The Sensitivity of the Regression Parameters

Marwa Esager

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
November 2018
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

_____

Assoc. Prof. Dr. Ali Hakan Ulusoy
Acting Director

I certify that this thesis satisfies 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. Evren Hınçal _____

2. Asst. Prof. Dr. Nidai Şemi _____

3. Asst. Prof. Dr. Yücel Tandoğdu _____

ABSTRACT

Theory of statistics and its application to data analysis in all fields of endeavor, forms the base of statistical analysis. While data collection and validation is not dealt with in this thesis, the importance of data on the analysis results is obvious. In general data are discrete observations in a continuous process. Number of variables involved is also very important. Hence, multivariate statistical analysis has gained importance, especially after computers could be used to process huge amounts of data.

In this thesis, simple and multivariate regression techniques, principal component analysis are explained in detail, and also used in the analysis of a real life data. Since the matrix algebra is implemented in all computations, a brief introduction to certain concepts of matrix algebra is also given under Chapter 3. Chapter 4 introduces some important concepts of multivariate linear regression theory, while Chapter 5 gives basic theoretical background to principal component analysis.

In the application section a data set consisting of 8 variables affecting the heating load of buildings is studied. Following careful examination of the variables and pairwise correlations, it was considered useful to reduce the number of variables to 5, all having a high correlation with the dependent variable. An attempt is made to estimate the predictor variables after the principal components were obtained. The methodology used proved to be a successful one as estimation errors were minimal.

Keywords: Matrix algebra, regression analysis, estimation, predictors, response, regression coefficients, principal components analysis, eigenvector, eigenvalue.

ÖZ

İstatistik teorisi ve her alandaki veri analizine olan uygulaması istatistiğin temelini oluşturur. Veri toplanması, temizlenmesi veya geçerliliğnin saptanması bu tezin kapsamına alınmadı. Ancak verilerin istatistik analizi ve analiz sonuçları üzerindeki etkisi ortadadır. Veriler sürekli bir olayın ayrık ölçümlerinden elde edilen değerlerdir. Bu nedenle çok deişkenli istatistiksel analiz gittikçe önem kazanmakta ve özellikle bilgsayar yazılımlarının istatistik analizindeki kullanımı bu önemi dahada artırmaktadır.

Bu tezde basit ve çok değişkenli regresyon teknikleri, temel bileşenler analizi detaylı olarak açıklanmıştır. Teorik kısmın uygulaması çok değişkenli gerçek bir veri üzerinde yapılmıştır. Tüm uygulamalarda matris cebiri teorilerinden yararlanıldığı için, matris cebiri ile ilgili bazı temel kavramlar kısaca açıklanmıştır. Dördüncü kısımda çok değişkenli lineer regresyon teorisi ile ilgili kavramlar, beşinci kısımda ise temel bileşenler analizi konusunda bazı temel teorik detaylar verilmiştir.

Uygulama alanında ise binaların ısı yüklenme kapasitelerini etkileyen 8 değişkenli bir işlemden elde edilen veriler incelenmiştir. Değişkenlerin titiz incelenmesi sonucunda ısı yükleme kapasinde en etkin olan 5 değişkenin analiz işlminde kullanılmasına karar verilmiştir. Elde edilen temel bileşenler kullanılarak etkin olduğu varsayılan 5 değişkenin tahmini yapılmıştır. Tahmin sonuçlarının geçerli olduğu tahmin hatalarının küçüklüğü ile kanıtlanmıştır.

Anahtar Kelimeler: Matris cebiri, regresyon analizi, tahmin, tahmin edici, tahmin edilen, regresyon katsayıları, temel bileşenler analizi, özvektör, özdeğer.

DEDICATION

I am dedicating my thesis to my family, my mother Hoda, my father's soul Winis, and my husband Tarek, also to my supervisor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\mathbf{A}^{-1}$	The Inverse of the Matrix $\mathbf{A}$
$\mathbf{A}'$	The Transpose of the Matrix $\mathbf{A}$
$E(Y\|X_i)$	A Linear Equation for the Response Y Using the Predictors X_i
e	Residual error
$\mathbf{e}$	Eigenvector
MSE	Mean Square of Error
PCs	Principal Components
PCA	Principal Components Analysis
$\mathbf{R}$	Sample Correlation Matrix
r_{Y,X_i}	Correlation between the Response Y and the Predictor X_i
RMSD	Root Mean Square Deviation
$\mathbf{S}$	Sample Covariance Matrix
s^2	Single Sample Variance
$\mathbf{x}$	Sampling Data Vector
$\bar{x}$	Sample Mean
$\bar{\mathbf{x}}$	Mean Data Vector
$\mathbf{X}$	Data Matrix
$\mathbf{\Sigma}$	Population Covariance Matrix
ρ	Population Correlation Coefficient
λ	Eigenvalue
ρ_{Y_i,X_j}	Correlation between Two Variables

μ	Population Mean
$\boldsymbol{\mu}$	Population Mean Vector
σ^2	Population Variance
$\sigma_{X_i X_j}$	Sample Covariance
β_i	Regression Coefficient
$\boldsymbol{\varepsilon}$	Random Error

Chapter 1

INTRODUCTION

In statistical analysis, regression has an important place. It involves a set of processes that establishes the relationship between dependent (response) and independent (predictor) variables. Regression analysis includes some techniques for modeling several variables, these techniques work on the relationship between dependent variable/s and one or more independent variables or predictors.

More specifically, regression analysis helps to realize how the value of the dependent variable changes if any one of the independent variables is varied, while the other predictors may be fixed.

In case the number of variables to be dealt with is very large, a technique known as Principal Component Analysis (PCA) can be used to reduce the number of variables to manageable levels, without much loss of the inherent characteristic of the process under study. The reduced number of variables or principal components that are made of linear combinations of the original variables can be used to obtain some idea about the process. They can also be associated with the original variables via linear regression.

This thesis deals with the multiple linear regression. Explains important parts of the multivariate regression, and explains how estimation is carried out, its accuracy and

associated errors. It also gives a short review of the theory of PCA and attempts to integrate the PCA results to multivariate regression.

Linear algebra plays a main role in multivariate statistics. Both in multivariate regression and the PCA application of algebraic concepts simplifies, otherwise tedious numeric computations. Hence a review of the matrix algebra is included in Chapter 3. Intense numeric computations necessitate the use of software such as SPSS and Matlab which are used in all computations in this thesis.

Since the regression analysis is based on the relationship between dependent and independent variables, the correlation coefficient between the variables is a necessity to determine the value of the regression analysis to be undertaken.

Theory related to linear simple and multivariate cases are explained in detail under Chapter 4. Derivation of the system that leads to the estimation of the regression parameters, and theorems proving some important properties of the regression process are given. The importance of the correlation coefficient between the response and explanatory variables is highlighted.

Principal components analysis is focused on explaining the variance-covariance structure of a set of variables by some linear combinations of these variables. The first aim of the principle components is dimension reduction and interpretation of the new variables named as principal components. Chapter 5 explains the theory and ideas behind this matter.

In Chapter 6 the analysis of a multivariate data with a response and 8 independent variables are analyzed using multivariate regression and PCA. The relation between the ratio of each principle component of total variation in the data, and the correlation coefficient and sum squares of error in linear regression are investigated The obtained results are summarized using tables and graphs, and interpreted.

Chapter 2

LITERATURE REVIEW

Work that relates to the technique of regression analysis can be traced back to the nineteenth century. The English mathematician Galton collected data of heights of individuals and the heights of their parents. After organizing frequency tables that classified these individuals, both by their height and by the average height of their parents, Galton came to the conclusion that tall parents usually had tall children and short parents usually had short children. He reached to the conclusion that the heights of children can be predicted by the heights of their parents [1].

In 1805 the idea of least-squares analysis was proposed by the French mathematician Adrien-Marie Legendre [2]. Similar ideas were also proposed by the American mathematician Robert Adrain in 1808, and by Gauss in 1809 [3]. In 1821 the theory of least squares was published by Guass. In addition, a version of the Gauss–Markov theorem was included with least of squares [4]. Udny Yule and Karl Pearson extended the work of Galton [1] and found that the response and explanatory variables represent the joint distribution that is assumed to be Gaussian [5] [6].

In 1922 and 1954 Fisher assumed that the conditional distribution of the response variable is Gaussian [7] [8], but the joint distribution was not. At this point the assumption of Fisher is close to the theory of Gauss in 1821 [4]. In the 1950s and 1960s electromechanical desk calculators were used to compute the regression

parameters. In the early years of electronic computation, a long time (as long as 24 hours) was needed to carry out simple regression analysis [9].

The problem of fitting a line when both *X* and *Y* variables have an error in measurement was mentioned by Karl Pearson. His solution to the regression problem in the presence of errors is named "major axis" of ellipse of data. *X* and *Y* co-variation was taken into account [10]. When the units of the X and Y variables were changed the major axis was not uniquely determined: The slope and intercept would differ if the scales are changed. That was mentioned by Kermack and Haldane and suggested the use of "reduced major axis" in case both *X* and *Y* were transformed to standardized variables where the mean $\mu = 0$ and the standard deviation $\sigma = 1$ [11].

A method of weighing was developed by York [12]. This method was called the least squares cubic because the slope of the regression equation needs the solution of a cubic equation to determine the regression parameters. The geometric mean regression and the formulae for the asymmetrical confidence limits for the slope of the geometric mean regression was shown by Ricker [13].

Sprint and Dolby gave some comments about the geometric mean regression, especially when the selection of the response and regressor variables is problematic. If the line splits the minor angle between the two models, *Y*-on-*X* and *X*-on-*Y,* this line is assumed as the least squares bisector While the geometric mean regression and the least squares bisector may have a small difference in slope that is probably not statistically significant [14].

Sokal and Rohlf wrote in great details a textbook is named “Biometry” about the issues of model-I (where *Y* is random and assumed to depend on *X* that can be random or fixed) versus model-II regression. Laws presented an extensive chapter about the techniques of model-II regression [15].

Regression analysis continues to be an area of active research. In recent decades, new methods have been developed for regression, regression involving correlated responses such as time series and growth curves, regression in which the predictor (independent variable) or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, nonparametric regression, Bayesian methods for regression, regression in which the predictor variables are measured with an error.

Principal component analysis (PCA) is essential in multivariate data analysis field. Karl Pearson proposed the principle components analysis in 1901 as an analogue of the principal axis theorem in mechanics [16]. In 1923 Fisher and MacKenzie mentioned PCA as more suitable than analysis of variance for the modelling of the response in regression analysis [17]. Harold Hotelling developed (PCA) independently and named it in the 1930s depending on the field of application [18].

In 1966 Fisher and MacKen- zie also outlined the NIPALS algorithm (Nonlinear estimation by iterative least squares) [19]. It is the most commonly used methods for calculating the principal components of a data set. It gives more accurate results when compared with the SVD of the covariance matrix [21] [22], but is slower in computational aspects.

Starting in late 1970s PCA started taking place in many textbooks on statistical data analysis [20]. In his book Jolliffe (2002) discusses the differences between PCA and factor analysis [23]. In most cases (PCA) is used as a tool in exploratory data analysis and for providing predictive models.

The terms of component scores or factor scores were usually used to discuss the results of the PCA. The variable values were transformed corresponding to a particular data point. Each standardized original variable should be multiplied to load the weight and get the component score [24].

PCA has a connection with canonical correlation analysis that is called (CCA). CCA defines coordinate systems that show the cross-covariance between two datasets whilst PCA defines a new orthogonal coordinate system that shows variance in a single dataset. [25] [26].

Chapter 3

MATRIX ALGEBRA APPLIED IN STATISTICS

The establishment of a relationship between the multivariate regression results, with those obtained after dimension reduction using principal component analysis is a challenging problem. While attempting to establish such a relationship, mathematical tools such as matrix algebra will simplify the otherwise long set of computations. Hence a review of matrix algebra as needed in this study is given.

3.1 Matrix Algebra and Statistics

When large volumes of data are to be processed, computer software packages such as Matlab and Mathematica are extensively used for the analysis of such data. The use of matrix operations in the design of algorithms for the analysis of such data is essential.

3.1.1 Definition of Vector and Matrix

Definition 3.1. The vector is an array of *n* real elements $x_1, x_2, \ldots, x_n$. Statistically, it represents the data values for a variable and is written as

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

Its transpose is $\mathbf{x}'_{1\times n} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$.

Definition 3.2. A matrix is made up of vectors of the same size, hence if *p* vectors of the same size are used the rectangular matrix of size $n \times p$ is obtained. It means the

matrix has *n* rows and *p* columns. Statistically it is very convenient to represent a multivariate data set using a rectangular matrix $\mathbf{X}_{n\times p}$.

$$\mathbf{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}=\begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \end{bmatrix}$$

3.1.2 Some Important Statistical Measures and their Representation by Vectors and Matrices

a) The sample mean

Given a set of observations $x_1, x_2, \ldots, x_n$, the sample mean is defined by

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

When *p* variables with *n* observations are available, then the mean vector $\bar{\mathbf{x}}$ can be represented in the given form

$$\bar{\mathbf{x}}=\begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_i \\ \vdots \\ \bar{x}_p \end{bmatrix}=\frac{1}{n}\mathbf{1}'\begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1j} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2j} & \ldots & x_{2p} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ x_{i1} & x_{i2} & \ldots & x_{ij} & \ldots & x_{ip} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nj} & \ldots & x_{np} \end{bmatrix}$$ Where $\mathbf{1}_{n\times 1}$ is the vector of 1s.

b) Sample variance/covariance and variance-covariance matrix

Theoretically the expectation of the squared deviation of a random variable from its mean is called variance and defined as $\sigma^2=\mathrm{E}\left[\left(x-\mu\right)^2\right]$. The single sample variance is given by $s^2=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n-1}$.

When there is more than one variable in question, the covariation between pairs of variables is to be considered. In probability theory the covariance between *p* pairs of random variables is defined as:

$Cov(X_i, X_j) = \sigma_{X_i X_j} = E[(X_i - \mu_{X_i})(X_j - \mu_{X_j})]$ Where $i, j = 1, 2, \ldots, p.$

It can be shown that $-\infty < \sigma_{ij} < \infty$ when $i \neq j,$ and $\sigma_{ij} \geq 0$ When $i = j$

When *i=j* $Cov(X_i, X_i) = \sigma_{X_i X_i} = \sigma^2_{X_i} = E[(X_i - \mu_{X_i})(X_i - \mu_{X_i})] = E[(X_i - \mu_{X_i})^2].$

$$\sum = E[(X - \mu)(X - \mu)']$$

$$= E\left(\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_p - \mu_p \end{bmatrix} \begin{bmatrix} x_1 - \mu_1, x_2 - \mu_2, \ldots, x_p - \mu_p \end{bmatrix} \right)$$

$$= E\begin{bmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) & \ldots & (X_1 - \mu_1)(X_P - \mu_P) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 & \ldots & (X_2 - \mu_2)(X_P - \mu_P) \\ \vdots & \vdots & \ddots & \vdots \\ (X_P - \mu_P)(X_1 - \mu_1) & (X_P - \mu_P)(X_2 - \mu_2) & \ldots & (X_P - \mu_P)^2 \end{bmatrix}$$

The result of the covariance computations can be expressed as the covariance matrix given below

$$\boldsymbol{\sigma}_{X_i X_j} = \boldsymbol{\sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

This is a diagonal and symmetric matrix with elements of the diagonal representing the variances of the particular variable, i.e. $\sigma_{ii} = \sigma_i^2.$ It can also be shown that when two random variables X_i, X_j are independent, $\sigma_{X_i X_j} = 0,$ but the vice versa case is not always true.

Statistically the covariance between *p* pairs of variables with *n* observations is given by

$$\mathrm{S}_{X_i X_j} = \frac{\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_i\right)\left(x_{kj}-\bar{x}_j\right)}{n-1} \text{ where } i,j=1,2,\ldots,p$$

Variance–covariance matrix or covariance matrix is a matrix where the elements in the *j, k* position are the covariances $\left(s_{jk}\right)$ between the j^{th} and k^{th} elements of a random vector, and the elements in a $j\times j$ position are the variances $\left(s_{jj}\right)$. Variance-covariance matrix maps a linear operator *c* of the random variables X onto a vector that is consisted of covariances for those variables.

$$\mathbf{S}=\begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1p} \\ s_{21} & s_{22} & \ldots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \ldots & s_{pp} \end{bmatrix}$$

$s_{jk}=s_{kj}$, since the covariance matrix is symmetric. In addition, it is positive-semidefinite matrix which means that, the eigenvalues of this matrix are either positive or zero

c) Correlation coefficient and correlation matrix

Correlation coefficient plays an important role in regression analysis as it is used to measure the strength of the relationship between the variables. The range of ρ is located in the interval [-1:1]. ρ *A* value close to -1 or 1 indicates the very strong linear relation between the variables. As ρ approaches 0 the strength of linear correlation diminishes and ρ = 0 means no linear correlation between the two variables.

The coefficient of determination ρ^2 is the squared value of ρ and it indicates the rate of the total variation in the values of the dependent variable y that can be calculated by a linear relationship with the values of the random variable x

The population correlation matrix is denoted as $\boldsymbol{\rho}$

$$\boldsymbol{\rho} = \frac{\sigma_{x_i x_j}}{\sqrt{\sigma_{x_i x_i}\sigma_{x_j x_j}}}$$

Where $\sigma_{x_i x_j}$ is the covariance between the variables X_i and X_j, while $\sigma_{x_i x_i}$ and $\sigma_{x_j x_j}$ are the variances of the i^{th} and j^{th} variables.

The amount of a linear association between the random variables $\left(X_i, X_j\right)$ is measured by the elements of the matrix $\boldsymbol{\rho}$.

Let the population correlation matrix be a $p \times p$ symmetric matrix.

$$\boldsymbol{\rho} = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix}$$

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}$$

The diagonal elements of the population correlation matrix indicate the correlations between the random variables with themselves, hence are $\rho_{ii} = 1.$ The off-diagonal

elements of $\boldsymbol{\rho}$ indicate the correlations between $\left(X_i, X_j\right)$where $i \neq j$. the matrix $\boldsymbol{\rho}$ is symmetric, i.e. $\rho_{ij} = \rho_{ji}$.

The sample correlation matrix is denoted by $\mathbf{R}$. It is obtained an analogous to the covariance matrix with correlations in place of covariances. Linear correlation between the ith and jth variables is computed by

$$r_{ij} = \frac{s_{x_i x_j}}{\sqrt{s_{x_i x_i} s_{x_j x_j}}} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}_i\right)\left(x_j - \bar{x}_j\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}_i\right)^2\left(x_j - \bar{x}_j\right)^2}}$$

$$R = \left(r_{jk}\right) = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

Example 3.1

A data set consisting of 53 observations for each of the 5 variables is used for the computation and explanation of the relationship between the variables.

X_1: Average of death per 1000 residents.

X_2: Availability of doctors per 100,000 residents.

X_3: Hospitals availability per 100,000 residents.

X_4: Income per person each year in thousand dollars.

X_5: Population density per square mile.

Data are given in Appendix, Table 3.1 (row data residents, income and density)

Mean for each variable:

$\bar{x}_1$ =9.3057, $\bar{x}_2$= 116.0943, $\bar{x}_3$=589.7925, $\bar{x}_4$=9.4358, $\bar{x}_5$= 110.6415.

Standard deviation for each variable:

s_1=1.6626 s_2=37.8866 s_3= 332.6183 s_4=1.0754 s_5=47.1797.

Then to compare the coefficient of variation, ($s/\bar{x}$) is calculated, the respective coefficients of the variation for the variables x_1, x_2, x_3, x_4, x_5 are 0.18, 0.33, 0.56, 0.11and 0.43. In general, the coefficient variation gives an idea about how the set of data points is distributed around the mean or it shows the measurement of the variation in points from the mean.

For instance, the coefficient variation for the first variable CV_1= 0.18 that means the average of death per 1000 residents has variation 18% from its mean ($\bar{x}_1$ =9.3057), and the same for the other variables. As a result the experimenter would choose CV_4 that indicates the income per person each year in thousand dollars because it gives the lowest variation (0.11) from its mean ($\bar{x}_4$=9.4358) so that the minimize risk is the best to be chosen. The correlation matrix for the same data is computed as:

$$\mathbf{R} = \begin{bmatrix} 1 & 0.1158 & 0.1106 & -0.1720 & -0.2776 \\ 0.1158 & 1 & 0.2956 & 0.4333 & -0.0199 \\ 0.1106 & 0.2956 & 1 & 0.0275 & 0.1866 \\ -0.1720 & 0.4333 & 0.0275 & 1 & 0.1287 \\ -0.2776 & -0.0199 & 0.1866 & 0.1287 & 1 \end{bmatrix}$$

The highest correlation between x_2 and x_4 $r_{2,4}$ = 0.4333 that means there is a nearly strong positive relationship between the availability of the doctors and the rate of the income per person each year. The lowest correlation between x_2 and x_5 $r_{2,5} = -0.0199$ that means almost there is no relationship between the doctors and the availability of the population density per square mile.

The covariance matrix is obtained

$$\mathbf{S}=\begin{bmatrix} 2.7640 & 7.2917 & 61.1550 & -0.3075 & -21.7749 \\ 7.2918 & 1435.4 & 3725.4 & 17.6542 & -35.6386 \\ 61.1550 & 3725.4 & 110630 & 9.8383 & 2928.5 \\ -0.3075 & 17.7645 & 9.8383 & 1.1566 & 6.5323 \\ -21.7749 & -35.6386 & 2928.5 & 6.5323 & 2225.9 \end{bmatrix}$$

From the previous matrix, the variances between the variables with themselves are located in diagonal elements. For instance, s_{11}=2.7640. The off-diagonal elements indicate the covariances between the variables $\left(x_i, x_j\right)$ where $i \neq j$. Remembering the fact that the magnitude of the covariance value does not indicate the strength of the relationship between the concerned variables, is evident in the **S** matrix. For example the variables with the highest linear correlation $r_{2,4}$=0.4333 does not record the highest covariance value which is evident from the **S** matrix.

3.2 Using Matrices in Statistics

3.2.1 Regression Analysis

Using matrices in regression analysis is essential to find the linear regression model in case the number of independent variables exceeds two. Matrix theory can simplify the mathematical operations in estimating the parameters of the multiple regression model.

3.2.2 General Linear Model by Using Matrices

The general linear model for a process governed by *k* variables is defined as

$$Y=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots+\beta_k x_{ki}+\varepsilon_i;\ i=1,\ 2,\ldots,n \qquad (3.1)$$

This regression model represents *n* equations, showing the dependent variable's (response) relation to the independent variables via the coefficients $\beta_j;\ j=1,2,\ldots,k,$

Then, the linear system given in equation (3.1) can be written in matrix format as

$$\mathbf{y} = \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ where}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

The vector $\mathbf{y}$ represents the dependent variables $y_i;\ i = 1,2,\dots,n.$ Then any

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

The matrix $\mathbf{X}$ must have 1's in the first column, the other columns representing the observations for the independent variables,

The vector $\boldsymbol{\beta}$ indicates the regression coefficients.

The vector $\boldsymbol{\varepsilon}$ indicates the random errors.

Using the data collected in the process under study, the coefficients $b_0, b_1, \dots, b_k$ are used to estimate the true but unknown coefficients $\boldsymbol{\beta}$. In doing so, it is desired to minimize the sum of the square of the errors between the estimated value of the response variables and their true values. The error can be expressed as

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \dots - b_k x_{ki})^2$$

In matrix format the SSE becomes

$$SSE = (\mathbf{y} - \mathbf{xb})'(\mathbf{y} - \mathbf{xb}) = (\mathbf{y}' - (\mathbf{xb})')(\mathbf{y} - \mathbf{xb})$$
$$= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{xb} - (\mathbf{xb})'\mathbf{y} + (\mathbf{xb})'\mathbf{xb} = \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{x}'\mathbf{y} + \mathbf{b}'\mathbf{x}'\mathbf{xb}$$

$$\mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{x}'\mathbf{y} + \mathbf{b}'\mathbf{x}'\mathbf{xb} = 0$$

$$-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{Xb} = 0$$
$$-\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{Xb} = 0$$
$$\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{y}$$

Then the result reduces to the solution of the parameters b in the equation

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$$

Let $\mathbf{A} = \mathbf{X}'\mathbf{X}$ and $\mathbf{g} = \mathbf{X}'\mathbf{y}$, then

$$\mathbf{A} = \mathbf{x}'\mathbf{x} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x^2_{1i} & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \sum_{i=1}^{n} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n} x^2_{ki} \end{bmatrix}$$

$$\mathbf{g} = \mathbf{X}'\mathbf{y} = \begin{bmatrix} g_0 = \sum_{i=1}^{n} y_i \\ g_1 = \sum_{i=1}^{n} x_{1i}y_i \\ \vdots \\ g_k = \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix}$$

Then the estimators $\boldsymbol{b}$ of $\boldsymbol{\beta}$ are calculated from: $\mathbf{b} = \mathbf{A}^{-1}\mathbf{g} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$

3.2.3 Some Properties of the Least Squares Method by Using Matrices

We obtain the estimators $(b_0, b_1, \ldots, b_k)$ under the assumption on the random errors $(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_k)$, if we assume those errors are distributed with mean 0 and variance σ^2, Then it is possible to show the parameters $(b_0, b_1, \ldots, b_k)$ as unbiased estimators to $(\beta_0, \beta_1, \ldots, \beta_k)$ respectively.

The matrix $\mathbf{A}^{-1}\sigma^2$ represents the variance-covariance matrix of the estimators, where the main diagonal elements display the variances of the estimators $(b_0, b_1, \ldots, b_k)$ and the covariances of the off-diagonal elements. In case we have *k* independent variables (predictors), The inverse of the matrix $\mathbf{A}$ gives

$$\mathbf{A} = (\mathbf{X}'\mathbf{X}) = \begin{bmatrix} c_{00} & c_{01} & \cdots & c_{0k} \\ c_{10} & c_{11} & \cdots & c_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k0} & c_{k1} & \cdots & c_{kk} \end{bmatrix}$$

Diagonal and off-diagonal elements can be defined as

$$\sigma_{bi}^2 = c_{ii}\sigma^2, i = 0,1,2,\ldots,k$$
$$\sigma_{bibj} = Cov(b_i, b_j) = c_{ij}\sigma^2, i \neq j.$$

S^2 Is unbiased estimator of σ^2 that can be obtained by : $s^2 = \frac{SSE}{n-k-1}, where$

$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. Where $\hat{y}_i = b_0 + b_1 x_1 + \ldots + b_k x_k$.

3.2.4 Principal Components Analysis

Principal components analysis (PCA) is the method used in dimension reduction for data sets with a large number of variables. This reduction is performed by employing various techniques provided by matrix algebra. Particularly correlation or variance-covariance matrix is used in determining the principal components. Eigenvalues and eigenvectors of these squares and symmetric matrices are computed and used in the process.

3.2.5 Principal Components Analysis by Using Variance - Covariance Matrix

Let the random vector $\mathbf{x} = [x_1, x_2, \ldots, x_p]$ represents a dataset with *p* variables, having a variance-covariance matrix denoted by $\boldsymbol{\Sigma}$. The eigenvalue **-** eigenvector pairs of $\boldsymbol{\Sigma}$ are $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. The principal components $(Y_1, Y_2, \ldots, Y_P)$ are then obtained as linear combinations initial variables as

$$Y_1 = \mathbf{e}_1' X = e_{11}X_1 + e_{12}X_2 + \ldots + e_{1p}X_P$$
$$Y_2 = \mathbf{e}_2' X = e_{21}X_1 + e_{22}X_2 + \ldots + e_{2p}X_P$$
$$\vdots$$
$$Y_P = \mathbf{e}_p' X = e_{p1}X_1 + e_{p2}X_2 + \ldots + e_{pp}X_P$$

Then the variance and covariance of a principal component are computed as

$$Var(Y_i) = \mathbf{e_i'}\boldsymbol{\Sigma}\mathbf{e_i} = \lambda_i, i = 1,2,\ldots,p$$
$$Cov(Y_i, Y_k) = \mathbf{e_i'}\boldsymbol{\Sigma}\mathbf{e}_k = 0, i \neq k$$

Chapter 4

REGRESSION ANALYSIS

4.1 General Linear Regression

In statistics, the general linear model (GLM) provides the flexibility of generalizing the ordinary linear regression. This is achieved by permitting the linear model to be associated with the response variable via a link function and by permitting the magnitude of the variance of every measurement to be a function of its expected value.

Regression analysis is a set of statistical operations to estimate the relationships between the dependent and independent variables. However, this should not be taken as to mean that the dependent variable is caused by the independent variable/s. The linear regression tries to calculate the intercept and slope/s of the linear line/s by minimizing the sum of the squares of the errors between each observation and its corresponding estimated value via the regression equation. If the scatter diagram of the data does not exhibit some kind of linearity, then the nonlinear regression model can be used. The independent variable/s can be either continuous or categorical.

4.1.1 Some Assumptions in Linear Regression

1. Linearity

The concept of linearity means that the mean of the dependent (response) variable can be written as a linear combination of the independent variables their coefficients being computed by the regression model.

2. Constant variance

This assumption leads to equality of the variances in the errors of the values of the dependent variable.

3. Independence of errors

This assumes that the errors of the dependent variables have no correlation with each other.

4.1.2 General Linear Regression Model

Let $x_1, x_2, \ldots, x_k$ be predictor variables and $y_1, y_2, \ldots, y_m$ response variables, then we have the following general linear model

$$
\begin{aligned}
Y_1 &= \beta_{01} + \beta_{11}x_1 + \ldots + \beta_{k1}x_k + \varepsilon_1 \\
Y_2 &= \beta_{02} + \beta_{12}x_1 + \ldots + \beta_{k2}x_k + \varepsilon_2 \\
\vdots \quad &= \qquad \vdots \qquad\qquad \vdots \\
Y_m &= \beta_{0m} + \beta_{1m}x_1 + \ldots + \beta_{km}x_k + \varepsilon_m
\end{aligned}
$$

where: $\beta_0, \beta_1, \ldots, \beta_k$ are regression coefficients, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$ are random errors.

4.1.3 Cases Where There is a Demand to Use the Linear Regression Analysis

There are many applications that fall into one of the following cases:

1.In case the target is prediction, forecasting or error reduction.

2.If the goal is to explain the variation in the dependent variable (response).

3.To measure the strength of the relationship between the response/s and predictor/s.

4.1.4 Three Types of Linear Regression Analysis

There are three types of the linear regression according to the number of variables (dependent and independent variable/s)

1. Simple linear regression: It includes one dependent variable that is called response and one independent variable that is called regressor.
2. Multiple linear regression: One dependent (response) variable and several independent (regressor) variables.

3. Multivariate multiple linear regression: Several dependent variables and several independent variables.

4.2 Simple Linear Regression

Definition: It is a statistical approach modeling the linear relationship between the response (*y*) and predictor (*x*) variables. Determination of the unknown parameters that establishes the linear function is crucial. However, simple linear regression can only determine the linear relationship between the response and one predictor variable.

4.2.1 Simple Linear Regression Model

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

where β_0 is the intercept, β_1 is the slope, ε is the random error, random error is assumed to be normally distributed with, $E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2$, where β, σ^2 are unknown parameters, σ^2 is called residual variance, ε is called random error that has constant variance. The existence of the random error in the system ensures that the model is not a simple deterministic equation, but has a distribution function that is associated with. Figure 4.1 illustrates the relationship between the predictor and the response variables.

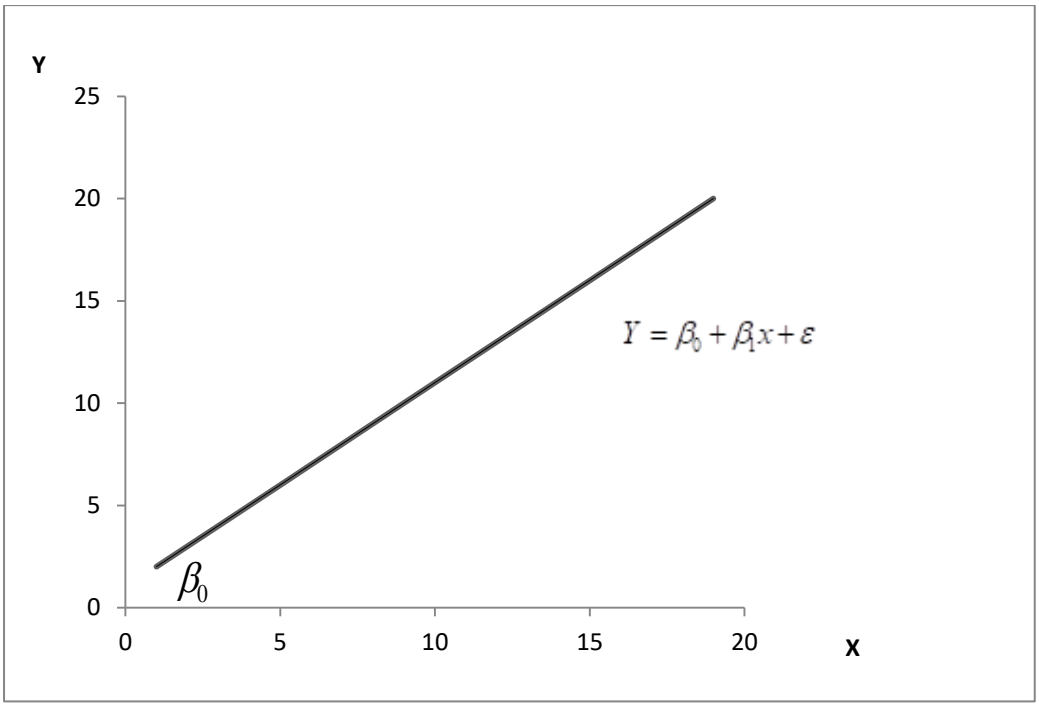

Figure 4.1: β_0 is the Intercept, β_1 is the Slope

4.2.2 Estimated Regression Model

While the true but unknown simple linear regression model is given by $Y = \beta_0 + \beta_1 x + \varepsilon,$ its estimator is

$y = b_0 + b_1 x_i + e_i,$

where $b_0,\ b_1$ are regression coefficients, e_i is the residual error. Using data collected from a certain process, then the estimated value of y and coefficients will be $\hat{y},\ \hat{b}_0,\ \hat{b}_1.$ Then, $\hat{y} = \hat{b}_0 + \hat{b}_1 x$ can be written. As the residual error $e_i = y_i - \hat{y}_i, i = 1, 2, \ldots, n$ has mean zero, it is omitted.

The error ε_i also called total error is made up of two components. Namely the explained error e_i and unexplained or random parts that cannot be explained by the regression equation obtained from the data. When the squares of these are considered

Sum of the squares of the explained errors: $\sum_{i=1}^{n} e_i^2 = SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

Sum of the squares of the unexplained errors: $SSR = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)$.

Sum of the squares of the total errors: $SST = \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2$.

Then the following can be written

$$SST = \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 = SSE + SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2$$

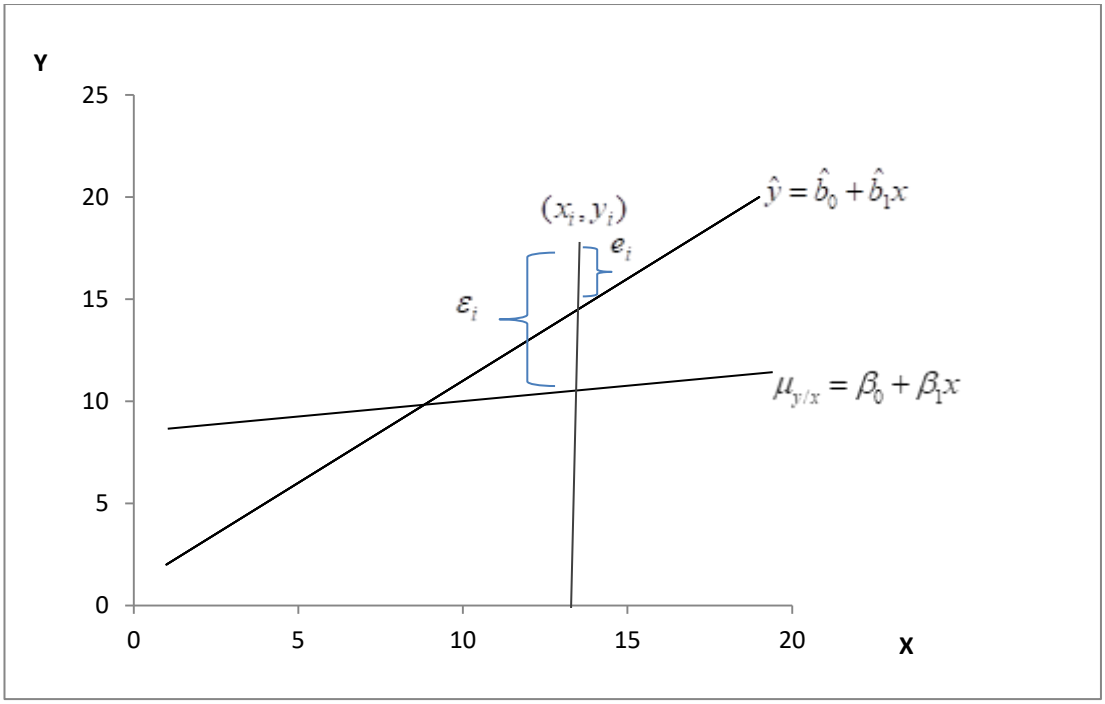

Figure 4.2: Illustrates the Difference between the Random and Residual Errors

4.2.3 The Method of Least Squares

Theorem 4.1

The linear equation to be obtained from available data must ensure that SSE value is minimized $Min\ SSE = Min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

Proof

$$\mu_{y/x} = \beta_0 + \beta_1 x$$
$$y_i = \mu_{y/x} + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1,2,3,\ldots,n$$
$$\varepsilon_i = y_i - \mu_{y/x} = y_i - \beta_0 - \beta_1 x_i$$
$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$
$$\frac{\partial(\text{SSE})}{\partial \beta_0} = -2\sum_{I=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right) = 0$$
$$\frac{\partial(\text{SSE})}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i \left(y_i - \beta_0 - \beta_1 x_i \right) = 0$$

Then

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$
$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad (4.1.1)$$

The equations (4.4.1) are called the normal equations. Solution of the normal equations gives the $\hat{\beta}_0, \hat{\beta}_1$ values that minimize SSE.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \qquad (4.1.2)$$

S_{xy} And S_{xx} can be rewritten as follows

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} \qquad (4.1.3)$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\sum_{i=1}^{n} x_i^2}{n} = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 \qquad (4.1.4)$$

Then

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (4.1.5)$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \; i = 1,2,\ldots,n \qquad (4.1.6)$$

Also the residual error can be found for each value of the estimator *y* by

$$e_i = y_i - \hat{y}_i, i = 1,2,\ldots,n \qquad (4.1.7)$$

Example 4.1

The following data given in Table 4.1 represents the height in cm and weight in kg of 12 people.

Table 4.1: Height, Weight Data for 12 Randomly Selected People.

	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{Y}_1$	150	150	150	155	155	155	155	160	160	175	175	175
$\hat{Y}_2$	50	61	54	54	63	59	61	68	65	77	82	72

The following will be done:

- Determination of the regression coefficients $\left(\hat{\beta}_0, \hat{\beta}_1\right) = \left(b_0, b_1\right)$ respectively
- The equation of the regression line.
- Draw the line on a scatter plot.

For the calculation of the parameters and simple linear regression equation see Appendix B, table 4.2

$\bar{x} = 63.92$, $\bar{y} = 159.58$.

We take the results from the Table (4.2), in Appendix and apply the following equations to determine the regression equation.

$$S_{xy} = 123365 - \frac{(1915)(767)}{12} = 964.583 \qquad (4.1.3)$$

$$S_{xx} = 306675 - \frac{(1915)^2}{12} = 1072.917 \qquad (4.1.4)$$

$$\hat{\beta}_1 = \frac{964.583}{1072.917} = 0.899 \qquad (4.1.2)$$

$$\hat{\beta}_0 = 63.92 - (0.899)(159.58) = -79.6 \qquad (4.1.5)$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$
$$\hat{y} = -79.6 + 0.899x \qquad (4.1.6)$$

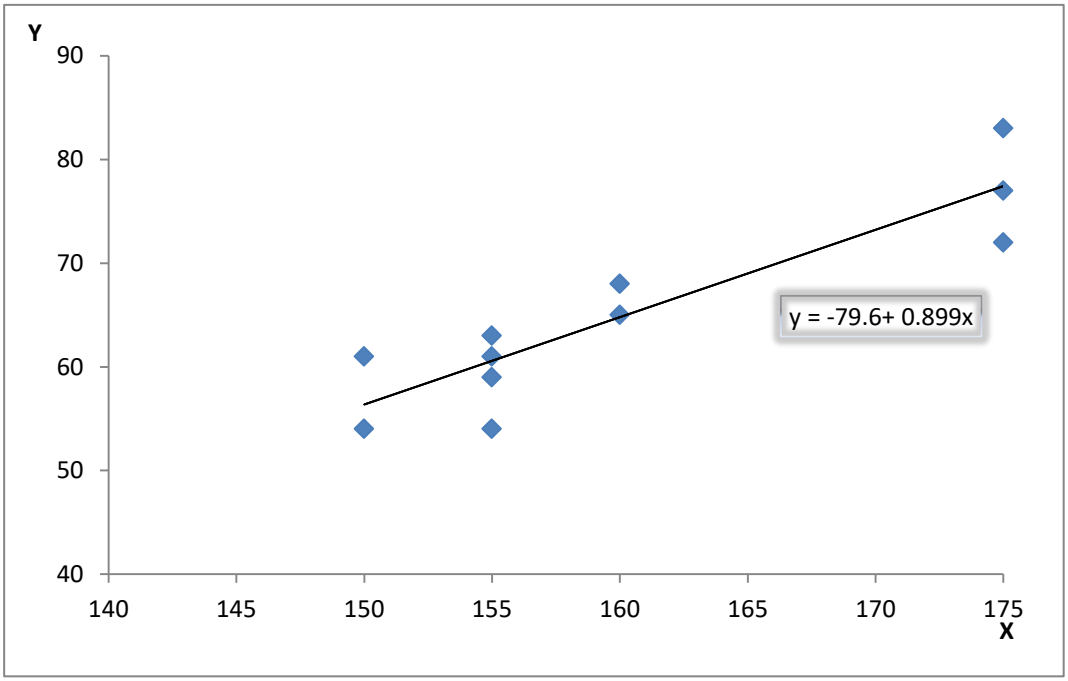

Figure 4.3: Scatter Plot for Data

4.2.4 Mean and Variance of Estimators

Theorem 4.2

Mean and variance of the estimators $\hat{\beta}_0, \hat{\beta}_1$ are

$$\mu\hat{\beta}_1 = \sum_{i=1}^{n} c_i \mu_i = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\sigma^2 \hat{\beta}_1 = \sum_{i=1}^{n} c_i^2 \sigma_i^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma_{Y_i}^2}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \sigma_{Y_i}^2}{\left[\sum_{i=1}^{n} (x_i - \bar{x})\right]^2}$$

Proof

If we have $x_1, x_2, \ldots, x_n$ with a normal distribution with means $\mu_1, \mu_2, \ldots \mu_n$ and variances, $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ respectively, then the random variable $Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n$ which is a linear combination of the random variables X_i has a normal distribution with, Mean, $\mu_Y = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n$

And variance, $\sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots + a_n^2\sigma_n^2$

Since the estimator

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} c_i Y_i$$

$$\text{Where } c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \text{i=1,2,...,}n$$

Based on the properties of linear combination of random variables, we can write

$$\text{Since } \beta_1 = c_1y_1 + c_2y_2 + \ldots + c_ny_n = \sum_{i=1}^{n} c_i Y_i$$

And since *Y* follows the normal distribution

$$\mu\beta_1 = c_1\mu_1 + c_2\mu_2 + \ldots + c_n\mu_n = \sum_{i=1}^{n} c_i\mu_i$$
$$\sigma^2\beta_1 = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \ldots + c_n^2\sigma_n^2 = \sum_{i=1}^{n} c_i^2\partial_i^2$$

Then $\hat{\beta}_1$ has $\mathrm{N}\left(\mu\hat{\beta}_1, \sigma^2\hat{\beta}_1\right)$, distribution parameters being

$$\mu\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(\beta_0 + \beta_1 x_i\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)}; \ \sigma^2\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\sigma_{Y_i}^2}{\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\right]^2}.$$

Following the same logic, it can be shown that β_0 is normally distributed with parameters

$$\mu\hat{\beta}_0 = \beta_0, \text{ and } \sigma^2\hat{\beta}_0 = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sigma^2$$

4.2.5 Correlation Coefficient

This is denoted by ρ and in simple linear regression it is used to measure the strength of the linear relationship between the response and the predictor variables. It also shows whether the relationship is positive or negative linear one.

Sample correlation coefficient r can be found using the equation (3.10).

4.3 Multiple Linear Regression

Multiple linear regression analysis applies when there is more than one explanatory variable and one continuous dependent variable. Each response variable in the linear regression equation has its own equation with coefficients β_i.

4.3.1 The Classical Model of Multiple Linear Regression

The general linear model is defined in (3.12).

In matrix format $\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times(k+1)} + \boldsymbol{\beta}_{(k+1)\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$,

where $\boldsymbol{\beta}$ is the regression coefficients vector, $\mathbf{y}$ represent the observations of the response variable, $\boldsymbol{\varepsilon}$ is the vector residuals.

In general, it is desired to have $(n > k)$.

4.3.2 Estimation of the Regression Coefficients (Method of Least Squares)

As we have mentioned before in chapter 3, it is not an easy job to estimate the parameters $b_0, b_1, \ldots, b_k$ in multiple linear regression by using the method of least squares (4.2.3), which gives *k* number of equations as following

$$
\begin{aligned}
& nb_0 + b_1\sum_{i=1}^{n} x_{1i} + b_2\sum_{i=1}^{n} x_{2i} + \ldots + b_k\sum_{i=1}^{n} x_{ki} = \sum_{i=1}^{n} y_i \\
& b_0\sum_{i=1}^{n} x_{1i} + b_1\sum_{i=1}^{n} x_{1i}^2 + b_2\sum_{i=1}^{n} x_{1i}x_{2i} + \ldots + b_1\sum_{i=1}^{n} x_{1i}x_{ki} = \sum_{i=1}^{n} x_{1i}y_i \\
& \vdots \\
& b_0\sum_{i=1}^{n} x_{ki} + b_1\sum_{i=1}^{n} x_{ki}x_{1i} + b_2\sum_{i=1}^{n} x_{ki}x_{2i} + \ldots + b_k\sum_{i=1}^{n} x_{ki}^2 + \sum_{i=1}^{n} x_{ki}y_i
\end{aligned}
$$

To estimate the regression coefficients, the linear equations (4.3) have to be solved simultaneously. For a large number of variables the solution of the system becomes difficult. Hence the use of matrix algebra makes the solution easier, as it is explained in the following section.

4.3.3 Least Square Estimation (by Using Matrices)

The purpose is to determine the values of $\boldsymbol{\beta}$ and error variance $\boldsymbol{\sigma}^2$ from available data. Assume $\mathbf{b}$ are the trail values for $\boldsymbol{\beta}$. Let $y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki}$ be the expected difference if the values of the vector $\mathbf{b}$ are correct. However, this difference is not zero since the response value fluctuates about their expected value resulting in errors. The method that determines the coefficients vector **b** such that the sum of the squares of the errors is minimum is called the "least squares".

Theorem 4.3

Let S (**b**) be the function representing the difference between the true and estimated values of a response variable in a multivariate linear regression problem.

$$S(\mathbf{b}) = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik} \right)^2$$

$S(\mathbf{b}) = (\mathbf{y} - \mathbf{Xb})' (\mathbf{y} - \mathbf{Xb})$ Can be written. Show that the least squares estimate **b** or $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ can be obtained when **X** has full rank $k+1 \leq n$ and is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Proof

The least square estimation of $\boldsymbol{\beta}$ in (4.4) can be given as $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

Let $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$, where $\mathbf{H}$ is the hat matrix. $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

The residuals can then be expressed as $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$

Satisfies $\mathbf{X'\hat{\varepsilon}=0}$ and $\mathbf{\hat{y}'\hat{\varepsilon}=0}$.

Residual sum of squares

$$\sum_{i=1}^{n}\left(y_i-b_0-b_1x_{1i}-b_2x_{2i}-\ldots-b_kx_{ki}\right)^2=\mathbf{\hat{\varepsilon}'\varepsilon'}$$

$$=\mathbf{y'[I-X(X'X)^{-1}X']y=y'y-y'X\hat{\beta}}$$

Let $\mathbf{\hat{\beta}=(X'X)^{-1}Xy}$

Then $\mathbf{\hat{\varepsilon}=y-\hat{y}=y-X\hat{\beta}=[I-X(X'X)^{-1}X]y}$

$\mathbf{[I-X(X'X)^{-1}X]}$ satisfies

1) $\mathbf{[I-X(X'X)^{-1}X]'=[I-X(X'X)^{-1}X]}$ symmetric.

2) $\mathbf{[I-X(X'X)^{-1}X'][I-X(X'X)^{-1}X]}$

$\mathbf{=I-2X(X'X)^{-1}X'+X(X'X)^{-1}X'X(X'X)^{-1}X'}$

$\mathbf{=[I-X(X'X)^{-1}X']}$ idempotent

3) $\mathbf{X'[I-X(X'X)^{-1}X']=X'-X'=0}$

So, $\mathbf{X'\hat{\varepsilon}=X'(y-\hat{y})=X'[I-X(X'X)^{-1}X]y=0}$, then $\mathbf{\hat{y}'\hat{\varepsilon}=\hat{\beta}'X'\hat{\varepsilon}=0}$.

In addition, $\mathbf{\hat{\varepsilon}'\varepsilon=y'[I-X(X'X')^{-1}X][I-X(X'X)^{-1}X']y=y'[I-X(X'X)^{-1}X]y}$

$\mathbf{=y'y-y'X\hat{\beta}}$.

To explain the equation for $\hat{\beta}$,*the* following can be written

$$\mathbf{y-Xb=y-X\hat{\beta}+X\hat{\beta}-Xb=y-X\hat{\beta}+X(\hat{\beta}-b)}$$

$$S(\mathbf{b})=\mathbf{(y-Xb)'(y-Xb)}$$

$$=\mathbf{(y-X\hat{\beta})'(y-X\hat{\beta})+(\hat{\beta}-b)'X'X(\hat{\beta}-b)+2(y-X\hat{\beta})'X(\hat{\beta}-b)}$$

$$=\mathbf{(y-X\hat{\beta})'(y-X\hat{\beta})+(\hat{\beta}-b)'X'X(\hat{\beta}-b)}$$

As $\mathbf{(y-Xb)'X=\hat{\varepsilon}'X=0'}$

The first part in $S(\mathbf{b})$ does not depend on $\mathbf{b}$ while the second part represents $\mathbf{X(\hat{\beta}-b)}$. Also $\mathbf{X(\hat{\beta}-b)\neq 0 \text{ if } \hat{\beta}\neq b}$ As $\mathbf{X}$ is full rank. This leads to the minimum sum of squares being unique and occurring for $\mathbf{b=\hat{\beta}=(X'X)^{-1}X'y}$. *QED*

4.3.4 Sampling Properties of the Estimators in the Classical Least Squares

1. In multiple linear regression the estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ has

   $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta},\ Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$

2. And the residual errors e_i have

   $E(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}, \text{cov}(\hat{\boldsymbol{\varepsilon}}) = \boldsymbol{\sigma}^2(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$.

3. $E(\hat{\boldsymbol{\varepsilon}}'\boldsymbol{\varepsilon}) = (n-k-1)\boldsymbol{\sigma}^2$.

Theorem 4.4

Given $\mathbf{y} = \mathbf{X} + \boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $E(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}, \text{cov}(\hat{\boldsymbol{\varepsilon}}) = \boldsymbol{\sigma}^2\mathbf{I}$. And **X** has full rank *k+1,* then for any vector of constants **h** the estimator $\mathbf{h}'\hat{\boldsymbol{\beta}} = h_0\hat{\beta}_0 + \cdots + h_k\hat{\beta}_k$ of $\mathbf{h}'\boldsymbol{\beta}$ yields the minimum variance among all linear estimators of the form

$\mathbf{h}'\mathbf{Y} = h_1Y_1 + \cdots + h_nY_n.$

Proof

Assume for any fixed **c**, $\mathbf{c}'\mathbf{Y}$ is an unbiased estimator of $\mathbf{h}'\boldsymbol{\beta}$. This means $E(\mathbf{c}'\mathbf{Y}) = \mathbf{h}'\boldsymbol{\beta}$ regardless the value of $\boldsymbol{\beta}$. We can also write $E(\mathbf{c}'\mathbf{Y}) = \text{E}(\mathbf{c}'\mathbf{X}\boldsymbol{\beta} + \mathbf{c}'\boldsymbol{\varepsilon}) = \mathbf{c}'\mathbf{X}\boldsymbol{\beta}$. From these expectations, we can write

$\mathbf{h}'\boldsymbol{\beta} = \mathbf{c}'\mathbf{X}\boldsymbol{\beta} \rightarrow (\mathbf{h}' - \mathbf{c}'\mathbf{X})\boldsymbol{\beta} = \mathbf{0}$ For all $\boldsymbol{\beta}$.

If $\boldsymbol{\beta} = (\mathbf{h}' - \mathbf{c}'\mathbf{X})'$ is taken, then $\mathbf{h}' = \mathbf{c}'\mathbf{X}$ for any unbiased estimator.

Let $\mathbf{h}'\hat{\boldsymbol{\beta}} = \mathbf{h}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{c}^{*\prime}\mathbf{Y}$ where $\mathbf{c}^* = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{h}$

From the unbiased condition $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ hence $\mathbf{h}'\hat{\boldsymbol{\beta}} = \mathbf{c}^{*\prime}\mathbf{Y}$ is an unbiased estimator of $\mathbf{h}'\boldsymbol{\beta}$, This leads to any **h** satisfying the unbiased condition $\mathbf{h}' = \mathbf{c}'\mathbf{X}$ giving

$Var(\mathbf{c}'\mathbf{Y}) = \text{Var}(\mathbf{c}'\mathbf{X}\boldsymbol{\beta} + \mathbf{c}'\boldsymbol{\varepsilon}) = \text{Var}(\mathbf{c}'\boldsymbol{\varepsilon}) = \mathbf{c}'\mathbf{I}\boldsymbol{\sigma}^2\mathbf{c}$

$$=\sigma^2(\mathbf{c}-\mathbf{c}^*+\mathbf{c}^*)(\mathbf{c}-\mathbf{c}^*+\mathbf{c}^*)=\sigma^2[(\mathbf{c}-\mathbf{c}^*)'(\mathbf{c}-\mathbf{c})'+\mathbf{c}^{*\prime}\mathbf{c}^*]$$

As $(\mathbf{c}-\mathbf{c}^*)'\mathbf{X}=\mathbf{c}\mathbf{X}-\mathbf{c}^{*\prime}\mathbf{X}=\mathbf{h}'-\mathbf{h}'=\mathbf{0}'$ we can write

$$(\mathbf{c}-\mathbf{c}^*)'\mathbf{c}^*=(\mathbf{c}-\mathbf{c}^*)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{h}=\mathbf{0}.$$

Since $\mathbf{c}^*$ is fixed and $(\mathbf{c}-\mathbf{c}^*)'(\mathbf{c}-\mathbf{c}^*)>0$ except when $\mathbf{c}=\mathbf{c}^*$, the variance of $\mathbf{cY}$ will be minimum when $\mathbf{c}^{*\prime}\mathbf{Y}=\mathbf{h}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}=\mathbf{h}'\hat{\boldsymbol{\beta}}$. *QED.*

By substituting $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ results in the best estimator of $\mathbf{h}'\boldsymbol{\beta}$ for any $\mathbf{h}$. The estimator $\mathbf{h}'\hat{\boldsymbol{\beta}}$ is named the minimum variance best linear unbiased estimator (BLUE).

4.3.5 Multicollinearity

Theoretically, each predictor in multiple linear regression has different effects on the response variable. It can be taken as each predictor having some degree of contribution in predicting the response variable's value. However, if the data matrix $\mathbf{X}$ is not of full rank, that means some linear combinations satisfy $\mathbf{aX}=\mathbf{0}$, $\mathbf{a}$ being the vector of coefficients, then the columns of $\mathbf{X}$ being *collinear* or there is a multicollinearity problem with the data, meaning $\mathbf{X}'\mathbf{X}$ does not have an inverse. In case the linear combinations of the columns of $\mathbf{X}$ being close to $\mathbf{0}$, meaning large diagonal elements in $(\mathbf{X}'\mathbf{X})^{-1}$ rendering the computation of $(\mathbf{X}'\mathbf{X})^{-1}$ unstable.

As a result, $\hat{\beta}_i$'s will have large variance. To alleviate the effect of the collinearity

1. If possible delete one pair of predictor variables that are strongly correlated.
2. Use the matrix of the principal components $\mathbf{F}$ obtained from the matrix $\mathbf{X}$.

4.3.6 Inferences on Regression Parameters

Determination of the sampling distribution of the regression parameters $\hat{\boldsymbol{\beta}}$ and error sum of squares $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ is very important. This is necessary in the assessment of the

importance of variables in a multivariate regression process. Hence the following theorems are used to clarify this point.

Theorem 4.5

Given $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ with full rank *r+1* and $\boldsymbol{\varepsilon}$ normally distributed with mean vector **0** and variance vector $\sigma^2\mathbf{I}$. Then the maximum likelihood estimator of $\boldsymbol{\beta}$ being equal to the least squares estimator of $\hat{\boldsymbol{\beta}}$. Further $\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ has a normal distribution with mean $\boldsymbol{\beta}$ and variance $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. This distribution is independent of the residuals, $\hat{\boldsymbol{\varepsilon}}=\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}$. Also $n\hat{\sigma}^2=\hat{\boldsymbol{\varepsilon}}'\boldsymbol{\varepsilon}'$ has distribution $\sigma^2\chi^2_{n-r-1}$ where $\hat{\sigma}^2$ is the maximum likelihood estimator of σ^2.

Theorem 4.6

Given $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ with full rank *r+1* and $\boldsymbol{\varepsilon}$ normally distributed with mean vector **0** and variance matrix $\sigma^2\mathbf{I}$, then $1-\alpha$ confidence area for $\boldsymbol{\beta}$ being

$$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\le(r+1)s^2F_{r+1,n-r-1(\alpha)},$$

where $F_{r+1,n-r-1(\alpha)}$ is the upper α^{th} percentile of the *F* distribution with degrees of freedom being *r+1 and* $n-r-1$.

This translates into $1-\alpha$ confidence interval for β_i being

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

Where $Var(\hat{\beta}_i)$ represents the diagonal element of $s^2(\mathbf{X}'\mathbf{X})^{-1}$

Proof

Assume we have the symmetric square-root matrix $(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}$

Let $\mathbf{V} = (\mathbf{X}'\mathbf{X})^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ where $E(\mathbf{V}) = \mathbf{0}$ and

$$\mathrm{Cov}(\mathbf{V}) = (\mathbf{X}'\mathbf{X})^{\frac{1}{2}}\,\mathrm{Cov}(\hat{\boldsymbol{\beta}})(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}$$

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})^{\frac{1}{2}} = \sigma^2\mathbf{I}$$

As $\mathbf{V}$ consists of a linear combination of the estimators $\hat{\beta}_i's$ then $\mathbf{V}$ is normally distributed. Hence

$$\mathbf{V}'\mathbf{V} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

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

Has a distribution that given as $\sigma^2\chi^2_{r+1}$. According to Theorem 4.5 $(n-r-1)s^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ has a distribution $\sigma^2\chi^2_{n-r-1}$, independent of $\hat{\boldsymbol{\beta}}$ and $\mathbf{V}$.

As a result kk

$[\chi^2_{r+1}/(r+1)]/[\chi^2_{n-r-1}/(n-r-1)] = [\mathbf{V}'\mathbf{V}/(r+1)]/s^2$ Has $F_{r+1,n-r-1}$ distribution.

From here the confidence ellipsoid for $\boldsymbol{\beta}$ can be obtained.

Note that, projecting the confidence ellipsoid of $\boldsymbol{\beta}$ for $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ benefiting from theory given in Ref [27] with $\mathbf{A}^{-1} = (\mathbf{X}'\mathbf{X})/s^2, c^2 = (r+1)F_{r+1,n-r-1}(\alpha)$ and $\mathbf{u}' = [0,\ldots,0,1,\ldots,0]$.

Where (**A**) indicates the ellipsoid $\left|\beta_i - \hat{\beta}_i\right| \le \sqrt{Var(\hat{\beta}_i)}\sqrt{(r+1)F_{r+1,n-r-1(\alpha)}}$. Here $Var(\hat{\beta}_i)$ is the diagonal element of $s^2(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to $\hat{\beta}_i$.

The next important point is, being able to do inference from the estimated regression function. That gives a set of predictor variables $\mathbf{x}_0' = [1, x_{01}, \cdots x_{0r}]$ and $\hat{\boldsymbol{\beta}}$, then the regression function $\beta_0 + \beta_1 x_{01} + \ldots + \beta_r x_{0r}$, and the value of Y at $\mathbf{x}_0$ can be estimated. In the classical regression model the expected value of Y_0 at $\mathbf{x}_0$ is given by $E(Y|\mathbf{x}_0) = \beta_0 + \beta_1 x_{01} + \ldots + \beta_r x_{0r} = \mathbf{x}_0'\boldsymbol{\beta}$, while its least squares estimate is $\mathbf{x}_0'\hat{\boldsymbol{\beta}}$.

Theorem 4.7

For the linear regression model $\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times(k+1)} + \boldsymbol{\beta}_{(k+1)\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$, $\mathbf{x}_0'\hat{\boldsymbol{\beta}}$ is the minimum variance unbiased linear estimator of $E(\mathbf{Y_0}|\mathbf{x_0})$ with variance $\text{Var}(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\sigma^2$. Then, the $1-\alpha$ confidence interval for $E(\mathbf{Y_0}|\mathbf{x_0}) = \mathbf{x}_0'\hat{\beta}$ is given by $\mathbf{x}_0'\hat{\beta} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{(\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)s^2}$ When the errors $\boldsymbol{\varepsilon}$ are normally distributed, with *n-r-1* d.f. and $\alpha/2$ upper percentile of the *t*-distribution.

Proof

When $\mathbf{x}_0$ is fixed, then $\mathbf{x}_0'\hat{\boldsymbol{\beta}}$ is a linear combination of the regression coefficients $\beta_i's$, as explained in Theorem 4.4

Also under the linear regression model it is known that,

$\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ thus

$\text{Var}(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \mathbf{x}_0'C\text{ov}\hat{\boldsymbol{\beta}}\mathbf{x}_0 = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\sigma^2$

According to Theorem 4.5 $\boldsymbol{\varepsilon}$ is assumed to be normally distributed which confirm that $\hat{\boldsymbol{\beta}}$ is $\text{N}_{r+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ independently of s^2/σ^2 that follows $x^2_{n-r-1}/(n-r-1)$

So the linear combination $\mathbf{x}_0'\hat{\boldsymbol{\beta}}$ is distributed normally as $N(\mathbf{x}_0'\hat{\boldsymbol{\beta}}, \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)$ and

$$\frac{(\mathbf{x}_0'\hat{\boldsymbol{\beta}} - \mathbf{x}_0'\boldsymbol{\beta})/\sqrt{\sigma^2\mathbf{x}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{x}_0}}{\sqrt{s^2/\sigma^2}} = \frac{(\mathbf{x}_0'\hat{\boldsymbol{\beta}} - \mathbf{x}_0'\boldsymbol{\beta})}{\sqrt{s^2(\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}}$$ follows the *t* distributed with *n-r-1* degrees of freedom.

Then the confidence interval will be $\mathbf{x}_0'\hat{\boldsymbol{\beta}} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{s^2(\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}$.

Theorem 4.8

Let $\mathbf{Y}_{(n\times 1)} = \mathbf{X}_{(n\times(r+1))}\boldsymbol{\beta}_{((r+1)\times 1)} + \boldsymbol{\varepsilon}_{(n\times 1)}$ be the linear regression model. Then by using this model to predict the a future value Y_0 can be achieved via the unbiased predictor given by

$$\mathbf{x}_0'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \ldots + \hat{\beta}_r x_{0r}$$

Then the variance of the prediction error $(Y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}})$ is

$$\mathrm{Var}(Y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \sigma^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)$$

Assuming the errors $\boldsymbol{\varepsilon}$ are normally distributed, the $1-\alpha$ prediction interval for Y_0 will be $\mathbf{x}_0'\hat{\boldsymbol{\beta}} \pm t_{n-r-1}(\alpha/2)\sqrt{s^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}$

Proof:

$\mathbf{x}_0'\hat{\boldsymbol{\beta}}$ Can be used to estimate Y_0, resulting in the estimation $E(Y_0 / z_0)$

From Theorem 4.7 $E(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$ and $\mathrm{Var}(\mathbf{x}_0'\hat{\beta}) = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\sigma^2$. can be written.

Then the prediction of error is $Y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}} = \mathbf{x}_0'\boldsymbol{\beta} + \varepsilon_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}} = \varepsilon_0 + \mathbf{x}_0'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$.

Hence $\mathrm{E}(Y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}) = E(\varepsilon_0) + E(\mathbf{x}_0'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})) = 0$ meaning the predictor is unbiased. Due to $\boldsymbol{\varepsilon}_0$ and $\hat{\boldsymbol{\beta}}$ being independent

$Var(Y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}) = Var(\varepsilon_0) + Var(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \sigma^2 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\sigma^2 = \sigma^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)$ can be written.

$\hat{\boldsymbol{\beta}}$ Is normally distributed which gives the error $\boldsymbol{\varepsilon}$ has a normal distribution. Therefore, the linear combination $Y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}$ is also normally distributed. As a result

$$(Y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}) / \sqrt{\sigma^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \qquad (4.1)$$

has standard normal distribution $N(0,1)$

Dividing the equation (4.1) by $\sqrt{s^2/\sigma^2}$. Where, $\sqrt{s^2/\sigma^2}$ having $\sqrt{\chi^2_{n-r-1}/(n-r-1)}$ a distribution with *n-r-1*d.f. Yields

$$\frac{(Y_0-\mathbf{x}_0'\hat{\boldsymbol{\beta}})/\sqrt{\sigma^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}}{\sqrt{s^2/\sigma^2}}=\frac{Y_0-\mathbf{x}_0'\hat{\boldsymbol{\beta}}}{\sqrt{s^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}}$$

that has *t* distribution with *n-r*-1 d.f.

Then the prediction interval can be obtained as

$$\mathbf{x}_0'\hat{\boldsymbol{\beta}}\ \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{s^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}\ \text{Q.}E.D.$$

4.4 Multivariate Multiple Regression

It is an important concept in a regression analysis that represents the relationship between two or more dependent variables with two or more predictors (independent variables). In other words, in this model the behavior of many response variables is determined by several independent variables.

4.4.1 Multivariate Multiple Linear Regression (MMLR) Model

$$\begin{aligned}
Y_1 &= \beta_{01}+\beta_{11}x_1+\ldots+\beta_{r1}x_r+\varepsilon_1\\
Y_2 &= \beta_{02}+\beta_{12}x_1+\ldots+\beta_{r2}x_r+\varepsilon_2\\
&\vdots \qquad\qquad \vdots\\
\mathrm{Y}_m &= \beta_{0m}+\beta_{1m}x_1+\ldots+\beta_{rm}x_r+\varepsilon_m
\end{aligned} \qquad (4.11)$$

The random error $\boldsymbol{\varepsilon}'=[\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_m]$ with properties

$$E(\boldsymbol{\varepsilon})=\mathbf{0},\ \mathrm{Var}(\boldsymbol{\varepsilon})=\boldsymbol{\Sigma}$$

Hence, the errors corresponding to different response variables may have some correlation to each other.

The following notation will be used.

$\mathbf{Y}_j'=[Y_{j1},Y_{j2},\ldots,Y_{jm}]$: The response variables.

$\left(x_{j0},x_{j1},\ldots,x_{jk}\right),\ j=1,2,\ldots,n.$ The observations of the predictor variables.

$\boldsymbol{\varepsilon}'_j = [\varepsilon_{j1}, \varepsilon_{j2}, \ldots, \varepsilon_{jm}]$: Errors associated with each response variable.

Then the design matrix is represented as in the single response regression model.

$$\mathbf{X}_{(n\times(k+1))} = \begin{bmatrix} x_{10} & x_{11} & \ldots & x_{1r} \\ x_{20} & x_{21} & \ldots & x_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \ldots & x_{nr} \end{bmatrix} \tag{4.12}$$

Response variables, regression coefficients and the error matrices are represented respectively as follows:

$$\mathbf{Y}_{(n\times m)} = \begin{bmatrix} Y_{11} & Y_{12} & \ldots & Y_{1m} \\ Y_{21} & Y_{22} & \ldots & Y_{2m} \\ \vdots & \vdots & \ddots & \\ Y_{n1} & Y_{n2} & \ldots & Y_{nm} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{(1)} & \vdots & \mathbf{Y}_{(2)} & \vdots & \cdots & \vdots & \mathbf{Y}_{(m)} \end{bmatrix} \tag{4.13}$$

$$\boldsymbol{\beta}_{((r+1)\times m)} = \begin{bmatrix} \beta_{01} & \beta_{02} & \ldots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \ldots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \ldots & \beta_{rm} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_{(1)} & \vdots & \boldsymbol{\beta}_{(2)} & \vdots & \cdots & \vdots & \boldsymbol{\beta}_{(m)} \end{bmatrix} \tag{4.14}$$

$$\boldsymbol{\varepsilon}_{(n\times m)} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \ldots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \ldots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \ldots & \varepsilon_{nm} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varepsilon}_{(1)} & \vdots & \boldsymbol{\varepsilon}_{(2)} & \vdots & \cdots & \vdots & \boldsymbol{\varepsilon}_{(m)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varepsilon}'_1 \\ \ldots \\ \boldsymbol{\varepsilon}'_2 \\ \ldots \\ \vdots \\ \ldots \\ \boldsymbol{\varepsilon}'_m \end{bmatrix} \tag{4.15}$$

MMLR model is written as

$$\mathbf{Y}_{(n\times m)} = \mathbf{X}_{(n\times(r+1))}\boldsymbol{\beta}_{((r+1)\times m)} + \boldsymbol{\varepsilon}_{(n\times m)} \tag{4.16}$$

$E(\boldsymbol{\varepsilon}_{(i)}) = 0$ and $\operatorname{cov}(\boldsymbol{\varepsilon}_{(i)}, \boldsymbol{\varepsilon}_{(k)}) = \sigma_{ik}\mathbf{I} \quad i,k=1,2,\ldots,m$

4.4.2 Least Square Estimation Method

Based on the single response regression concept, the least squares estimates of the regression coefficients are

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

Then by taking into account *m* set of regression coefficients we have

$$\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_{(1)} \vdots \hat{\boldsymbol{\beta}}_{(2)} \vdots \cdots \vdots \hat{\boldsymbol{\beta}}_{(m)}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{y}_{(1)} \vdots \mathbf{y}_{(2)} \vdots \cdots \vdots \mathbf{y}_{(m)}]$$

Or

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

Obtained estimators $\hat{\boldsymbol{\beta}}$ are used in the prediction of the response values $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Then the residuals can be computed by

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} \qquad (4.17)$$

The residuals, predicted values, and the columns of $\mathbf{Z}$ in MMLR satisfy the orthogonality condition which is also valid in classical linear regression. These become possible from the facts

$\mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \mathbf{X}' - \mathbf{X}' = \mathbf{0}$ And $\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = \mathbf{0}$ This means the residuals $\hat{\boldsymbol{\varepsilon}}_{(i)}$ as $\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}$, $\mathbf{Y}'\mathbf{Y} = (\hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}})'(\hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}) = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} + \mathbf{0} + \mathbf{0}'$ or

From this relationship we can also write

$$\underset{\substack{\text{Tot sum of squares}\\ \text{and cross products}}}{\mathbf{Y}'\mathbf{Y}} = \underset{\substack{\text{predicted sum of square}\\ \text{and cross products}}}{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} + \underset{\substack{\text{residual sum of squares}\\ \text{and cross products}}}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}$$

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{x}\hat{\boldsymbol{\beta}}$$

4.4.3 Hypothesis Test in Multiple Linear Regression

Individual regression coefficients follow a normal distribution. Hence, the confidence interval for these coefficients can be computed and tested via hypothesis testing. From section (3.2.3), $b_j, j = 1,2,\ldots,k$ are normally distributed with mean β_j and variance $c_{jj}\sigma^2$. Where c_{jj} is the diagonal elements of the matrix $(\mathbf{X}'\mathbf{X})$

Note that s^2 is an unbiased estimator of σ^2 which is given by

$$s^2 = \frac{SSE}{n-k-1}$$

Then the T statistic $t = \frac{b_j - \beta_{j0}}{s\sqrt{c_{jj}}}$ has $(n-k-1)$ degrees of freedom. Then the hypothesis test on β_j can be written as

$$H_0 : \beta_j = \beta_{j0}$$
$$H_1 : \beta_j \neq \beta_{j0}$$

Then, if the test statistic t is $-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}$, H_0 is not rejected. It means $\beta_j = \beta_{j0}$ is acceptable.

4.4.4 Confidence Interval for Mean Response $\mu_Y | x_{10}, x_{20}, \ldots, x_{k0}$

The $1-\alpha$ confidence interval for the mean response $\mu_Y | x_{10}, x_{20}, \ldots, x_{k0}$ is

$$\hat{y}_0 - t_{\frac{\alpha}{2}} s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} < \mu_Y | x_{10}, x_{20}, \ldots, x_{k0} < \hat{y}_0 + t_{\frac{\alpha}{2}} s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$$

where $t_{\frac{\alpha}{2}}$ has $(n-k-1)$, degrees of freedom of the t-distribution.

4.4.5 The Effect of Some Factors on the Width of the Confidence Interval

The confidence interval for the mean response can be computed by using an appropriate software. It is considered necessary just to list some of the factors that affect this confidence interval. The width of the confidence interval decreases as the mean square error (MSE) decreases.

- The width of the confidence interval decreases as the level of the confidence level decrease (since the t-multiplier decreases when the confidence interval decreases).
- The width of the confidence interval decreases as the sample size (n) increases.

Chapter 5

PRINCIPAL COMPONENTS ANALYSIS

5.1 Introduction to Principal Components Analysis

5.1.1 Definition of Principal Components

When a process is governed by a large number of variables (p), the processing of data belonging to so many variables becomes a problem. One way of efficiently eliminating this handicap, is to find some linear combination of these variables in such a way that only a few of the linear combinations will be able to represent the process adequately. Given the p random variables $X_1, X_2, \ldots, X_p$, their linear combination will be in a new coordinate system via the rotation of the original system. This new coordinate system points to the directions where maximum variable exists in the original data. These linear combinations are called Principal Components (PC). The covariance $\boldsymbol{\Sigma}$ or correlation $\boldsymbol{\rho}$ matrices obtained from $X_1, X_2, \ldots, X_p$ are used in determining the coefficients of the PCs.

In this chapter, the theoretical concepts of the principal components analysis (PCA) will be discussed and illustrated by some theorems and proofs.

5.1.2 Computation of the Principal Components

Let $\boldsymbol{\Sigma}$ be the covariance matrix obtained from the data matrix $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$.

Let the eigenvalues of $\boldsymbol{\Sigma}$ in descending order be $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

Then the eigenvectors corresponding to the eigenvectors are $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ respectively.

Using the elements of eigenvectors and the random variables forming the data matrix, the following linear combinations can be written for the population.

$$\begin{aligned} Y_1 &= \mathbf{e_1'X} = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ Y_2 &= \mathbf{e_2'X} = e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ &\vdots \qquad\qquad\qquad \vdots \\ Y_p &= \mathbf{e_p'X} = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned} \tag{5.1}$$

These linear combinations (PCs) expressed as $y_i = \mathbf{e}_i'\mathbf{x}_i;\ i = 1, 2, \ldots, p$ expressed in complete form in (5.2).

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}_{(p\times 1)} = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1p} \\ e_{21} & e_{22} & \cdots & e_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ e_{p1} & e_{p2} & \cdots & e_{pp} \end{bmatrix}_{(p\times p)} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_P \end{bmatrix}_{(p\times 1)} \tag{5.2}$$

The mean and covariance of the PCs can be written as in (5.3)

$$\begin{aligned} \boldsymbol{\mu}_Y &= E(\mathbf{y}) = E(\mathbf{e}_i'\mathbf{x}) = \mathbf{e}_i'\boldsymbol{\mu}_x \\ \boldsymbol{\Sigma}_Y &= Cov(Y_i, Y_k) = Cov(\mathbf{e}'\mathbf{x}) = \mathbf{e}'\boldsymbol{\Sigma}_X\mathbf{e} \end{aligned} \tag{5.3}$$

From equation (5.3) we can obtain

$$\begin{aligned} Var(Y_i) &= \mathbf{e_i'\Sigma e}_i \qquad i = 1, 2, \ldots, p \\ \mathrm{Cov}(Y_i, Y_k) &= \mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{e}_i = \mathbf{0} \quad i = 1, 2, \ldots, p \end{aligned}$$

The PCs $Y_1, Y_2, \ldots, Y_p$ are uncorrelated linear combinations obtained by using the eigenvectors of the covariance matrix. Each PC has maximum variance subject to the condition the eigenvectors are normal. That is

$Var(\mathbf{e}_i'\mathbf{X})$ is max. if $\mathbf{e}_i'\mathbf{e}_i = 1$ And

$\mathrm{Cov}(\mathbf{e}_i'\mathbf{X}, \mathbf{e}_k'\mathbf{X}) = 0 \quad$ for $\mathrm{k} < i$

Theorem 5.1

Given the random vector $\mathbf{X}' = [x_1, x_2, x_3, \ldots x_p]$ with covariance matrix $\mathbf{\Sigma}$ and the eigenvalue–eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ the PC can be written as

$Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{ip}X_p, \quad i = 1, 2, \ldots, p$ Where

$\text{Var}(Y_i) = \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_i = \lambda_i \quad i = 1, 2, \ldots, p$
$\text{Cov}(Y_i, Y_k) = \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_k = 0 \quad i \neq k$

Proof

Given a positive definite *p*x*p* square matrix **A** with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ and corresponding normalized eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$, it is known that the maximization of quadratic forms with points on the unit sphere can be achieved via $\max_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_i$, when $\mathbf{x} = \mathbf{e}_i$ and **A** is the covariance matrix $\mathbf{\Sigma}$. Then

$\max_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{e}'\mathbf{\Sigma}\mathbf{e}}{\mathbf{e}'\mathbf{e}} = \mathbf{e}'\mathbf{\Sigma}\mathbf{e} = Var(Y_i).$

Similarly

$\max_{a \perp e_1, e_2, \ldots, e_k} \frac{\mathbf{x}'\mathbf{\Sigma}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_{k+1}$ Can be written. Setting $\mathbf{x} = \mathbf{e}_{k+1}, k = 1, 2, \ldots, p-1$ with $\mathbf{e}_{k+1}'\mathbf{e}_i = 0, \text{for } i = 1, 2, \ldots, k \text{ and } k = 1, 2, \ldots, p-1$ we get

$\mathbf{e}_{k+1}'\mathbf{\Sigma}\mathbf{e}_{k+1} / \mathbf{e}_{k+1}'\mathbf{e}_{k+1} = Var(Y_{k+1})$.

However $\mathbf{e}_{k+1}'(\mathbf{\Sigma}\mathbf{e}_{k+1}) = \lambda_{k+1}\mathbf{e}_{k+1}'\mathbf{e}_{k+1} = \lambda_{k+1}$ so $Var(Y_{k+1}) = \lambda_{k+1}$. That means

$\mathbf{e}_i$ And $\mathbf{e}_k$ are perpendicular (orthogonal) which means $\mathbf{e}_i'\mathbf{e}_k = 0, i \neq k,$ giving

$\text{Cov}(Y_i, Y_k) = 0$. Then

1. If all the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ are distinct, the eigenvectors of $\mathbf{\Sigma}$ will be orthogonal.

2. If the eigenvalues are not all distinct, the eigenvectors corresponding to common eigenvalues may be selected to be orthogonal. So, for any two eigenvectors $\mathbf{e_i}, \mathbf{e_k}$ where $\mathbf{e_i'e_k} = 0, i \neq k.$

$\mathbf{\Sigma e_k} = \boldsymbol{\lambda}_k \mathbf{e_k}$ Multiplying by $\mathbf{e_i}'$ giving

$$Cov(Y_i, Y_k) = \mathbf{e}_i' \mathbf{\Sigma e}_k = \mathbf{e'} \boldsymbol{\lambda_k} \mathbf{e}_k = \boldsymbol{\lambda_k} \mathbf{e'e}_k = 0$$

For any $i \neq k.$

Note: It must be stressed that some of the coefficient vectors $\mathbf{e}_i$ and so Y_i will not be unique, if some eigenvalues λ_i are equal.

Theorem 5.2

Given a random vector $\mathbf{X'} = [X_1, X_2, X_3, \ldots, X_p]$ that has covariance matrix $\mathbf{\Sigma}$

With eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ corresponding to the eigenvectors $\mathbf{e_1, e_2, \ldots, e_p}$. and PCs $Y_1 = \mathbf{e}_1' \mathbf{X}$, $Y_2 = \mathbf{e}_2' \mathbf{X}, \ldots, Y_P = \mathbf{e}_P' \mathbf{X}$, it can be shown that

$$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \sum_{i=1}^{p} Var(\mathbf{X}_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{p} Var(Y_i) \qquad (5.2.1)$$

Proof

It is known that $tr(\mathbf{A}) = \sum_{i=1}^{k} a_{ii}$ where $\mathbf{A} = \{a_{ij}\}$ indicates a $k \times k$ square matrix.

Applying this to the covariance matrix $\mathbf{\Sigma}$ yields

$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = tr(\mathbf{\Sigma})$, then by using $\mathbf{A} = \sum_{i=1}^{k} \lambda_i \mathbf{e}_{i_{(k\times 1)}} \mathbf{e}'_{i_{(1\times k)}} = \mathbf{P}_{(k\times k)} \mathbf{\Lambda}_{(k\times k)} \mathbf{P}'_{(k\times k)}$ with

$\mathbf{A} = \mathbf{\Sigma}$

$\mathbf{\Sigma} = \mathbf{P \Lambda P'}$ Where $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues and, $\mathbf{P} = [e_1, e_2, \ldots, e_p]$

Then $\mathbf{P'P = PP' = I}.$

$tr(\mathbf{\Sigma}) = tr(\mathbf{P\Lambda P'}) = tr(\mathbf{\Lambda PP'})$, Since $tr(AB) = tr(BA)$.

Then, $tr(\mathbf{\Lambda PP'}) = tr(\mathbf{\Lambda}) = \lambda_1 + \lambda_2 + \ldots + \lambda_p$.

Then, $\sum_{i=1}^{p} Var(\mathbf{X}_i) = tr(\mathbf{\Sigma}) = tr(\mathbf{\Lambda}) = \sum_{i=1}^{p} Var(\mathbf{Y}_i)$. *Q.E.D.*

Theorem 5.3

Given principal components, $Y_1 = \mathbf{e}_1'\mathbf{X},\ Y_2 = \mathbf{e}_2'\mathbf{X}, \ldots, Y_P = \mathbf{e}_P'\mathbf{X},$ then the correlation coefficients between the variables X_k and the principal components Y_i are

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i = 1, 2, \ldots, p,$$

Proof

$\mathbf{a}_k' = [0, \ldots, 0, 1, 0, \ldots, 0]$, is given, so $X_k = \mathbf{a_k'X}$ and $\mathrm{Cov}(X_k, Y_i) = Cov(\mathbf{a_k'X}, \mathbf{e}_i'\mathbf{X})$. Also $Cov(\mathbf{a_k'X}, \mathbf{e}_i'\mathbf{X}) = \mathbf{a_k'\Sigma e}_i$ based on the maximization of quadratic forms on the unit sphere concept.

Remembering that $Cov(X_k, Y_i) = \mathbf{a_k'\lambda e}_i = \lambda_i e_{ik}$ and $\mathbf{\Sigma e}_i = \lambda_i \mathbf{e}_i$, then

$Var(Y_i) = \mathbf{e_i'\Sigma e}_k = \lambda_i$ Yields $Var(Y_i) = \lambda_i$ and, $Var(\mathbf{X}_k) = \sigma_{kk}$ which gives

$$\rho_{Y_i, X_k} = \frac{Cov(Y_i, X_k)}{\sqrt{Var(Y_i)}\sqrt{Var(X_k)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i, k = 1, 2, \ldots, p \quad Q.E.D.$$

5.2 Optimum Number of the Principal Components

There are some important points that are used to judge on the optimum number of PCs which offer valuable information and summarize the data.

The guidelines are explained as follows:

1. Keeping enough principal components to account for a high proportion of the total variance. By using the following formula

$$\frac{\sum_{j=1}^{k}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \qquad (5.1.1)$$

Accepting a threshold value for this proportion the value of k can be determined. This will be illustrated in the example 5.1.

2. Keeping the principal components that have eigenvalues which are greater than the average of the eigenvalues.
3. Using the scree plot where λ_i (the eigenvalues) versus i (the number of the eigenvalues) starting with the largest towards the smallest λ_i are plotted. The optimum number of the PCs is taken to be those before the point where an elbow occurs in the graph including the elbow point. Figure 5.1 shows the scree plot obtained in example 5.1. The optimum number of the principal components is considered 2 since the elbow occurs at point 2.

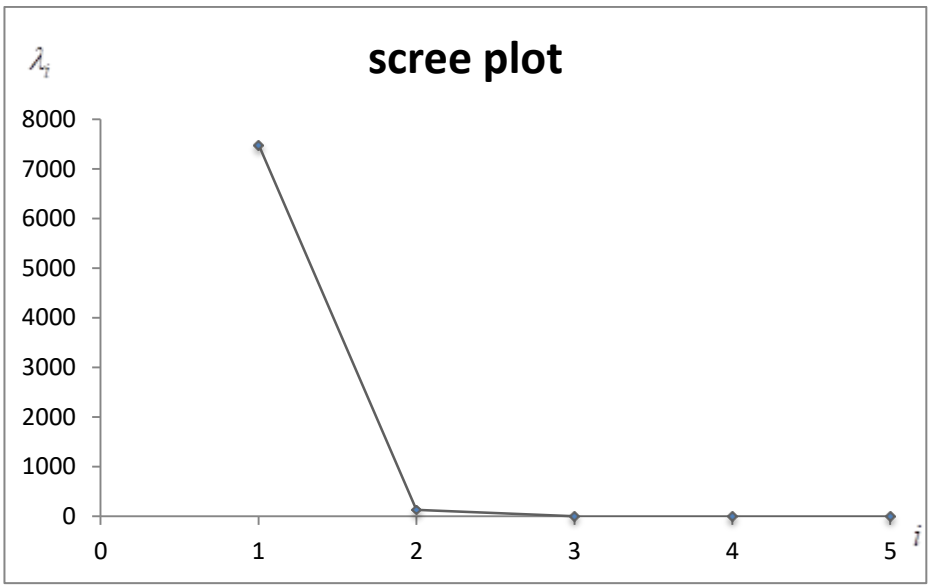

Figure 5.1: Scree Plot to Determine the Number of PCs

5.3 Interpretation of the Principal Components

5.3.1 Rotation

Essentially, we obtain the principal components by doing a rotation of the axes orderly to line up with natural extension of the system, then the obtained variables are changed to be uncorrelated with making a reflection to the directions of the maximum variance.

In case the interpretation of the new principal components was not satisfactory, then the angle of rotation can be changed and re-examine the variance direction relation.

5.3.2 The Correlation between the Variables and the Principal Components

It is recommended to use the correlation coefficient that is between the dependent variable and the principal components to get an interpretation of the components.

While the magnitude of the coefficients of a PC indicates the contribution of that variable to the PC, the correlation between a variable and a PC may not be in direct association to the contribution of the variable to the PC.

Example 5.1

Source of Dataset

The data were collected from the database of National Agency Savreio Devito, (saverio.devito'@'enea.it).

Information about the Dataset

The dataset consists of 48 observations and 5 variables indicate responses of an oxidizing chemical gas Multisensor device was installed in polluted area on the field in some city in Italy. The data were obtained in the period from March 2004 till February 2005. The variables used are as explained below.

X_1: True hourly averaged concentration CO in mg/m^3

X_2: True hourly averaged overall Non Metanic Hydro Carbons concentration in $microgram/m^3$.

X_3: True hourly averaged Benzene concentration in $microgram/m^3$

X_4: Temperature C°

X_5: Relative Humidity (%).

See Appendix, Table 5.1: Raw Data for Responses of an Oxidizing Chemical Gas Multisensor Device

To find the principle components, $\boldsymbol{\Sigma}$ matrix is calculated by Matlab as follows

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.0014 & 0.0992 & 0.0066 & 0.0011 & -0.0022 \\ 0.0992 & 7.4434 & 0.4781 & 0.1057 & -0.2802 \\ 0.0066 & 0.4781 & 0.0325 & 0.0077 & -0.0208 \\ 0.0011 & 0.1057 & 0.0077 & 0.0087 & -0.0302 \\ -0.0022 & -0.2802 & -0.0208 & -0.0302 & 0.1383 \end{bmatrix}$$

Eigenvalues and corresponding eigenvectors are:

$\lambda_1 = 7487.8 \qquad \mathbf{e}_1 = [-0.0133,\ -0.9970,\ -0.0641,\ -0.0143,\ 0.0383]$

$\lambda_2 = 133 \qquad \mathbf{e}_2 = [0.0120,\ 0.0415,\ -0.0195,\ -0.2029,\ 0.9780]$

$\lambda_3 = 2.1 \qquad \mathbf{e}_3 = [-0.0643,\ 0.0496,\ -0.6990,\ -0.6925,\ -0.1589]$

$\lambda_4 = 1.3 \qquad \mathbf{e}_4 = [-0.1373,\ 0.0417,\ -0.6975,\ 0.6901,\ 0.1292]$

$\lambda_5 = 0.0215 \qquad \mathbf{e}_{15} = [0.9883,\ -0.0049,\ -0.1430,\ 0.0530,\ -0.0037]$

Then by using equation (5.1) the five principal components are obtained. See Appendix, Table 5.2: Row Data PCs. To determine the optimum number of the principal components, it is determined from the scree plot that 2 PCs are enough. It is also evident that the first two PCs represent 99% of total variation in the data

according to equation 5.1.1, which strongly supports the result obtained from the scree plot.

$\hat{Y}_1$=0.98, $\hat{Y}_2$=0.017 of the total variation, It means the five variables can be replaced by only first two principle components without much loss information since they explain approximately 0.99 of the total variation. Then the first 2 PCs will be

$$Y_1 = -0.0133X_1 - 0.997X_2 - 0.0641X_3 - 0.0143X_4 + 0.0383X_5$$
$$Y_2 = 0.012X_1 + 0.041X_2 - 0.0195X_3 - 0.2029X_4 + 0.9780X_5 .$$

Here the first PC which represents about 99% of total variation is very important. It is also evident that the second variable with the highest negative coefficient, means true hourly averaged overall Non Metanic Hydro Carbons concentration has a major negative influence on the process in the direction where the biggest variation exists. On the second PC relative humidity seems to be the leading variable influencing the process in the second main direction with second larger variation occurs. Now the correlation coefficients between the first two principal components and the independent variables are calculated.

Table 5.3: Correlation Coefficients between Variables and PCs ($\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$).

	X_1	X_2	X_3	X_4	X_5
$\hat{Y}_1$	-0.972	-1.000	-0.973	-0.421	0.281
$\hat{Y}_2$	0.117	0.006	-0.039	-0.796	0.959

By checking table 5.3 it is obvious that the first principal component $\hat{Y}_1$ has high correlations with most of the variables since 98% of the total variation was explained by $\hat{Y}_1$, and the vice versa for the second one $\hat{Y}_2$ which represents 0.017 of the total variation.

Chapter 6

APPLICATIONS

6.1 The Target of the Application

Topics and theory explained in Chapter 4 regarding multivariate regression, and PCA as explained in Chapter 5 are used in the analysis of multivariate data. A study of the relationship between the linear regression and the PCA is undertaken. For this purpose the comparison of the MSE values obtained from multivariate regression, and MSE values obtained by treating each of the data variables as dependent variable and PCAs as the predictor variables.

Some factors which have a functional impact on the linear regression performance:

- The correlation coefficients between the dependent variable (response) and the independent variables (predictors).
- The number of the independent variables.
- Repeating the values in the independent variables.
- The size of the selected samples of the data.

6.2 Data Description

The data is related to the variables that affects the heat retaining capacity of a building. Therefore the heat retaining capacity is considered as the dependent (response) variable, onto 8 independent (predictor) variables.

6.2.1 Source of the Data

The dataset was created by Angeliki Xifara (angxifara '@' gmail.com, Civil/Structural Engineer) and was processed by Athanasios Tsanas (tsanasthanasis '@' gmail.com, Oxford Centre for Industrial and Applied Mathematics, Oxford University, UK).

6.2.2 Information about the Data

The dataset consists of eight independent variables or features $(X_1, X_2, X_3, \ldots, X_8)$ and a response variable (Y_1). In the original data the number of the observations was 768 for each variable. Description of the variables is as follows.

X_1 : Relative compactness

X_2 : Surface area

X_3 : Wall area

X_4 : Roof area

X_5 : Overall height

X_6 : Orientation

X_7 : Glazing area distribution

X_8 : Glazing area distribution

Y_1 : Heating load.

6.2.3 Cleaning or Validating the Data

A preliminary study of the data via simple linear regression and the linear correlation r_{Y,X_i}; $i = 1, 2, \ldots, 8$ between the response and each predictor variable was undertaken. Mean Square Errors (MSE) and r_{Y,X_i} are given in Table 6.1.

Table 6.1: Simple Linear Regression and the Linear Correlation between the response and each predictor variable for the data in 6.2.2.

Linear regression equation	**The correlation coefficient**	**MSE**
$E(Y/X_1)$	$r_{Y,X_1}=0.4024$	24.5755
$E(Y/X_2)$	$r_{Y,X_2}=0.4529$	49.4473
$E(Y/X_3)$	$r_{Y,X_3}=0.1824$	73.8964
$E(Y/X_4)$	$r_{Y,X_4}=0.744$	23.1398
$E(Y/X_5)$	$r_{Y,X_5}=0.8024$	17.8573
$E(Y/X_6)$	$r_{Y,X_6}=0.0431$	82.2517
$E(Y/X_7)$	$r_{Y,X_7}=0.0431$	76.1569
$E(Y/X_8)$	$r_{Y,X_8}=0.0026$	79.2872

Based on the results given in Table 6.1, it was considered necessary to reduce the number of variables from 8 to 5, as the variables X_6, X_7, and X_8 have very low correlation with the response variable and also resulted in high MSE values. Therefore, it was decided to discard these variables from the study, and using the data consisting of the following variables.

X_1 : Relative compactness.

X_2 : Surface area.

X_3 : Wall area.

X_4 : Roof area.

X_5 : Overall height.

Y_1 : Heating load.

It was also observed that in the original data under most variables, data values were either repeating or missing. Hence, it was decided to select a subset or a pilot data set of 50 observations Where repeated or missing values did not exist.

See Appendix, Table 6.2: Row Data about the Heating Load Process

6.2.4 Application of Simple and Multivariate Regression to Pilot Data

As a first step the correlation matrix **R** for the pilot data is computed to identify the correlation level between the response and predictor variables. This is given below.

$$R=\begin{bmatrix} r_{y,y} & r_{y,x_1} & r_{y,x_2} & r_{y,x_3} & r_{y,x_4} & r_{y,x_5} \\ r_{x_1,y} & r_{x_1,x_1} & r_{x_1,x_2} & r_{x_1,x_3} & r_{x_1,x_4} & r_{x_1,x_5} \\ r_{x_2,y} & r_{x_2,x_1} & r_{x_2,x_2} & r_{x_2,x_3} & r_{x_2,x_4} & r_{x_2,x_5} \\ r_{x_3,y} & r_{x_3,x_1} & r_{x_3,x_2} & r_{x_3,x_3} & r_{x_3,x_4} & r_{x_3,x_5} \\ r_{x_4,y} & r_{x_4,x_1} & r_{x_4,x_2} & r_{x_4,x_3} & r_{x_4,x_4} & r_{x_4,x_5} \\ r_{x_5,y} & r_{x_5,x_1} & r_{x_5,x_2} & r_{x_5,x_3} & r_{x_5,x_4} & r_{x_5,x_5} \end{bmatrix}=$$

$$\begin{bmatrix} 1 & 0.6232 & -0.6704 & 0.3649 & -0.8365 & 0.8998 \\ 0.6323 & 1 & -0.9922 & -0.2437 & -0.8945 & 0.8410 \\ -0.6704 & -0.9922 & 1 & 0.2381 & 0.9048 & -0.8727 \\ 0.3649 & -0.2437 & 0.2381 & 1 & -0.1981 & 0.2011 \\ -0.8365 & -0.8945 & 0.9048 & -0.1981 & 1 & -0.9689 \\ 0.8998 & 0.8410 & -0.8727 & 0.2011 & -0.9689 & 1 \end{bmatrix}$$

The previous matrix was checked so that, from the first row (or first column) the correlation values between the response and predictors were obtained.

The diagonal values indicate the correlation between the variables themselves.

Simple linear regression values between the response and predictor variables $E(Y|X_i);\ i=1,2,\ldots,5$ are obtained (Table 6.3) from the pilot data have significantly improved.

Table 6.3: Simple linear Regression and the Linear Correlation between the Response and Each Predictor Obtained from the Pilot Data.

$E(Y\|X_i)$	**The correlation coefficient**	**MSE**
$E(Y\|X_1)$	$r_{Y,X_1}=0.634$	52.14
$E(Y\|X_2)$	$r_{Y,X_2}=-0.678$	81.83
$E(Y\|X_3)$	$r_{Y,X_3}=0.533$	73.89
$E(Y\|X_4)$	$r_{Y,X_4}=-0.841$	25.59
$E(Y\|X_5)$	$r_{Y,X_5}=0.900$	16.22

It was then decided to check for multiple linear regression, initially using all possible 2 predictor pairs to estimate the response variable $E(Y|X_i, X_j);\ i, j = 1,2,\ldots,5;\ i < j.$ Then obtained MSE results from each pair of predictors, and correlation between these predictors and the response variable are presented in Table 6.4.

Table 6.4: MSE Results from Each Pair of predictors, and Correlation between these Predictors and the Response Variable.

$E(Y\|X_i, X_j)$	**Correlation coefficient**		**MSE**
$E(Y\|X_2, X_4)$	r_{Y,X_2} =-0.678	r_{Y,X_4} = -0.841	25.73
$E(Y\|X_1, X_3)$	r_{Y,X_1} = 0.634	r_{Y,X_3} = 0.533	28.35
$E(Y\|X_1, X_5)$	r_{Y,X_1} = 0.634	r_{Y,X_5} =0.900	11.25
$E(Y\|X_3, X_5)$	r_{Y,X_3} = 0.533	r_{Y,X_5} =0.900	13.69
$E(Y\|X_1, X_4)$	r_{Y,X_1} = 0.634	r_{Y,X_4} = -0.841	19.32
$E(Y\|X_2, X_3)$	r_{Y,X_2} = -0.678	r_{Y,X_3} = 0.533	74.43
$E(Y\|X_2, X_5)$	r_{Y,X_2} =-0.678	r_{Y,X_5} =0.900	16.31
$E(Y\|X_3, X_4)$	r_{Y,X_3} = 0.533	r_{Y,X_4} = -0.841	22.54
$E(Y\|X_4, X_5)$	r_{Y,X_4} = -0.841	r_{Y,X_5} =0.900	14.79
$E(Y\|X_1, X_2)$	r_{Y,X_1} = 0.634	r_{Y,X_2} =-0.678	14.79

Subsequently, multiple linear regression for all tuples with 3 predictors and the response variable is carried out $E(Y|X_i, X_j, X_k);\ i, j, k = 1,2,\ldots,5;\ i < j < k.$ MSE results from each triplet of predictors, and correlation between these predictors and the response variable are presented in Table 6.5.

Table 6.5: MSE Results from Each Triplet of Predictors, and Correlation between these Predictors and the Response Variable.

$E(Y\|X_i, X_j, X_k)$	**The correlation coefficient**			**MSE**
$E(Y\|X_1, X_2, X_3)$	r_{Y,X_1} = 0.634	r_{Y,X_2} =-0.678	r_{Y,X_3} = 0.533	29.13
$E(Y\|X_1, X_2, X_4)$	r_{Y,X_1} = 0.634	r_{Y,X_2} =-0.678	r_{Y,X_4} = -0.841	19.73
$E(Y\|X_1, X_2, X_5)$	r_{Y,X_1} = 0.634	r_{Y,X_2} =-0.678	r_{Y,X_5} =0.900	11.48
$E(Y\|X_2, X_3, X_4)$	r_{Y,X_2} =-0.678	r_{Y,X_3} = 0.533	r_{Y,X_4} = -0.841	22.9
$E(Y\|X_2, X_3, X_5)$	r_{Y,X_2} =-0.678	r_{Y,X_3} = 0.533	r_{Y,X_5} =0.900	13.78
$E(Y\|X_3, X_4, X_5)$	r_{Y,X_3} = 0.533	r_{Y,X_4} = -0.841	r_{Y,X_5} =0.900	11.91
$E(Y\|X_3, X_4, X_1)$	r_{Y,X_3} = 0.533	r_{Y,X_4} = -0.841	r_{Y,X_1} = 0.634	14.79
$E(Y\|X_4, X_5, X_1)$	r_{Y,X_4} = -0.841	r_{Y,X_5} =0.900	r_{Y,X_1} = 0.634	11.5
$E(Y\|X_5, X_4, X_2)$	r_{Y,X_5} =0.900	r_{Y,X_4} = -0.841	r_{Y,X_2} =-0.678	14.85
$E(Y\|X_1, X_3, X_5)$	r_{Y,X_1} = 0.634	r_{Y,X_3} = 0.533	r_{Y,X_5} =0.900	11.49

As expected, an increase in the number of predictors with high correlation with the response variable results in a decrease in the MSE values. This is clearly evident from Tables 6.3, 6.4, and 6.5.

Also in case four and five variables is shown in Tables 6.6 and 6.7 and have the result as same as previous Tables.

Table 6.6: MSE Results from Each Tuple of 4 Variables of Predictors, and Correlation between these Predictors and the Response Variable.

$E(Y\|X_i, X_j, X_k)$	**The correlation coefficient**				**MSE**
$E(Y\|X_1, X_2, X_3, X_4)$	r_{Y,X_1} = 0.634	r_{Y,X_2} = -0.678	r_{Y,X_3} = 0.533	r_{Y,X_4} = -0.841	14.87
$E(Y\|X_1, X_2, X_3, X_5)$	r_{Y,X_1} = 0.634	r_{Y,X_2} = -0.678	r_{Y,X_3} = 0.533	r_{Y,X_5} =0.900	11.72
$E(Y\|X_1, X_2, X_4, X_5)$	r_{Y,X_1} = 0.634	r_{Y,X_2} = -0.678	r_{Y,X_4} = -0.841	r_{Y,X_5} =0.900	11.73
$E(Y\|X_1, X_3, X_4, X_5)$	r_{Y,X_2} = -0.678	r_{Y,X_3} = 0.533	r_{Y,X_4} = -0.841	r_{Y,X_5} =0.900	11.62
$E(Y\|X_2, X_3, X_4, X_5)$	r_{Y,X_2} = -0.678	r_{Y,X_3} = 0.533	r_{Y,X_4} = -0.841	r_{Y,X_5} =0.900	12.17

Table 6.7: Five Variables with the Response.

$E(Y\|X_i, X_j, X_k)$	**The correlation coefficient**					**MSE**
$E(Y\|X_1, X_2, X_3, X_4, X_5)$	r_{Y,X_1} = 0.634	r_{Y,X_2} = -0.678	r_{Y,X_3} = 0.533	r_{Y,X_4} = -0.841	r_{Y,X_5} =0.900	11.83

6.2.5 Application of PCA to the Pilot Data

As required by the PCA theory, the correlation matrix **R** for the 5 response variables is computed. Use of the **R** matrix is due to the facts explained under the relevant theory. That is a significant difference between the data values of different variables, and different units between the variables.

Then the obtained correlation matrix is given below.

$$\mathbf{R} = \begin{bmatrix} 1.000 & -0.9922 & -0.2437 & -0.8945 & 0.8410 \\ -0.9922 & 1.000 & 0.2381 & 0.9048 & -0.8727 \\ -0.2437 & 0.2381 & 1.000 & -0.1981 & 0.2011 \\ -0.8945 & 0.9048 & -0.1981 & 1.000 & -0.9689 \\ 0.8410 & -0.8727 & 0.2011 & -0.9689 & 1.000 \end{bmatrix}$$

The eigenvalues and corresponding eigenvectors are as follows,

$\lambda_1 = 3.7381,\ \mathbf{e}_1' = [0.4998,\ -0.5054,\ -0.0159,\ -0.5030,\ 0.4914]$

$\lambda_2 = 1.1986,\ \mathbf{e}_2' = [-0.2036,\ 0.1916,\ 0.9109,\ -0.2059,\ 0.2229]$

$\lambda_3 = 0.0596,\ \mathbf{e}_3' = [0.4968,\ -0.0835,\ 0.2761,\ -0.2053,\ -0.7924]$

$\lambda_4 = 0.0037,\ \mathbf{e}_4' = [0.6796,\ 0.4901,\ 0.0828,\ 0.4583,\ 0.2846]$

$\lambda_5 = -0.00,\ \mathbf{e}_5' = [-0.000,\ 0.6788,\ -0.2948,\ -0.6726,\ -0.000]$

Subsequently the five principal components can be written as follows

$$Y_1 = 0.4998X_1 - 0.5054X_2 - 0.0159X_3 - 0.5030X_4 + 0.4914X_5$$
$$Y_2 = -0.2036X_1 + 0.1916X_2 + 0.9109X_3 - 0.2059X_4 + 0.2229X_5$$
$$Y_3 = 0.4968X_1 - 0.0835X_2 + 0.2761X_3 - 0.2053X_4 - 0.7924X_5$$
$$Y_4 = 0.6796X_1 + 0.4901X_2 + 0.0828X_3 + 0.4583X_4 + 0.2846X_5$$
$$Y_5 = -0.0000X_1 + 0.6788X_2 - 0.2948X_3 - 0.6726X_4 - 0.000X_5$$

Using these PCs and the pilot data, the obtained PC values are given in Appendix, Table: 6.8

However, based on the eigenvalues it is observed that 99% of the total variation are represented by the first two eigenvalues. That is $\frac{\lambda_1+\lambda_2}{\sum_{i=1}^{5}\lambda_i} = \frac{4.9367}{5} = 0.98734.$ This means the first two PCs $(\hat{Y}_1, \hat{Y}_2)$ are enough to represent the data (Theorm 5.2). Since the PC's are used to reduce the number of the variables, then those five independent variables can be represented by these two principal components.

6.2.6 Linear Regression Using the PCs as Predictors

Correlation coefficient values between the response and the PCs has surprisingly produced results in absolute terms, very close to the percent of variation each PC represents over the total variation in the data. Subsequently the linear regression between the response variable and each PC generated MSE results as given in Table

6.9. It is an expected result to observe higher MSE when correlation is low, as in $E(Y/\hat{Y}_2)$.

Table 6.9: Correlation and Regression Results between the Response and the PCs.

Linear regression equation	**The correlation coefficient**	**Mean square of error**
$E(Y/\hat{Y}_1)$	$r_{Y,Y_1} = 0.74$	38.614
$E(Y/\hat{Y}_2)$	$r_{Y,Y_2} = -0.224$	80.978

6.2.7 Using the PCs as Predictors to Estimate X_1, X_2, X_3, X_4, X_5

As the PCs are a linear combination of the variables, X_1, X_2, X_3, X_4, X_5. And in the pilot data set the first two PCs represent 75% and 24% of total variation in the data respectively. Hence, it is logical to expect a dependence of each *X* variable onto each PC. In order to decide which *X* variables to be used as a response variable, and which PC to be used as a predictor, the correlations between each of the variables X_1, X_2, X_3, X_4, X_5 and the PCs (Y_1, Y_2) are computed and presented in Table 6.10.

Table 6.10: Correlation values between (Y_1, Y_2) and each of X_1, X_2, X_3, X_4, X_5.

$r_{Y_i, Xj}$	X_1	X_2	X_3	X_4	X_5
Y_1	0.981	-0.990	-0.102	-0.955	0.923
Y_2	-0.409	0.405	0.985	-0.023	0.033

Based on the correlation coefficients between each PC and each variable X_1, X_2, X_3, X_4, X_5, linear regression models $E(X_i/Y_j);\ i = 1, 2, \ldots, 5;\ j = 1, 2$ are

created for the five pairs with highest correlation. Obtained MSE results are given in Table 6.11.

Table 6.11: Correlation Coefficients and MSE Values Obtained from Using Each of X_1, X_2, X_3, X_4, X_5 as a Response and Each PC as a Predictor.

$E(X_i \mid Y_j)$	**The correlation coefficient**	**MSE**	**RMSE**	**RMSED/** $\bar{X}_i$
$E(X_1 \mid Y_1)$	$r_{X_1,Y_1} = 0.981$	0.0005	0.022	0.029
$E(X_2 \mid Y_1)$	$r_{X_2,Y_1} = -0.990$	164.261	12.81	0.019
$E(X_4 \mid Y_1)$	$r_{X_4,Y_1} = -0.955$	186.137	13.64	0.079
$E(X_5 \mid Y_1)$	$r_{X_5,Y_1} = 0.923$	0.468	0.68	0.13
$E(X_3 \mid Y_2)$	$r_{X_3,Y_2} = 0.74$	49.875	7.06	0.022

The first PC is $Y_1 = 0.4998X_1 - 0.5054X_2 - 0.0159X_3 - 0.5030X_4 + 0.4914X_5$. With the exception of X_3 the coefficients of other variables are very close to each other in absolute terms. That means the contribution of these variables to Y_1 are almost the same. Correlation values between Y_1 and each of the variables X_1, X_2, X_3, X_4, X_5 are also very close to each other in absolute terms. However, MSE values are also reasonable, considering the fact that Y_1 represents 75% of total variation in the process.

The root mean square deviation (RMSD) is expressed as a percentage of the average for the variables $(\bar{X}_i)$ gives a good idea about the magnitude of the error committed in estimating data variables using PCs.

For instance, RMSED/$\bar{X}_1$=0.029 for the first model $E(X_1 / Y_1)$ in Table (6.11) which means the magnitude of the error in the estimation of the response X_1 by using the first PC is very small.

Chapter 7

CONCLUSION

The study concentrated on initially explaining the theoretical background on multivariate regression and PCA, together with some important theorems related to the topics. Establishing a relationship between the results of linear regression and PCA is attempted using a pilot data set.

On the other hand, it is very useful to realize the factors that have an impact on the accuracy of the linear regression model and avoiding big errors in the prediction process. Furthermore, focusing on the relationship between the PCs and the accuracy of the linear regression model is very functional.

In application chapter, by studying group of data and analyze it by applying the linear regression model, whether is simple or multiple (with different numbers of the predictors), some results were obtained as given under Chapter 6.

Some important factors that affect the linear regression model and the residuals are summarized as;

1. The level of the correlation between the dependent and independent variables.
2. Existence of missing observations in dataset.
3. Existence of repeated values in the data set.

Existence of missing and/or repeated values is a separate topic of research and is not dealt with, in this thesis. However, in selecting the pilot data set from the large data, only observations with no missing data and/or not pronounced repeat cases were included. In the regression analysis of the data careful examination of the correlations between the predictor and response variables was carried out to avoid high error margins.

In the application of PCA to the pilot data again the correlation matrix **R** between the variables used as predictor variables X_1, X_2, X_3, X_4, X_5, were used, since the use of covariance matrix will result in large discrepancies in estimation, due to big differences between the magnitude of the data under different variables. Following the determination of the PCs based on the **R** it is clear that the first two PCs represents 99% of total variation in the pilot data.

The main conclusion of this study was to use the idea of estimating the predictor variables X_1, X_2, X_3, X_4, X_5, now treated as response variables regressed on each PC. The correlation between each $X_i;\ i = 1,2,\ldots,5$ and the PCs is examined. Based on this the regression between each X_i and the first PC Y_1, and the regression $X_3 | Y_2$ are computed with very satisfactory error levels as given in Table 6.7.

REFERENCES

[1] Galton, F. (1989). Kinship and correlation. *Statistical Science*, 4(2), 81-86.

[2] Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot.

[3] Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium* (Vol. 7). Perthes et Besser.

[4] Gauss, C. F. (1821). *Anzeige, Theoria Combinationis observationum erroribus minimis obnoxiae*, pars prior. Carl Friedrich Gauss Werke, 4, 95-100.

[5] Yule, G. U. (1897). On the theory of correlation. *Journal of the Royal Statistical Society*, 60(4), 812-854.

[6] Pearson, K., Yule, G. U., Blanchard, N., & Lee, A. (1903). The law of ancestral heredity. *Biometrika*, *2*(2), 211-236.

[7] Fisher, R. A. (1922). The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, *85*(4), 597-612.

[8] Edwards, A. W. (2005). RA Fischer, statistical methods for research workers, (1925). In *Landmark Writings in Western Mathematics 1640-1940* (pp. 856-870).

[9] Ramcharan, R. (2006). Regressions: why are economists obsessed with them. *Finance and development*, *43*(1), 1-5.

[10] Pearson, K., Yule, G. U., Blanchard, N., & Lee, A. (1903). The law of ancestral heredity. *Biometrika*, *2*(2), 211-236.

[11] Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572.

[12] Kermack, K. A., & Haldane, J. B. (1950). Organic correlation and allometry. *Biometrika*, *37*(1/2), 30-41.

[13] York, D. (1966). Least-squares fitting of a straight line. *Canadian Journal of Physics*, *44*(5), 1079-1086.

[14] Ricker, W. E. (1973). Linear regressions in fishery research. *Journal of the fisheries board of Canada*, *30*(3), 409-434.

[15] Sprent, P., & Dolby, G. R. (1980). Query: the geometric mean functional relationship. *Biometrics*, *36*(3), 547-550.

[16] Laws, E. A. (1997). *Mathematical methods for oceanographers: An introduction*. John Wiley & Sons.

[17] Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572

[18] Fisher, R. A., & Mackenzie, W. A. (1923). Studies in crop variation. II. The manurial response of different potato varieties. *The Journal of Agricultural Science*, 13(3), 311-320.

[19] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6), 417.

[20] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika,* 28 (3/4), 321-377.

[21] Wold, H. (1966). Nonlinear Estimation by Iterative Least Squares Procedures in: David, FN (Hrsg.), Festschrift for J. *Neyman: Research Papers in Statistics, London*.

[22] Gnanadesikan, R. (2011). *Methods for statistical data analysis of multivariate observations* (Vol. 321). John Wiley & Sons.

[23] Golub, G. H., & Van Loan, C. F. (2012). *Matrix computations* (Vol. 3). JHU Press.

[24] Mandel, J. (1982). Use of the singular value decomposition in regression analysis. The American Statistician, 36(1), 15-24.

[25] Jolliffe, I. (2011). Principal component analysis. *In International encyclopedia of statistical science* (pp. 1094-1096). Springer, Berlin, Heidelberg.

[26] Shaw, P. J. (2009). *Multivariate statistics for the environmental sciences*. Wiley.

[27] Johnson, R. A., & Wichern, D. WD (2007). *Applied Multivariate Statistical Analysis*. Prentice Hall.

APPENDIX

Table 3.1: Row Data Residents, Income and Density

No	X_1	X_2	X_3	X_4	X_5
1	8	78	284	9.1	109
2	9.30000019	68	433	8.7	144
3	7.5	70	739	7.2	113
4	8.89999962	96	1792	8.9	97
5	10.1999998	74	477	8.3	206
6	8.30000019	111	362	10.9	124
7	8.80000019	77	671	10	152
8	8.80000019	168	636	9.1	162
9	10.6999998	82	329	8.7	150
10	11.6999998	89	634	7.6	134
11	8.5	149	631	10.8	292
12	8.30000019	60	257	9.5	108
13	8.19999981	96	284	8.8	111
14	7.9000001	83	603	9.5	182
15	10.3000002	130	686	8.7	129
16	7.4000001	145	345	11.2	158
17	9.60000038	112	1357	9.7	186
18	9.30000019	131	544	9.6	177
19	10.6000004	80	205	9.1	127
20	9.69999981	130	1264	9.2	179
21	11.6000004	140	688	8.3	80
22	8.10000038	154	354	8.4	103
23	9.80000019	118	1632	9.4	101
24	7.4000001	94	348	9.8	117
25	9.39999962	119	370	10.4	88
26	11.1999998	153	648	9.9	78
27	9.10000038	116	366	9.2	102
28	10.5	97	540	10.3	95
29	11.8999996	176	680	8.9	80
30	8.39999962	75	345	9.6	92
31	5	134	525	10.3	126
32	9.80000019	161	870	10.4	108
33	9.80000019	111	669	9.7	77
34	10.8000002	114	452	9.6	60
35	10.1000004	142	430	10.7	71
36	10.8999996	238	822	10.3	86
37	9.19999981	78	190	10.7	93
38	8.30000019	196	867	9.6	106
39	7.30000019	125	969	10.5	162
40	9.39999962	82	499	7.7	95
41	9.39999962	125	925	10.2	91
42	9.80000019	129	353	9.9	52
43	3.59999991	84	288	8.4	110
44	8.39999962	183	718	10.4	69
45	10.8000002	119	540	9.2	57
46	10.1000004	180	668	13	106
47	9	82	347	8.8	40
48	10	71	345	9.2	50
49	11.3000002	118	463	7.8	35
50	11.3000002	121	728	8.2	86
51	12.8000002	68	383	7.4	57
52	10	112	316	10.4	57
53	6.69999981	109	388	8.9	94

Table 4.2: Calculating the Parameters and Simple Linear Regression Equation

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
150	50	22500	2500	7500
150	61	22500	3721	9150
150	54	22500	2916	8100
155	54	24025	2916	8370
155	63	24025	3969	8765
155	59	24025	3481	9145
155	61	24025	3721	9455
160	68	25600	4624	10880
160	65	25600	4225	10400
175	77	30625	5929	13475
175	83	30625	6889	14525
175	72	30625	5184	12600
$\sum_{i=1}^{12} x_i$ =1915	$\sum_{i=1}^{12} y_i$ =767	$\sum_{i=1}^{12} x_i^2$ =306675	$\sum_{i=1}^{12} y_i^2$ =50075	$\sum_{i=1}^{12} x_i y_i$ =123365

Table 5.1: Raw Data for Responses of an Oxidizing Chemical Gas Multisensor Device

No	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$
1	2.6	150	11.9	13.6	48.9
2	2	112	9.4	13.3	47.7
3	2.2	88	9.0	11.9	54.0
4	2.2	80	9.2	11.0	60.0
5	1.6	51	6.5	11.2	59.6
6	1.2	38	4.7	11.2	59.2
7	1.2	31	3.6	11.3	56.8
8	1	31	3.3	10.7	60.0
9	0.9	24	2.3	10.7	59.7
10	0.6	19	1.7	10.3	60.2
11	1	14	1.3	10.1	60.5
12	0.7	8	1.1	11.0	56.2
13	0.7	16	1.6	10.5	58.1
14	1.1	29	3.2	10.2	59.6
15	2	64	8.0	10.8	57.4
16	2.2	87	9.5	10.5	60.6
17	1.7	77	6.3	10.8	58.4
18	1.5	43	5.0	10.5	57.9
19	1.6	61	5.2	9.5	66.8
20	1.9	63	7.3	8.3	76.4
21	2.9	164	11.5	8.0	81.1
22	2.2	79	8.8	8.3	79.8
23	2.2	95	8.3	9.7	71.2
24	2.9	150	11.2	9.8	67.6
25	4.8	307	20.8	10.3	64.2
26	6.1	401	24.0	9.6	67.8
27	3.9	197	12.8	9.1	64.0
28	1.5	61	4.7	8.2	63.4
29	1	26	2.6	8.2	60.8
30	1.7	55	5.9	8.3	58.5
31	1.9	53	6.4	7.7	59.7
32	1.4	40	4.1	7.1	61.8
33	0.8	21	1.9	7.0	62.3
34	1	10	1.1	6.1	65.9
35	0.6	7	1.0	6.3	65.0
36	0.8	17	1.8	6.8	62.9
37	1.4	33	4.4	6.4	65.1
38	4.4	202	17.9	7.3	63.1
39	3.1	208	14.0	13.2	41.7
40	2.7	166	11.6	14.3	38.4
41	2.1	114	10.2	15.0	36.5
42	2.5	140	11.0	16.1	34.5

43	2.7	169	12.8	16.3	35.7
44	2.9	185	14.2	15.8	37.0
45	2.8	165	12.7	15.9	37.2
46	2.4	133	11.7	16.9	34.3
47	3.9	233	19.3	15.1	39.6
48	3.7	242	18.2	14.4	43.4

Table 5.2: PCs Obtained from the Data in Table 5.1

No	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_5$
1	-51.2958	-6.55886	-0.71852	0.486873	-0.07102
2	-13.2883	-9.19938	-0.48149	0.286172	-0.14154
3	10.92332	-3.76615	-1.44858	-0.61339	0.123353
4	19.1301	1.972395	-2.32466	-0.92397	0.056385
5	48.20622	0.367261	-1.86954	-0.11958	0.014408
6	61.27249	-0.53828	-1.17857	0.610047	-0.04442
7	68.22813	-3.18514	-0.47894	0.873281	0.172366
8	68.38279	0.104629	-0.31623	1.075778	-0.02553
9	75.41526	-0.48042	0.112678	1.432308	0.059821
10	80.46869	-0.08528	0.498217	1.4733	-0.1452
11	85.49178	0.0198	0.602083	1.450214	-0.02292
12	91.30817	-4.61843	0.482642	1.418906	0.148377
13	83.38254	-2.27696	0.601319	1.273893	0.000521
14	70.37346	-0.27035	0.035788	0.665405	0.074105
15	35.06707	-1.13928	-1.68892	-1.23481	0.110077
16	12.16195	2.94574	-1.93635	-1.11286	-0.0484
17	22.25349	0.326824	0.001645	0.671607	0.018008
18	56.23252	-1.4166	-0.42989	0.002421	0.169158
19	38.61827	8.180982	-0.47473	0.994041	0.057752
20	36.87886	17.91423	-2.48843	0.055122	-0.06456
21	-63.9237	26.72317	-1.0652	1.508279	-0.21049
22	20.94877	21.84709	-3.38511	0.022975	-0.09596
23	4.677818	13.78364	-1.80796	0.859323	0.017674
24	-50.4828	12.56047	-0.56733	0.68511	0.057903
25	-207.802	15.42761	0.486806	0.134771	-0.19429
26	-301.597	22.92727	2.767571	1.587757	0.139016
27	-97.5887	11.05685	1.612268	0.445571	0.57609
28	38.5421	5.16245	1.344803	-0.02524	-0.02067
29	73.47937	1.209124	1.460391	-0.22801	-0.02553
30	44.26238	0.09448	0.921924	-1.58942	0.06085
31	46.27435	1.249953	0.685031	-2.31366	0.154229
32	59.47322	2.947638	1.695152	-1.40048	0.015653
33	78.59231	2.691527	2.433926	-0.55224	-0.15028

34	89.7634	5.971768	2.480882	-0.57658	-0.13831
35	92.71788	4.905603	2.363833	-0.68419	-0.21121
36	82.60945	3.230938	2.281146	-0.7858	-0.14735
37	66.57454	5.992453	1.14925	-1.8708	-0.03295
38	-102.913	10.72592	-0.3985	-4.29991	0.173653
39	-109.538	-11.09	2.00457	0.080806	-0.15887
40	-67.6461	-16.2441	1.383758	0.401827	0.067493
41	-15.7882	-20.4353	-0.36651	-0.45317	-0.0401
42	-41.8603	-21.5636	-0.07351	0.492976	0.183013
43	-70.8465	-19.204	-0.21888	0.700529	-0.01778
44	-86.8285	-17.1995	-0.24299	0.289267	-0.11809
45	-66.7921	-17.8679	-0.31095	0.50056	0.087007
46	-34.9407	-22.1494	-1.42233	0.259372	0.047786
47	-134.926	-12.6893	-1.50634	-1.60781	-0.18364
48	-143.667	-8.35297	-0.36307	-0.46784	-0.3053

Table 6.2: Row Data about the Heating Load Process

No	Y_1	X_1	X_2	X_3	X_4	X_5
1	21.33	0.98	514.50	294.00	110.25	7.00
2	28.28	0.90	563.50	318.50	122.50	7.00
3	27.30	0.86	588.00	294.00	147.00	7.00
4	24.93	0.82	612.50	318.50	147.00	7.00
5	37.73	0.79	637.00	343.00	147.00	7.00
6	29.40	0.76	661.50	416.50	122.50	7.00
7	10.90	0.74	68600	245.00	220.50	3.50
8	11.67	0.71	710.50	269.50	220.50	3.50
9	11.74	0.69	735.00	294.00	220.50	3.50
10	12.14	0.66	759.50	318.50	220.50	3.50
11	16.78	0.64	784.00	343.00	220.50	3.50
12	12.04	0.62	808.50	367.50	220.50	3.50
13	26.47	0.98	514.50	294.00	110.25	7.00
14	34.33	0.90	563.50	318.50	122.50	7.00
15	30.89	0.86	588.00	294.00	147.00	7.00
16	28.51	0.82	612.50	318.50	147.00	7.00
17	41.68	0.79	637.00	343.00	147.00	7.00
18	33.67	0.76	661.50	416.50	122.50	7.00
19	13.43	0.74	686.00	245.00	220.50	3.50
20	14.27	0.71	710.50	269.50	220.50	3.50
21	14.28	0.69	735.00	294.00	220.50	3.50
22	13.65	0.66	759.50	318.50	220.50	3.50
23	19.37	0.64	784.00	343.00	220.50	3.50
24	14.27	0.62	808.50	367.50	220.50	3.50
25	25.95	0.98	514.50	294.00	110.25	7.00
26	34.20	0.90	563.50	318.50	122.50	7.00
27	30.91	0.86	588.00	294.00	147.00	7.00
28	28.79	0.82	612.50	318.50	147.00	7.00
29	41.07	0.79	637.00	343.00	147.00	7.00
30	14.11	0.71	710.50	269.50	220.50	3.50
31	13.51	0.69	735.00	294.00	220.50	3.50
32	14.86	0.66	759.50	318.50	220.50	3.50
33	19.30	0.64	784.00	343.00	220.50	3.50
34	14.37	0.62	808.50	367.50	220.50	3.50
35	26.14	0.98	514.50	294.00	110.25	7.00
36	34.14	0.90	563.50	318.50	122.50	7.00
37	27.40	0.86	588.00	294.00	147.00	7.00
38	28.68	0.82	612.50	318.50	147.00	7.00
39	34.29	0.79	637.00	343.00	147.00	7.00
40	33.85	0.76	661.50	416.50	122.50	7.00
41	13.49	0.74	686.00	245.00	220.50	3.50
42	14.14	0.71	710.50	269.50	220.50	3.50

43	14.40	0.69	735.00	294.00	220.50	3.50
44	13.46	0.66	759.50	318.50	220.50	3.50
45	19.29	0.64	784.00	343.00	220.50	3.50
46	14.09	0.62	808.50	367.50	220.50	3.50
47	25.87	0.98	514.50	294.00	110.25	7.00
48	29.34	0.90	563.50	318.50	122.50	7.00
49	25.81	0.86	588.00	294.00	147.00	7.00
50	36.87	0.79	637.00	343.00	147.00	7.00

Table 6.8: Row Data PCs

No	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_5$
1	-316.2185	345.047	10.509	329.697	188.382
2	-347.573	374.247	10.627	361.302	206.178
3	-371.908	351.588	-3.233	382.481	213.553
4	-384.700	378.608	1.464	396.491	222.959
5	-397.486	405.625	6.1683	410.508	232.365
6	-398.729	482.322	29.430	417.354	243.801
7	-459.406	309.841	-37.323	459.062	245.085
8	-472.193	336.859	-32.620	473.078	254.490
9	-484.974	363.875	-27.911	487.101	263.896
10	-497.760	390.892	-23.208	501.118	273.302
11	-510.542	417.908	-18.499	515.141	282.707
12	-523.323	444.924	-13.791	529.165	292.113
13	-316.218	345.047	10.509	329.697	188.382
14	-347.573	374.247	10.627	361.302	206.178
15	-371.908	351.588	-3.233	382.481	213.553
16	-384.700	378.6081	1.464	396.491	222.959
17	-397.486	405.625	6.168	410.508	232.365
18	-398.729	482.322	29.430	417.354	243.801
19	-459.406	309.8417	-37.323	459.062	245.085
20	-472.193	336.859	-32.620	473.078	254.490
21	-484.974	363.875	-27.911	487.101	263.896
22	-497.760	390.892	-23.208	501.118	273.302
23	-510.542	417.908	-18.499	515.141	282.707
24	-523.323	444.924	-13.791	529.165	292.113
25	-316.218	345.047	10.509	329.697	188.382
26	-347.573	374.247	10.627	361.302	206.178
27	-371.908	351.588	-3.233	382.481	213.553
28	-384.700	378.608	1.464	396.491	222.959
29	-397.486	405.625	6.168	410.508	232.365
30	-472.193	336.859	-32.620	473.078	254.490
31	-484.974	363.875	-27.911	487.101	263.896
32	-497.760	390.892	-23.208	501.118	273.302
33	-510.542	417.908	-18.499	515.141	282.707
34	-523.323	444.924	-13.791	529.165	292.113
35	-316.218	345.047	10.509	329.697	188.382
36	-347.573	374.247	10.627	361.302	206.178
37	-371.908	351.588	-3.233	382.481	213.553
38	-384.700	378.608	1.464	396.491	222.959
39	-397.486	405.625	6.168	410.508	232.365
40	-398.729	482.322	29.430	417.354	243.801
41	-459.406	309.841	-37.323	459.062	245.085
42	-472.193	336.859	-32.620	473.078	254.490
43	-484.974	363.875	-27.911	487.101	263.896
44	-497.760	390.892	-23.208	501.118	273.302

45	-510.542	417.908	-18.499	515.141	282.707
46	-523.323	444.924	-13.791	529.165	292.113
47	-316.218	345.047	10.509	329.697	188.382
48	-347.573	374.247	10.627	361.302	206.178
49	-371.908	351.588	-3.233	382.481	213.553
50	-397.486	405.625	6.168	410.508	232.365