associated with Simple Regression

Origin of variation	(SS)	(D.F)	Mean square	F-statistic
REGRESSION	(SSR)	1	MSR= $\frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
ERROR	(SSE)	n-2	MSE= $\frac{SSE}{n-2}$	
TOTAL	(SST)	n-1		

Explained variation divided by the total variation is the coefficient of determination and is given by

$$r^2 = \frac{SSR}{SST} \qquad 0 \leq r^2 \leq 1$$

The greater or bigger coefficient of determination r^2, the more accurate the fitted model will be.

For a numeric example see Appendix A, Example 1.

3.4.2 Multiple Linear Regression

The regression process in the presence of more than one independent variable X_1, X_2 ,…,X_p upon which a single response variable Y depends, is named as the multiple linear regression analysis. This model can be expressed as $\underset{n\times 1}{\mathbf{y}} = \underset{n\times(p+1)}{\mathbf{X}} + \underset{(p+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$. Here

$\boldsymbol{\beta}$: Coefficients of the regression vector, $\mathbf{y}$: Observations of dependent variable, $\boldsymbol{\varepsilon}$: Vector residuals.

Writing the elements of the system openly we have

$$\mathbf{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X}=\begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix}, \boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Then an individual element of the vector **y** will be

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

The true but unknown coefficients $\boldsymbol{\beta}$ are estimated using the coefficients $b_0, b_1, \ldots, b_k$ that are computed from the data collected from a process of interest. Error sum of squares must be minimum. The sum of square of the errors is

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 ,$$

Expressing in matrix format SSE and carrying out necessary algebraic manipulation (Tandogdu Y. & Esager M., 2018) $\left(\mathbf{X}'\mathbf{X}\right)\mathbf{b} = \mathbf{X}'\mathbf{y}$ is obtained. From here the parameters b_i can be found as $\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$.

The estimators $\left(b_0, b_1, \ldots, b_k\right)$ of $\left(\beta_0, \beta_1, \ldots, \beta_k\right)$ are obtained by assuming that the random errors $\left(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_k\right)$, are independent and all having the same distributed with mean $E(\varepsilon_i) = 0$, variance $Var(\varepsilon_i) = \sigma^2$. Hence, it can be shown that the parameters $\left(b_0, b_1, \ldots, b_k\right)$ are unbiased estimators to $\left(\beta_0, \beta_1, \ldots, \beta_k\right)$ respectively.

The variance-covariance matrix of the estimators $\left(b_0, b_1, \ldots, b_k\right)$ is $\mathbf{A}^{-1}\sigma^2$. Variance of the estimators are on the main diagonal and the off-diagonal elements are the covariances. When there are *p* independent variables (predictors), inverse of $\mathbf{A}$ can in general be written as

$$\mathbf{A}^{-1} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} c_{00} & c_{01} & \dots & c_{0p} \\ c_{10} & c_{11} & \cdots & c_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p0} & c_{p1} & \dots & c_{pp} \end{bmatrix}$$

Then we have

$\sigma_{b_i}^2 = c_{ii}\sigma^2, \;\; i = 1, 2, \ldots, p$

$\sigma_{b_i b_j} = Cov(b_i, b_j) = c_{ij}\sigma^2, \;\; i \neq j$

To emphasize the importance of certain multiple regression related concepts some theorems are given.

Theorem 4.1. In a multivariate linear regression problem subtracting the estimated values from the true ones of a response variable is expressed as a function $S(\mathbf{b})$. Then

$S(\mathbf{b}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ or $S(\mathbf{b}) = (\mathbf{y} - \mathbf{Xb})' (\mathbf{y} - \mathbf{Xb})$.

It can be shown that the estimate $\mathbf{b}$ or $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtainable when $\mathbf{X}$ is full rank $k+1 \leq n$ with $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

The following properties of the estimators in the classical least squares can be written.

1. $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ has $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. (Tandogdu Y. & Esager M., 2018)
2. Expectation of error vector $\hat{\boldsymbol{\varepsilon}}$ is zero and covariance $\text{Cov}(\hat{\boldsymbol{\varepsilon}}) = \boldsymbol{\sigma}^2(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$.
3. Similarly $E(\hat{\boldsymbol{\varepsilon}}'\boldsymbol{\varepsilon}) = (n-k-1)\boldsymbol{\sigma}^2$.

Theorem 4.2: If $\mathbf{y} = \mathbf{X} + \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $E(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}, \text{cov}(\hat{\boldsymbol{\varepsilon}}) = \boldsymbol{\sigma}^2\mathbf{I}$. and $\mathbf{X}$ has full rank *k+1*, then for any vector of constants $\mathbf{h}$ the estimator $\mathbf{h}'\hat{\boldsymbol{\beta}} = h_0\hat{\beta}_0 + \cdots + h_k\hat{\beta}_k$ of $\mathbf{h}'\boldsymbol{\beta}$ gives the

minimum variance as $\mathbf{h}'\mathbf{Y} = h_1 Y_1 + \cdots + h_n Y_n$. For proof see (Tandogdu Y. & Esager M., 2018)

Regression parameters $\hat{\boldsymbol{\beta}}$ and error sum of squares $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ have sampling distributions which are indisputably important for assessing the influence of independent variables in regression analysis. Theorems 4.3 and 4.4 are given without proof to further clarify this concept.

Theorem 4.3. Let the multiple linear regression $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with full rank $p+1$ be given. Error vector $\boldsymbol{\varepsilon}$ has normal distribution (mean vector $\mathbf{0}$ and variance vector $\sigma^2\mathbf{I}$). $\boldsymbol{\beta}$'s maximum likelihood estimator corresponds to the least squares estimator of $\hat{\boldsymbol{\beta}}$. It can also be shown that $\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{X}'\mathbf{Y}$ is normally distributed with $\boldsymbol{\beta}$ being its mean and e $\sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}$ its variance. There is no dependence between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}$. Additionally $n\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}'\boldsymbol{\varepsilon}'$ follows the $\sigma^2\chi^2_{n-r-1}$ distribution. It must be remembered that the maximum likelihood estimator of σ^2 is $\hat{\sigma}^2$.

Theorem 4.4 Let $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *r+1* being the full rank, and $\boldsymbol{\varepsilon}$ normally distributed $N(\mathbf{0}, \sigma^2\mathbf{I})$ Then for $\boldsymbol{\beta}$ the $1-\alpha$ confidence area can be determined as

$\mathbf{U} = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{Z}'\mathbf{Z}$

$\mathbf{V} = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq (r+1)s^2 F_{r+1,n-r-1(\alpha)}$,

UV

here $F_{r+1,n-r-1(\alpha)}$ is the F value above which the probability is α in the *F* distribution. Interval within which the β_i values will fall with probability $1-\alpha$ is

$$\hat{\beta}_i \pm \sqrt{\sigma^2_{\hat{\beta}_i}}((r+1)F_{r+1,n-r-1(\alpha)})^{0.5},\ i=1,2,\ldots,r$$

If $\mathbf{T}=\mathbf{Z}'\mathbf{Z}$ then $\sigma^2_{\hat{\beta}_i}$ is the is represented by the diagonal elements of the sample error variance multiplied by the inverse of matrix **T.**

An example to highlight the concepts discussed in this section see Example 2 in Appendix A.

3.4.3 Multivariate Multiple Linear Regression (MMLR)

Sometimes on may need to determine the linear relationship between multiple response (dependent) variables with multiple predictors (independent) variables. The technique employed to determine the regression equations is named multivariate multiple linear regression (MMLR). That is the model determines the values of several dependent variables under the influence of predictor.

For each dependent variable the MMLR model can explicitly be expressed as

$$Y_l=\beta_{01}+\beta_{11}z_1+\ldots+\beta_{r1}z_r+\varepsilon_1;\ \ l=1,2,...,\mathrm{m}$$ (Tandogdu Y. & Esager M., 2018)

$\boldsymbol{\varepsilon}=\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_m\end{bmatrix}$ is the random error vector with $E(\boldsymbol{\varepsilon})=\mathbf{0}$, Var($\boldsymbol{\varepsilon}$)=$\boldsymbol{\Sigma}$

It becomes evident that for different response variables, corresponding errors to can have the same correlation.

The vector of response variables are as follows.

$$\mathbf{Y}_j'=[Y_{j1},Y_{j2},\ldots,Y_{jm}]$$

The data represented by the predictor variables is,

$$\left(x_{j0},x_{j1},\ldots,x_{jk}\right),\ \ j=1,2,\ldots,n.$$

Vector of errors for each response variable is

$\boldsymbol{\varepsilon}'_j = [\varepsilon_{j1}, \varepsilon_{j2}, \ldots, \varepsilon_{jm}]$

The matrix with data is written similar to the multivariate regression model.

$$\mathbf{X}_{(n\times(r+1))} = \begin{bmatrix} x_{10} & x_{11} & \ldots & x_{1r} \\ x_{20} & x_{21} & \ldots & x_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \ldots & x_{nr} \end{bmatrix}$$

Similarly the response variables expressed as a $\mathbf{Y}_{(n\times m)}$ matrix, (Tandogdu Y. & Esager M., 2018)

Regression coefficients matrix will be of size

$\boldsymbol{\beta}_{((r+1)\times m)}$,

and the error matrix will be $\boldsymbol{\varepsilon}_{(n\times m)}$.

Applying the least square estimation method to the MMLR system the regression coefficients can expressed as

$\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_{(i)}$

Keeping in mind that there exists *m* regression coefficients

$\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{B} = \mathbf{X}'\mathbf{Y}$

$\hat{\boldsymbol{\beta}} = \mathbf{AB}$ can be written.

Response values $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and errors are $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$

In MMLR the columns of **X,** predicted values $\hat{\mathbf{Y}}$, the residuals $\hat{\boldsymbol{\varepsilon}}$, must satisfy the orthogonality condition.

Analysis of variance in multivariate multiple regression can be summarized as given in Table 3.3.

Table 3. 3: Multivariate Multiple Analysis of Variance

source of variation (source)	(SSCP)	(D.F)	mean square	F-STATISTIC
Treatment	**SSCP between treatment**	r-1	MST= $\frac{\mathbf{SSCPT}}{\mathrm{r-1}}$	$F = \frac{\mathrm{MST}}{\mathrm{MSE}}$= $\frac{9190.88}{\mathbf{3450.55}} = 2.66$
Error	**SSCP between treatment**	n-r	MSE= $\frac{\mathbf{SSCPE}}{\mathrm{n-r}}$	
TOTAL	(SSCT) CORRECTED	n-1		

3.4.4 Assumptions under the Linear Regression

- Linearity

  This implies that a linear combination of the predictor variables gives the mean of the dependent variable and the coefficients are determined from the fitted model

- Linearly independent predictor/s

  Assume that observations are independent of each other. Thus correlation between sequential observations, or auto-correlation, turns out to be an issue associated with time series data that is.

- Probability distribution of the errors has constant variance
- Homoscedasticity of errors which refers to equal variance around the regression line.

3.4.5 Nonlinear Regression

While the topic covered in this thesis is predominantly linear regression, inclusion of a brief note is deemed necessary to remind the future reader that regression does not start and end with linear regression. From linear regression we know that the regression model requires the relationship between variables be linear thus been seen as the best fitted straight line relationship between the response and predictor variables.

In the preliminary study of available data the statistician may be convinced the relationship between the response and predictor variables is not linear. Depending on the behavior exhibited in the for instance scatter plot, one may decide to use an exponential, a polynomial, or some other kind of model to represent the expected relationship $\mu_{Y|X}$, $\mu_{Y|X_1,X_2,\ldots,X_k}$, $\mu_{Y_1,Y_2,\ldots,Y_l|X_1,X_2,\ldots,X_k}$ in the simple, multiple, or multivariate multiple cases respectively.

For nonlinear regression ordinary least squares (OLS) approach can be used to obtain the curve that best fits to the data, with minimized sum of squares and expected value of residuals being zero.

Sometimes using transformation techniques, the nonlinear relation between the response and predictor can be converted to linear, enabling the application of regression techniques used in linear regression.

For example given the following exponential model representing the relation between the response (Y) and predictor (X)

$$y = ae^{bx}u, \; a \text{ and } u \text{ real non-zero constants}$$

a logarithmic transformation will yield

$$\ln(\mathrm{y}) = \ln(a) + \mathrm{bx} + \ln(\mathrm{u})$$

This is a linear relation between Y and *X*, and linear regression rules can be applied. However, obtained results will have to be back transformed by applying the anti-log process, to enable their proper interpretation.

Chapter 4

MOVING AVERAGES

Moving average technique was initially used for trend generation in time or space dependent variables. In essence it's a process that can also be applied to time series. But in place of time any space interval can also be used. Taking any set of observations of a variable measured at successive period of time or space results in a time or space dependent sequence Y_t, where t belongs to the set of integers and denotes the time/space steps. A time series is deterministic, if it is expressed as a known function of time $Y_t = \mathrm{f}(\mathrm{t})$. If X is a random variable and $Y_t = X(t)$, then Yt is called a stochastic time series. Analysis of data in time series aims at prediction, description and control of the process under study. Trend generation is a process that generates a graphical or functional behavior of the variable under study over a period of time or an interval of space. Observations may be spaced out at fixed time or space intervals. They may be spaced out at unequal time or space intervals as well. Lag is the difference in time or space between an observation and a previous one. Thus Y_{t-k} lags Y_t by k periods. When observations are at fixed distance apart, the lag is also fixed. In the case of irregularly spaced data, different approaches have to be taken regarding the lag. For example taking a fixed length as a lag, such that there is at least one observation in it, and in case of more than one observation averaging these and assigning the average to the lag.

For trend generation some of widely used methods are kernel, splines, moving average smoothing techniques. In this study we opted for the use of the moving average technique. Its effect is smoothing the trajectory of a raw data resulting in a trend line depicting an estimate of the expected trend for the variable under consideration. Different approaches are taken in the generation of the moving average trend. Some approaches in used are simple, cumulative, weighted, exponential moving average techniques, just to mention a few.

Finding the average from a starting point up to some final point is required, the *cumulative moving* average can be used. In certain applications higher weights are desired to be allocated for more recent data. In such cases the *weighted moving* average is used. Different weight determination techniques are employed. Determination of weights can be achieved using techniques based on linear or exponentially decreasing order. (Durbin, 1959). Application fields include finance, economy, medical applications, just to mention some.

In a way the moving average is a smoothing method, encompassing a range of sophisticated moving average methods. One such method is the moving average as a trend generator on a trajectory where a smooth trend is generated, such that a smooth value can correspond to every raw data value with the same time or space coordinate. Hence the name Stretched Interpolated Moving Average (SIMA).

The underlying function f in a random process $X(\cdot)$ is unknown. Using data collected at p different locations an idea can be acquired about the function *f*. Prediction of the underlying random function governing the random process from available observations is a hard work, as data tends to include random errors due to various

reasons. Assuming the data does not include any errors some simple linear interpolation methods may adequately represent $X(\cdot)$. Some kind of smoothing is necessary when the errors $\boldsymbol{\varepsilon}$ are present in order to have a more accurate appreciation of the process $X(\cdot)$.

Assume the process $X(\cdot)$ be represented by $Y_i = X(t_i) + \varepsilon_i,\ i = 1,\ldots,n$. The data matrix $\mathbf{Y}$ has $\sigma_{\mathbf{Y}}^2 = \boldsymbol{\Sigma}_e = \sigma^2 \mathbf{I}$.

It must be pointed out that the fitting of ordinary least squares function is in fact a kind of smoothing. However, all data points are assumed to have equal weights that becomes a handicap when data are not observed at regular time or space intervals.

In this thesis the stretched moving average method which is a simple version of the stretched interpolated moving average SIMA (Tandoğdu Y., Çıdar İ. Ö., 2013) Will be used. Interpolation part is not considered necessary, as the aim is to smooth the data using different lag intervals to find out the level of smoothing by comparing error measures such as root mean square deviation (RMSD) or relative root mean square deviation (RRMSD).

In the SIMA method the obtained moving average values at a given lag interval are stretched to cover the whole length of the trajectory, hence averaged values coordinates will be different to those of the data values. Then smoothed average values corresponding to actual data coordinates can be computed by interpolation.

Let random variable Y represent the obtained average values. Assume p_i data values X_{ij} are collected at equal distance apart on the i^{th} trajectory. Then MA is given by

$$Y_{il} = \frac{1}{m}\sum_{j=l}^{m+l-1} X_{ij},\ 2 \le m \le p_i,\ i = 1,\ldots,n,\ l = 1,\ldots,M\ .$$

Where each lag contains $m+1$ data values, and m is data points used for averaging. M is the number of obtained average values and is given by $M = p - m + 1$.

Here the obtained M moving average values are not assigned any coordinate on the trajectory, except they are computed sequentially starting from the first data point. The purpose is just to compare the effect of smoothing against regression smoothing.

Example 4.1: A data set consisting of 20 observations with 2 independent and a response variable is used. Predictor variables are;

X_1: Area of a house in ft^2 x(100).

X_2: Assessed value in thousand dollars

Y: Price the house sold.

A moving average with m = 4 is used resulting in a smoothed data set of 17 smooth observations. These are given in Table 4.1.

Table 4. 1: Raw data with 20 observations and moving average data with $m = 4$

	Raw data				Moving aver data m = 4		
	X_1	X_2	Y		X_1	X_2	Y
1	15.31	57.30	74.80		15.27	60.88	72.93
2	15.20	63.80	74.00		15.09	65.28	72.95
3	16.25	65.40	72.90		15.62	65.13	73.45
4	14.33	57.00	70.00		15.18	63.83	73.23
5	14.57	74.90	74.90		15.32	64.00	74.10
6	17.33	63.20	76.00		15.49	59.38	74.00
7	14.48	60.20	72.00		14.63	57.48	73.38
8	14.91	57.70	73.50		14.81	58.08	73.25
9	15.25	56.40	74.50		14.69	59.50	72.63
10	13.89	55.60	73.50		14.60	60.45	73.73
11	15.18	62.60	71.50		15.78	63.35	76.98
12	14.44	63.40	71.00		15.79	61.98	76.10
13	14.87	60.20	78.90		18.62	68.53	83.85
14	18.63	67.20	86.50		19.66	70.63	85.13
15	15.20	57.10	68.00		18.85	68.85	80.75
16	25.76	89.60	102.00		19.56	71.15	85.75
17	19.05	68.60	84.00		17.21	65.20	79.25
18	15.37	60.10	69.00				
19	18.06	66.30	88.00				
20	16.35	65.80	76.00				

Smoothing effect of the moving averaging process is clearly visible in Figure 4.1 showing the raw and moving average graphs for the variable assessed value for the houses (X_2).

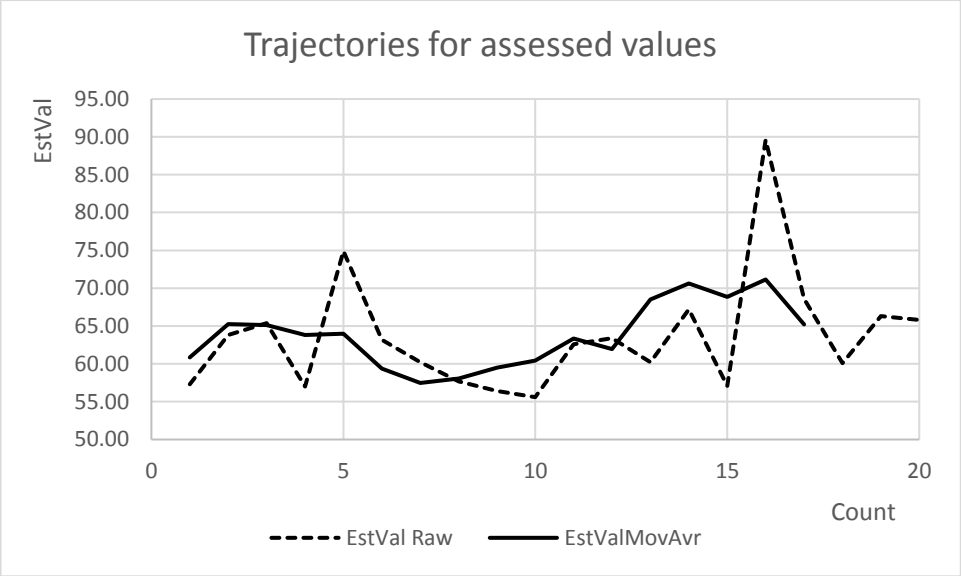

Figure 4. 1: Raw and moving average graphs for the assessed values of houses

Due to the smoothing effect of the moving average process, when regression is applied to the smooth data, obtained regression equation will result in lower error values. This is an expected effect and in the case of repeated sampling the moving average manifests itself as an efficient trend generator (Tandoğdu Y., Çıdar İ. Ö., 2013; Tandoğdu Y., Çıdar İ. Ö., 2013).

Obtained error measures MSD, RMSD, and RRMSD results from the application of simple linear regression (Y on X_1, Y on X_2), and multiple linear regression (Y on X_1, X_2), techniques to the raw and smoothed data are given in Table 4.1 and 4.2 respectively. In all cases it is evident that the error levels obtained from the smoothed data is lower.

Table 4. 2: Obtained error measures from applying regression to raw data.

SOURCE	R	SSE	MSD	RSMD	RRMSD
y *on* x_1	0.913	205.29765	10.26488	3.203886	0.041854
y *on* x_2	0.786	472.8031	23.64015	4.862114	0.063516
y *on* x_1, x_2	0.915	201.2602	10.06301	3.172225	0.04144

Table 4. 3: Obtained error measures from applying regression to smoothed data.

SOURCE	R	SSE	MSD	RSMD	RRMSD
y *on* x_1	0.972	18.807505	1.106324	1.051819	0.013739
y *on* x_2	O.841	99.76221	5.868365	2.422471	0.031644
y *on* x_1, x_2	0.974	17.793482	1.046675	1.023072	0.013364

Chapter 5

APPLICATION

5.1 Case Study of the Application

Multivariate regression and moving averages topics are explained in chapters 3 and 4. The main topic of the thesis is linear regression. To highlight the effect of smoothing on the results of regression one of the smoothing techniques, namely stretched moving averaging is used for smoothing the data. Then regression techniques are applied to both raw and smoothed data. This will bring into focus from application point of view all topics covered in this thesis. Main yard sticks to be used for the assessment of results are mean square error (MSE), root mean square deviation (RMSD), and relative root mean square deviation (RRMSD).

Expressing these parameters openly

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$$

$$RMSD = \sqrt{MSE}$$

The RRMSD is especially useful as it expresses the RMSD value as a percentage of the average response values $\bar{y}$, and is expected to be 0 < RMSD < 1.

$$RRMSD = \frac{RMSD}{\bar{y}}$$

Depending on the value of RMSD the following can be said about the quality of the fitted regression model. (Milan Despotovic, Vladimir Nedic, Danijela Despotovic, Slobodan Cvetanovic, 2015)

RMSD value	Quality of fitted model
RMSD < 0.1	Excellent
0.1 < RMSD < 0.2	Good
0.2 < RMSD < 0.3	Fair
RMSD > 0.3	Poor

5.2 Data used in this Case Study

This data was created by Paulo Cortez and Anibal Morais (Univ. Minho) in 2007. (data.world, 2002). P. Cortez and A. Morais worked on the data (Paulo Cortez, Anibal Morais1, 2007)

This data relates to forest fire of an area in northern Portugal with 517 observations. The fire weather indexes are considered as the dependent variables. These are fine fuel moisture code (FFMC), duff moisture code (DMC), drought code(DC) and initial spread index (ISI) as y_1, y_2, y_3 *and*, y_4 respectively. 5 independent variables are temperature, relative humidity, wind, rain and area represented by x_1 , x_2, x_3, x_4 *and* x_5 respectively.

From the original data temperature (x_1) and relative humidity (x_2) are selected as independent variables, while fine fuel moisture code (FFMC (y_1)) and duff moisture code (DMC (y_2)) were selected as the dependent variables. For many statistical

computations a sample size of over 30 is considered adequate, hence for this case study 50 observations were taken into account. See Appendix II for this data.

5.2 Descriptive Statistics for Raw and Moving Average Data

A glance at some simple statistics for the variables FFMC (y_1), DMC (y_2), TEMP (x_1), RH (x_2) gives a preliminary idea about the nature of the data. Table 5.1 is a summary of these statistics for the raw data.

Table 5. 1: Summary statistics about raw and moving average data

	Variable	Mean $\bar{x}$	StDev s	Median $\tilde{x}$	Q_1	Q_3
Raw data	FFMC (y_1)	90.286	4.748	91.00	89.825	92.500
	DMC (y_2)	82.69	38.89	84.00	43.70	120.05
	Tmp (x_1)	17.962	5.160	18.25	14.475	21.825
	R.H (x_2)	45.46	18.51	41.00	33.00	51.75
M.Av. m=3	FFMC (y_1)	90.399	2.680	91.183	89.642	92.283
	DMC (y_2)	84.63	31.43	84.20	61.20	108.18
	Tmp (x_1)	18.237	3.011	18.32	16.017	19.833
	R.H (x_2)	45.53	9.64	43.83	39.33	50.17
M.Av. m=6	FFMC (y_1)	90.379	1.711	90.717	89.242	91.825
	DMC (y_2)	86.68	23.67	88.95	66.08	102.49
	Tmp (x_1)	18.434	2.002	18.30	17.058	19.817
	R.H (x_2)	45.25	6.90	42.33	40.67	49.50

From the above data it can be seen that the standard deviation for all Raw data > M.Av. (m=3) > M.Av. ($m=6$). This is a direct result of smoothing the data via the moving averaging process.

From these statistics for the raw data it can be observed that the variables

FFMC (y_1), DMC (y_2), and temperature (x_1) are fairly symmetrically distributed as their $\bar{x} \approx \tilde{x}$.

R.H (x_2) is positively skewed (skewed towards right) as $\bar{x} > \tilde{x}$.

After smoothing with *m = 3*, FFMC (y_1), DMC (y_2), and temperature (x_1) are fairly symmetrically distributed as their $\bar{x} \approx \tilde{x}$.

R.H (x_2) is positively skewed (skewed towards right) as $\bar{x} > \tilde{x}$.

After smoothing with *m = 6*, DMC (y_2), and temperature (x_1) are fairly symmetrically distributed as their $\bar{x} \approx \tilde{x}$.

R.H (x_2) is still positively skewed (skewed towards right) as $\bar{x} > \tilde{x}$.

5.3 Multivariate Analysis for Raw and Data Obtained from Moving Averaging

Graphs are drawn for all (x_i, y_i) pairs in order to see the effect of moving average smoothing on the trend of the data. Figure 5.2 is the graph of raw x_1 versus the raw, and moving average values obtained from *m* = 3 and *m* = 6 smooth cases of y_1. Similarly Figures 5.2, 3, and 4 are the graphs of x_1 versus y_2 , x_2 versus y_1 , and x_2 versus y_2 for the same variables. Effect of smoothing due to moving averaging process is clearly visible in all graphs. Numerical assessment of the smoothing becomes more

evident when regression is applied to the smooth data, results of which are summarized in Table 5.1.

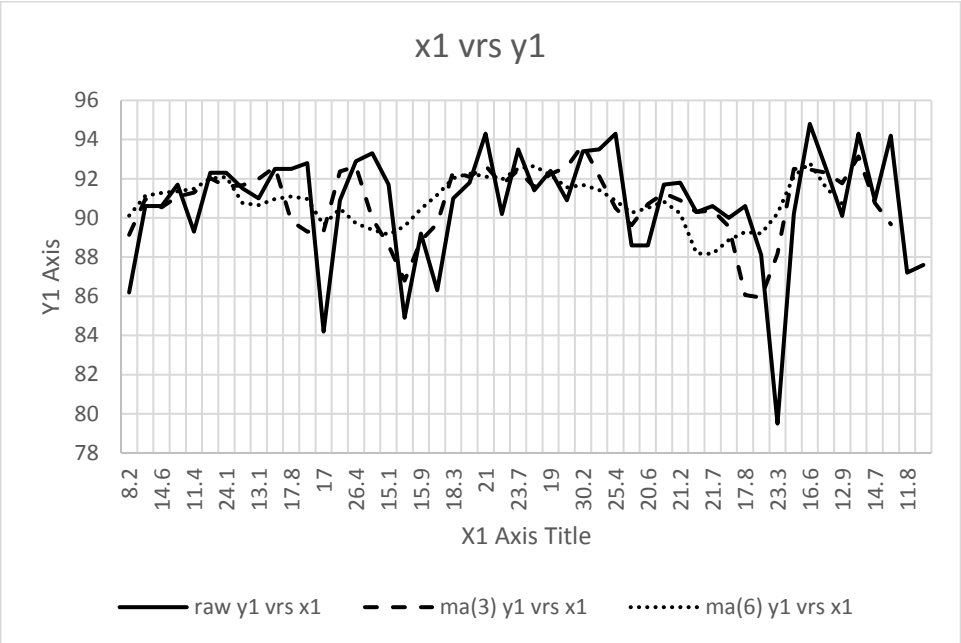

Figure 5. 1: $y_1|x_1$ versus the raw, and moving average values obtained from m = 3 and m = 6 smooth cases of y1

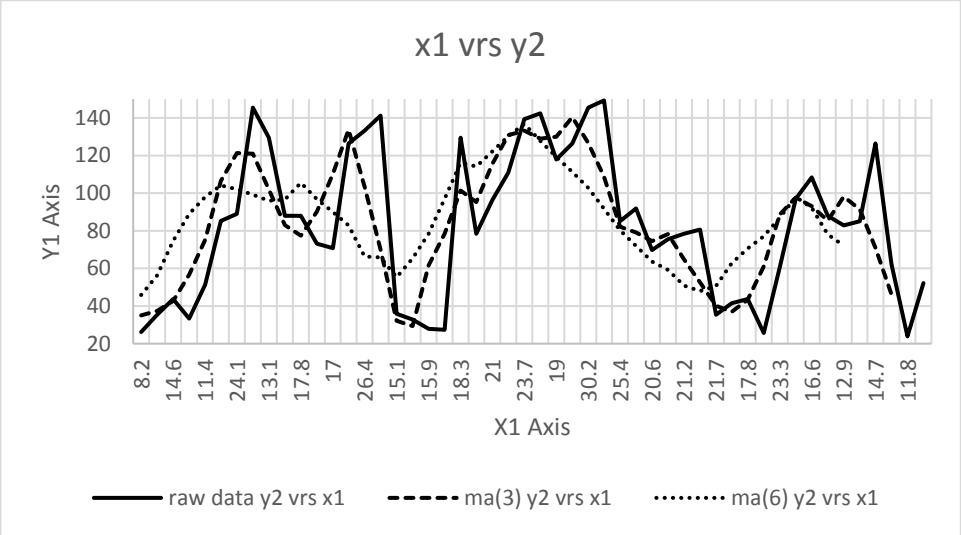

Figure 5. 2: $y_2|x_1$ between the raw, and moving average values obtained from m = 3 and m = 6 smooth

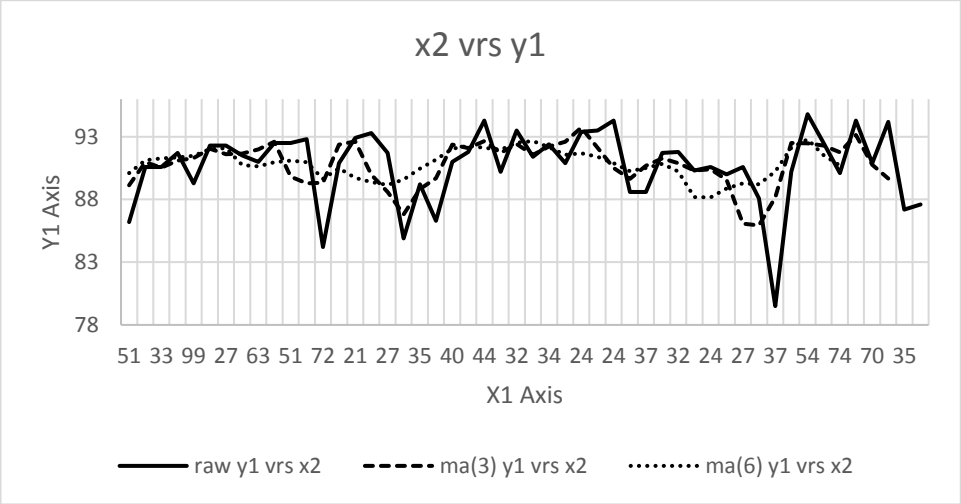

Figure 5. 3: $y_1|x_2$ between the raw, and moving average values obtained from m = 3 and m = 6 smooth cases

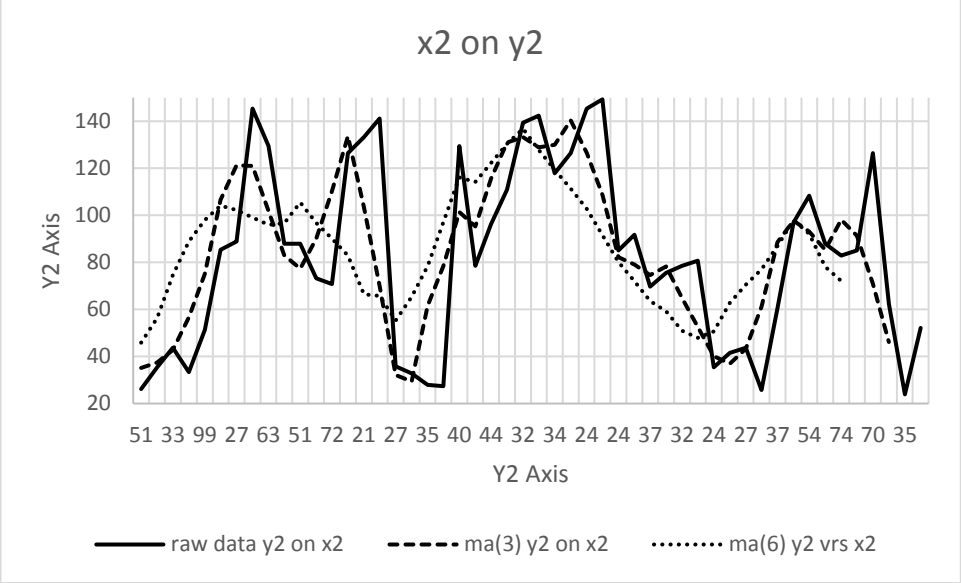

Figure 5. 4: $y_2|x_2$ between the raw, and moving average values obtained from m = 3 and m = 6 smooth case

Regression analysis of the response variables $y_1, y_2|x_1$, $y_1, y_2|x_2$, and $y_1, y_2|x_1, x_2$ are undertaken and parameters used for the assessment of the quality of regression are given in Table 5.2

Table 5. 2: R, MSD, RMSD, and RRMSD values from raw and smooth data for the regression cases

		$y_1, y_2 \mid x_1$	$y_1, y_2 \mid x_2$	$y_1, y_2 \mid x_1, x_2$
Raw data	R	0.268 0.427	0.257 0.032	0.91 0.61
	MSD	20.49 1213.12	20.63 1480.83	20.27 924.75
	RMSD	4.53 34.83	4.54 38.48	4.50 30.41
	RRMSD	0.050 0.421	0.050 0.465	0.050 0.368
M. Avr. m= 3	R	0.224 0.673	0.032 0.089	0.806 0.763
	MSD	6.684 529.032	7.027 958.908	6.579 0.404
	RMSD	2.585 23.001	2.651 30.966	2.565 20.088
	RRMSD	0.029 0.272	0.029 0.366	0.028 0.237
M. Avr. m = 6	R	0.164 0.707	0.071 0.055	0.261 0.835
	MSD	2.787 273.923	2.850 546.278	2.669 165.520
	RMSD	1.669 16.551	1.688 23.373	1.633 1.634
	RRMSD	0.018 0.191	0.019 0.270	0.018 0.148

From the above statistic it can been seen that the MSD and associated parameters RMSD and RRMSD are maximum in the regression undertaken using the raw data. They decrease as the number of values used in the moving average process increase as expected. This trend is valid for all 3 regressions ($y_1, y_2 \mid x_1$, $y_1, y_2 \mid x_2$, and $y_1, y_2 \mid x_1, x_2$) used.

Chapter 6

CONCLUSION

This study was aimed at exploring some theoretical characteristics of linear regression starting with simple linear regression, proceeding up to the multivariate multiple linear regression. As regression inherently harbors smoothing, it was considered informative to compare its performance when data is smoothed using one of the smoothing techniques that is widely used. Moving average technique was implemented for smoothing with 2 different lags, namely using lag with $m = 3$ and lag with $m = 6$ values for averaging. Clearly the larger the lag of the moving average, the smoother the obtained trend will be.

This idea of seeing the relation between the regression and the smoothed regression a case study was undertaken in chapter 5. The graphical representation of smoothing raw data via moving average technique is clearly visible in Figures 5.1, 2, 3, and 4. Following the application of multiple and multivariate regression techniques on the same raw and smoothed data sets has clearly shown the effect of smoothing by means of reduced errors as summarized in Table 5.2.

Noteworthy benefits of smoothing the data before applying regression are

i. Number of observations should be large or very large, i.e. in the order of hundreds or thousands to realize the real benefit of smoothing.

ii. High fluctuations in the observed values, especially in the response variable/s yielding high variation from the general trend of the data is partially eliminated.

iii. Reduced error level obtained from the regression of smooth data compared to error levels obtained from regressing the raw data enables more sound estimation and projections for the future.

Therefore, it can be recommenced that especially in the case where large data sets are to be used for regression study, some smoothing can be beneficial as it will result in reduced estimation errors.

REFERENCES

Forest Fire.(2002). Retrieved from data.world: https://data.world/uci/forest-fires

Galton, F. (1989). Kinships and corelation. *statistical science*, 81-86.

Guass, C. F. (1821). Anzeige. in c. f. werke, *anzeige theoria combinationis observationum erroribus minimis obnoxiae* (pp. 95-100). pars prior.

Karen A. Randolph & Laura L. Myers. (2013). *Basic statistics in multivariate analysis.* united state of america: oxford university press.

Karl Pearson, G. U. Yule,Blanchard, Norman; Lee,Alice. (1903). The law of ancestral heredity. *biometrika trust*, : 211–236.

Keenan Pituch & James P. Stevens. (2016). *Applied multivariate statistics for social sciences.* new york: routledge.

Milan Despotovic, Vladimir Nedic, Danijela Despotovic, Slobodan Cvetanovic. (2015). Renewable and sustainable energy reviews. *elsevier*, 12.

Minitab, S. S. (2011). *Mintab software for statistics.* wikipedia.

Nasha, K. J. (2015). Estimation of multivarite multiple regression. 12.

Neil, H. T. (2002). *Applied multivariate analysis.* new york: springer-verlag.

Paulo Cortez, Anibal Morais1. (2007). A data mining approach to predict forest fires. *pcortez*, 12.

Richard R.A & Wichen D.W. (2007). Linear regression . in *applied multivariate statistics analysis* . new jersey: upper saddle river.

Ricker, W. E. (1984). Computation and uses of central trend lines. *can. j. zool*, 62:1897-905.

Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers,Keying Ye. (2011). Mathematical expectations. in *probability & statistics for engineers* (pp. 111-142). boylston: prentice.

Stigler, S. M. (1980). Gauss & the invention of least squares. *institute of mathematical statistics*, vol. 9 n.3 pp.465-474.

Tandogdu Y. & Esager M. (2018). The sensitivity of the regression parameter.

Tandoğdu Y., Çıdar I. Ö. (2013). Stretched interpolated moving average. *pak. j. statist.*, 16.

Yates, F. (1966). Computer the second revolution in statistics. *rothamshed experimental station*.

York, D. (1966). Least square fitting of a straight line. *canadian journal of physics* , 1079-1086.

APPENDICES

Appendix A: Examples

Example 1: A data containing 5 observation with 2 variables is used for the computation where *X* represent temperature and where *Y* is representing rented bikes.

Rented Bike Count **y**	Temperature(°C) **x**
507	-0.4
390	-1.4
402	-2.2
389	-2.7
259	-3.2

For the above data the fitted regression equation is computed as

$$\hat{y} = 528.243 + 70.12275(\boldsymbol{x})$$

$$\overline{y} = 389.4$$

Now calculating sum of square regression

Y	**X**	$\hat{\boldsymbol{y}}$	$(\hat{\boldsymbol{y}} - \overline{\boldsymbol{y}})$	$(\hat{\boldsymbol{y}} - \overline{\boldsymbol{y}})^{\boldsymbol{2}}$
507	-0.4	500.1939	110.7939	12275.29
390	-1.4	430.07115	40.67115	1654.142
402	-2.2	373.97295	-15.42705	237.9939
389	-2.7	338.911575	-50.488425	2549.081
259	-3.2	303.8502	-85.5498	7318.768
1947		1946.99978	-0.000225	24035.27

Therefore the sum of square regression = 24035.27

Now we calculate total sum of square

Y	**X**	$(\mathbf{y}-\overline{\boldsymbol{y}})$	$(\mathbf{y}-\overline{\boldsymbol{y}})^{\mathbf{2}}$
507	-0.4	117.6	13829.76
390	-1.4	0.6	0.36
402	-2.2	12.6	158.76
389	-2.7	-0.4	0.16
259	-3.2	-130.4	17004.16
1947	-1.98	0	30993.2

Therefore the total sum of square is = 30993.2

Now calculating sum of square error

Y	**X**	$\hat{\boldsymbol{y}}$	$(\boldsymbol{y}-\hat{\boldsymbol{y}})$	$(\mathbf{y}-\hat{\boldsymbol{y}})^{\mathbf{2}}$
507	-0.4	500.1939	6.8061	46.323
390	-1.4	430.0712	-40.0712	1605.697
402	-2.2	373.973	28.02705	785.5155
389	-2.7	338.9116	50.08842	2508.85
259	-3.2	303.8502	-44.8502	2011.54
1947		1947	0.000225	6957.926

Therefore the sum of square error = 6957.926

Since we know the values of SSE, SSR and SST already we then use them to find the rest.

SSE=6957.926 SSR=24035.27 and SST=30993.2 n=5

Therefore our table becomes

Table A1: ANOVA associated with Simple Regression for the example

(SOURCE)	(SS)	(D.F)	MEAN SQUARE	F-STATISTIC
REGRESSION	(SSR)= 24035.27	1	MSR= 24035.27	F = 10.363
ERROR	(SSE)= 6957.926	3	MSE= 2319.308	
TOTAL	(SST)= 30993.2	4		

$$r^2 = \frac{24035.27}{30993.2} = 0.775501$$

Computations of the example related with the analysis of variance in *multiple linear regression.*

Example 2: A data containing 5 observation with 4 variables is used for the computation where x1 represent temperature, x2 represent humidity, x3 represent humidity and where y is representing rented bikes.

Table A2: Raw data used

Rented Bike Count	Temperature (°C)	Humidity (%)	Wind speed
507	-0.4	47	1.1
390	-1.4	47	2.1
402	-2.2	46	1.8
389	-2.7	48	3.5
259	-3.2	50	1.6

$$\boldsymbol{X} = \begin{bmatrix} 1 & -0.4 & 47 & 1.1 \\ 1 & -1.4 & 47 & 2.1 \\ 1 & -2.2 & 46 & 1.8 \\ 1 & -2.7 & 48 & 3.5 \\ 1 & -3.2 & 50 & 1.6 \end{bmatrix} \qquad \boldsymbol{X}^{\boldsymbol{T}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -0.4 & -4.4 & -2.2 & -2.7 & -3.2 \\ 47 & 47 & 46 & 48 & 50 \\ 1.1 & 2.1 & 1.8 & 3.5 & 1.6 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 507 \\ 390 \\ 402 \\ 389 \\ 259 \end{bmatrix}$$

$$\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -0.4 & -4.4 & -2.2 & -2.7 & -3.2 \\ 47 & 47 & 46 & 48 & 50 \\ 1.1 & 2.1 & 1.8 & 3.5 & 1.6 \end{bmatrix} \times \begin{bmatrix} 1 & -0.4 & 47 & 1.1 \\ 1 & -1.4 & 47 & 2.1 \\ 1 & -2.2 & 46 & 1.8 \\ 1 & -2.7 & 48 & 3.5 \\ 1 & -3.2 & 50 & 1.6 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & -9.9 & 238 & 10.1 \\ -9.9 & 24.49 & -475 & -21.91 \\ 238 & -475 & 11338 & 481 \\ 10.1 & 21.91 & 481.2 & 23.67 \end{bmatrix}$$

$$\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}$$

$$\begin{bmatrix} 425.6 & -9.3 & -9.1 & -4.8 \\ -9.2 & 0.5 & 0.2 & 0.3 \\ -9.1 & 0.2 & 0.2 & 0.1 \\ -4.8 & 0.3 & 0.1 & 0.4 \end{bmatrix}$$

$$\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -0.4 & -4.4 & -2.2 & -2.7 & -3.2 \\ 47 & 47 & 46 & 48 & 50 \\ 1.1 & 2.1 & 1.8 & 3.5 & 1.6 \end{bmatrix} \times \begin{bmatrix} 507 \\ 390 \\ 402 \\ 389 \\ 259 \end{bmatrix} = \begin{bmatrix} 1947 \\ -3512.3 \\ 92273 \\ 3876.2 \end{bmatrix}$$

$$\hat{\beta} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

$$\begin{bmatrix} 425.6 & -9.3 & -9.1 & -4.8 \\ -9.2 & 0.5 & 0.2 & 0.3 \\ -9.1 & 0.2 & 0.2 & 0.1 \\ -4.8 & 0.3 & 0.1 & 0.4 \end{bmatrix} \times \begin{bmatrix} 1947 \\ -3512.3 \\ 92273 \\ 3876.2 \end{bmatrix} = \begin{bmatrix} 1183.5 \\ 66.7 \\ -14.9 \\ 23.7 \end{bmatrix}$$

$$\hat{Y} = 1183.468 + 66.68709(x_1) - 14.9123(x_2) + 23.66191(x_3)$$

The vector of the fitted value is

$$\hat{y} = X\hat{\beta}$$

$$\begin{bmatrix} 1 & -0.4 & 47 & 1.1 \\ 1 & -1.4 & 47 & 2.1 \\ 1 & -2.2 & 46 & 1.8 \\ 1 & -2.7 & 48 & 3.5 \\ 1 & -3.2 & 50 & 1.6 \end{bmatrix} \times \begin{bmatrix} 1183.5 \\ 66.7 \\ -14.9 \\ 23.7 \end{bmatrix} = \begin{bmatrix} 481.9 \\ 438.9 \\ 393.4 \\ 370.4 \\ 262.3 \end{bmatrix}$$

The residual becomes

$$\hat{e} = y - \hat{y}$$

$$\begin{bmatrix} 507 \\ 390 \\ 402 \\ 389 \\ 259 \end{bmatrix} - \begin{bmatrix} 481.9 \\ 438.9 \\ 393.4 \\ 370.4 \\ 262.3 \end{bmatrix} = \begin{bmatrix} 25.1 \\ -48.9 \\ 8.6 \\ 18.6 \\ -3.3 \end{bmatrix}$$

And residual sum of square is

$$\hat{\varepsilon}^2 = \hat{\varepsilon}^T \hat{\varepsilon}$$

$$\begin{bmatrix} 25.1 & -48.9 & 8.6 & 18.6 & -3.3 \end{bmatrix} \times \begin{bmatrix} 25.1 \\ -48.9 \\ 8.6 \\ 18.6 \\ -3.3 \end{bmatrix} = 3450.6$$

$$\hat{\varepsilon}^2 = \hat{\varepsilon}^T \hat{\varepsilon} = 3450.567$$

Table 3.5

Y	**x**	$(\mathbf{y}-\overline{\boldsymbol{y}})$	$(\mathbf{y}-\overline{\boldsymbol{y}})^{\mathbf{2}}$
507	-0.4	117.6	13829.76
390	-1.4	0.6	0.36
402	-2.2	12.6	158.76
389	-2.7	-0.4	0.16
259	-3.2	-130.4	17004.16
1947	-1.98	0	**30993.2**

Therefore the total sum of square **= 30993.2**

Now we calculate sum of square regression

$$\hat{Y} = 1183.468 + 66.68709(\boldsymbol{x_1}) - 14.9123(\boldsymbol{x_2}) + 23.66191(\boldsymbol{x_3})$$

Rented Bike Count	Temperature	Humidity	Wind speed	$\widehat{\boldsymbol{y}}$	$\widehat{\boldsymbol{y}}-\overline{\boldsymbol{y}}$	$(\widehat{\boldsymbol{y}}-\overline{\boldsymbol{y}})^{\mathbf{2}}$
507	-0.4	47	1.1	481.9432	92.54317	8564.237
390	-1.4	47	2.1	438.918	49.51799	2452.031
402	-2.2	46	1.8	393.382	3.98204	15.85664
389	-2.7	48	3.5	370.4391	-18.9609	359.5141
259	-3.2	50	1.6	262.3134	-127.087	16151.01
389.4						**27542.65**

Therefore the sum of square regression = 27542.65

$$r^2 = \frac{27542.65}{30993.2} = 0.8886675142$$

Table A3: ANOVA for Multiple Regression for the example

(SOURCE)	(SS)	(D.F)	MEAN SQUARE	F-STATISTIC
Regression	(SSR)=	3	MSR= $\frac{27542.65}{3}$ $= 9190.88$	$F = \frac{\text{MSR}}{\text{MSE}}$= $\frac{9190.88}{\mathbf{3450.55}} = 2.66$
Error	(SSE)=	1	MSE= $\frac{\mathbf{3450.55}}{1}$	
TOTAL	(SST)=	4		

Example highlighting the concepts in multivariate multiple linear regression (MMLR).

Example 3: In business environment it is common practice that the amount purchased of a certain product depends on many characteristics represented by variables. Similarly the quality of a product also depends on various variables. In this example it is given that the amount purchased (y_1) and the quality (y_2) depends on the palatability (x_1) and the texture (x_2) of the product. A small data set of 5 observations are taken to highlight the application of MMLR technique.

Table A4: Amount purchased (y_1) and the quality (y_2) as dependent, palatability (x_1) and the texture (x_2) as independent variables of a product.

purchase (y_1)	quality (y_2)	palatability (x_1)	Texture (x_1)
507	44	47	57
390	58	47	66
402	28	46	41
389	44	48	74
259	59	50	49

Expressing the dependent and independent variables as matrices **Y** and **X**.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 47 & 57 \\ 1 & 47 & 66 \\ 1 & 46 & 41 \\ 1 & 48 & 74 \\ 1 & 50 & 49 \end{bmatrix} \qquad \boldsymbol{X^T} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 47 & 47 & 46 & 48 & 50 \\ 57 & 66 & 41 & 74 & 49 \end{bmatrix} \qquad \boldsymbol{Y} = \begin{bmatrix} 507 & 44 \\ 390 & 58 \\ 402 & 28 \\ 389 & 44 \\ 259 & 59 \end{bmatrix}$$

Proceeding with necessary computations using matrix operations to obtain the regression parameters

$$\left(\boldsymbol{X^T X}\right)$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 47 & 47 & 46 & 48 & 50 \\ 57 & 66 & 41 & 74 & 49 \end{bmatrix} \times \begin{bmatrix} 1 & 47 & 57 \\ 1 & 47 & 66 \\ 1 & 46 & 41 \\ 1 & 48 & 74 \\ 1 & 50 & 49 \end{bmatrix} = \begin{bmatrix} 5 & 238 & 287 \\ 238 & 11338 & 13669 \\ 287 & 13669 & 17163 \end{bmatrix}$$

$$\left(\boldsymbol{X^T X}\right)^{-1}$$

$$\begin{bmatrix} 246.9 & -5.15 & -0.02 \\ -5.15 & 0.12 & -0.001 \\ -0.02 & -0.001 & 0.001 \end{bmatrix}$$

$$\boldsymbol{X^T Y}$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 47 & 47 & 46 & 48 & 50 \\ 57 & 66 & 41 & 74 & 49 \end{bmatrix} \times \begin{bmatrix} 507 & 44 \\ 390 & 58 \\ 402 & 28 \\ 389 & 44 \\ 259 & 59 \end{bmatrix} = \begin{bmatrix} 1947 & 233 \\ 92273 & 11144 \\ 112598 & 13631 \end{bmatrix}$$

Finally the solution of the matrix equation using the data he parameters $\hat{\boldsymbol{\beta}}$ are found.

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

$$\begin{bmatrix} 246.9 & -5.15 & -0.02 \\ -5.15 & 0.12 & -0.001 \\ -0.02 & -0.001 & 0.001 \end{bmatrix} \times \begin{bmatrix} 1947 & 233 \\ 92273 & 11144 \\ 112598 & 13631 \end{bmatrix} = \begin{bmatrix} 2451.16 & -233.9 \\ -45.40 & -0.45 \\ 1.73 & 0.31 \end{bmatrix}$$

From obtained $\hat{\boldsymbol{\beta}} = \begin{bmatrix} 2451.16 & -233.94 \\ -45.40 & -0.45 \\ 1.73 & 0.31 \end{bmatrix}$ the fitted regression equations for the response variables Y_1 and Y_2 are written as

$$\hat{y}_1 = 2451.16 - 45.404x_1 + 1.732953x_2$$
$$\hat{y}_2 = -233.938 - 0.45217x_1 + 0.310137x_2$$

Using matrix operations the predicted value are computed as

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

$$\begin{bmatrix} 1 & 47 & 57 \\ 1 & 47 & 66 \\ 1 & 46 & 41 \\ 1 & 48 & 74 \\ 1 & 50 & 49 \end{bmatrix} \times \begin{bmatrix} 2451.16 & -233.9 \\ -45.40 & -0.45 \\ 1.73 & 0.31 \end{bmatrix} = \begin{bmatrix} 415.9 & 43.2 \\ 431.5 & 46.0 \\ 433.6 & 32.7 \\ 400 & 54.0 \\ 265.9 & 57.2 \end{bmatrix}$$

RESIDUALS SUM OF SQURES CROSS PRODUCT

$$\hat{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

$$\begin{bmatrix}507\\390\\402\\389\\259\end{bmatrix}-\begin{bmatrix}415.9492\\431.5458\\433.626\\400.0054\\265.8735\end{bmatrix}=\begin{bmatrix}91.05077\\-41.5458\\-31.626\\-11.0054\\-6.87353\end{bmatrix}$$

$$\begin{bmatrix}44\\58\\28\\44\\59\end{bmatrix}-\begin{bmatrix}43.16415\\45.95538\\32.68228\\53.95615\\57.24205\end{bmatrix}=\begin{bmatrix}0.835855\\12.04462\\-4.68228\\-9.95615\\1.757954\end{bmatrix}$$

$$\hat{\boldsymbol{\varepsilon}}^{2}=\hat{\boldsymbol{\varepsilon}}^{T}\hat{\boldsymbol{\varepsilon}}$$

$$\begin{bmatrix}91.05077 & -41.5458 & -31.626 & -11.0054 & -6.87353\end{bmatrix}\begin{bmatrix}91.05077\\-41.5458\\-31.626\\-11.0054\\-6.87353\end{bmatrix}=11184.87$$

$$\begin{bmatrix}0.835855 & 12.04462 & -4.68228 & -9.95615 & 1.757954\end{bmatrix}\begin{bmatrix}0.835855\\12.04462\\-4.68228\\-9.95615\\1.757954\end{bmatrix}=269.9105$$

Appendix B: Data

Raw data used in the case stuy in Chapter 5.

FFMC(y1)	DMC(y2)	Temp(x1)	RH(x2)
86.2	26.2	8.2	51.0
90.6	35.4	18	33.0
90.6	43.7	14.6	33.0
91.7	33.3	8.3	97.0
89.3	51.3	11.4	99.0
92.3	85.3	22.2	29.0
92.3	88.9	24.1	27.0
91.5	145.4	8.0	86.0
91.0	129.5	13.1	63.0
92.5	88.0	22.8	40.0
92.5	88.0	17.8	51.0
92.8	73.2	19.3	38.0
63.5	70.8	17.0	72.0
90.9	126.5	21.3	42.0
92.9	133.3	26.4	21.0
93.3	141.2	22.9	44.0
91.7	35.8	15.1	27.0
84.9	32.8	16.7	47.0
89.2	27.9	15.9	35.0
86.3	27.4	9.3	44.0
91.0	129.5	18.3	40.0
91.8	78.5	19.1	38.0
94.3	96.3	21.0	44.0
90.2	110.9	19.5	43.0
93.5	139.4	23.7	32.0
91.4	142.4	16.3	60.0
92.4	117.9	19.0	34.0
90.9	126.5	19.4	48.0
93.4	145.4	30.2	24.0
93.5	149.3	22.8	39.0
94.3	85.1	25.4	24.0
88.6	91.8	11.2	78.0
88.6	69.7	20.6	37.0
91.7	75.6	17.7	39.0
91.8	78.5	21.2	32.0
90.3	80.7	18.2	62.0

90.6	35.4	21.7	24.0
90.0	41.5	11.3	60.0
90.6	43.7	17.8	27.0
88.1	25.7	14.1	43.0
79.5	60.6	23.3	37.0
90.2	96.9	18.4	42.0
94.8	108.3	16.6	54.0
92.5	88.0	19.6	48.0
90.1	82.9	12.9	74.0
94.3	85.1	25.9	24.0
90.9	126.5	14.7	70.0
94.2	62.3	23.0	36.0
87.2	23.9	11.8	35.0
87.6	52.2	11.0	46.0

MOVING AVERAGE OF ORDER 3

Y1	Y2	X1	X2
89.1	35.1	13.6	39.0
90.9	37.5	13.6	54.3
90.5	42.8	11.4	76.3
91.1	56.6	14.	75.0
91.3	75.2	19.2	51.7
92.0	106.5	18.1	47.3
91.6	121.3	15.1	58.7
91.7	121.0	14.6	63
92	101.8	17.9	51.3
92.6	83.1	20.0	43.0
82.9	77.3	18.0	53.7
82.4	90.2	19.2	50.7
82.4	110.2	21.6	45.0
92.4	133.7	23.5	35.7
92.6	103.4	21.5	30.7
89.9	69.9	18.2	39.3
88.6	32.2	15.9	36.3
86.8	29.4	14.	42.0
88.8	61.6	14.5	39.7
89.7	78.5	15.6	40.7
92.4	101.4	19.5	40.7
92.1	95.2	19.9	41.7
92.7	115.5	21.4	39.7
91.7	130.9	19.8	45.0
92.4	133.2	19.7	42.0

91.6	128.9	18.2	47.3
92.2	129.9	22.9	35.3
92.6	140.4	24.1	37.0
93.7	126.6	26.1	29.0
92.1	108.7	19.8	47.0
90.5	82.2	19.1	46.3
89.6	79.0	16.5	51.3
90.7	74.6	19.8	36.0
91.3	78.3	19.0	44.3
90.9	64.9	20.4	39.3
90.3	52.5	17.1	48.7
90.4	40.2	16.9	37.0
89.6	37.0	14.4	43.3
86.1	43.3	18.4	35.7
85.9	61.1	18.6	40.7
88.2	88.6	19.4	44.3
92.5	97.7	18.2	48.0
92.5	93.1	16.4	58.7
92.3	85.3	19.5	48.7
91.8	98.2	17.8	56.0
93.1	91.3	21.2	43.3
90.8	70.9	16.5	47.0
89.7	46.1	15.3	39.0

MOVING AVERAGE OF ORDER 6

y1	y2	x1	x2
90.1	45.9	13.8	57.0
91.1	56.3	16.4	53.0
91.3	74.7	14.8	61.8
91.4	89.0	14.5	66.8
91.5	98.1	16.9	57.3
92.1	104.2	18.0	49.3
92.1	102.2	17.5	50.8
87.3	99.2	16.3	58.3
87.2	96.0	18.6	51.0
87.6	96.6	20.8	44.0
87.7	105.5	20.8	44.7
87.6	96.8	20.3	40.7
86.2	90.1	19.9	42.2
90.5	82.9	19.7	36.0
89.7	66.4	17.7	36.3
89.4	65.8	16.4	39.5

89.2	55.3	15.73	38.5
89.6	65.4	16.7	41.3
90.5	78.4	17.2	40.7
91.2	97.0	18.5	40.2
92.0	116.2	19.7	42.8
92.3	114.2	19.8	41.8
92.1	122.2	19.8	43.5
92.0	130.4	21.4	40.2
92.5	136.8	21.9	39.5
92.7	127.8	22.2	38.2
92.2	119.3	21.3	41.2
91.6	111.3	21.6	41.7
91.7	102.9	21.3	40.2
91.4	91.7	19.8	41.5
90.9	80.2	19.1	45.3
90.3	72.0	18.4	45.3
90.5	63.6	18.5	42.3
90.8	59.2	18.0	40.7
90.2	51.0	17.4	41.3
88.2	47.9	17.7	42.2
88.7	50.6	17.8	38.8
88.9	62.8	16.9	43.8
89.3	70.5	18.3	41.8
89.2	77.1	17.5	49.7
90.3	87.0	19.5	46.5
92.2	98.0	18.0	52.0
92.8	92.2	18.8	51.0
91.5	78.1	18.0	47.8
90.8	72.2	16.6	47.5

Graphs

FFMC(y1)	DMC(y2)	temp(x1)	RH(x2)
86.2	26.2	8.2	51.0
90.6	35.4	18.0	33.0
90.6	43.7	14.6	33.0
91.7	33.3	8.3	97.0
89.3	51.3	11.4	99.0
92.3	85.3	22.2	29.0
92.3	88.9	24.1	27.0
91.5	145.4	8.0	86.0
91.0	129.5	13.1	63.0.0

92.5	88.0	22.8	40.0
92.5	88.0	17.8	51.0
92.8	73.2	19.3	38.0
63.5	70.8	17.0	72.0
90.9	126.5	21.3	42.0
92.9	133.3	26.4	21.0
93.3	141.2	22.9	44.0
91.7	35.8	15.1	27.0
84.9	32.8	16.7	47.0
89.2	27.9	15.9	35.0
86.3	27.4	9.3	44.0
91.0	129.5	18.3	40.0
91.8	78.5	19.1	38.0
94.3	96.3	21.0	44.0
90.2	110.9	19.5	43.0
93.5	139.4	23.7	32.0
91.4	142.4	16.3	60.0
92.4	117.9	19.0	34.0
90.9	126.5	19.4	48.0
93.4	145.4	30.2	24.0
93.5	149.3	22.8	39.0
94.3	85.1	25.4	24.0
88.6	91.8	11.2	78.0
88.6	69.7	20.6	37.0
91.7	75.6	17.7	39.0
91.8	78.5	21.2	32.0
90.3	80.7	18.2	62.0
90.6	35.4	21.7	24.0
90.0	41.5	11.3	60.0
90.6	43.7	17.8	27.0
88.1	25.7	14.1	43.0
79.5	60.6	23.3	37.0
90.2	96.9	18.4	42.0
94.8	108.3	16.6	54.0
92.5	88.0	19.6	48.0
90.1	82.9	12.9	74.0
94.3	85.1	25.9	24.0
90.9	126.5	14.7	70.0
94.2	62.3	23.0	36.0
87.2	23.9	11.8	35.0
87.6	52.2	11.0	46.0