Credit Scoring Problem Based on Regression Analysis

Bashar Suhil Jad Allah Khassawneh

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
July 2014
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Elvan Yılmaz
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. Nazım 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. Ersin Kuset Bodur
Supervisor

Examining Committee

1. Prof. Dr. Rashad Aliyev
2. Asst. Prof. Dr. Ersin Kuset Bodur
3. Asst. Prof. Dr. Müge Saadetoğlu

ABSTRACT

This thesis provides an explanatory introduction to the regression models of data mining and contains basic definitions of key terms in the linear, multiple and logistic regression models. Meanwhile, the aim of this study is to illustrate fitting models for the credit scoring problem using simple linear, multiple linear and logistic regression models and also to analyze the found model functions by statistical tools.

Keywords: Data mining, linear regression, logistic regression.

ÖZ

Bu tez çalışması regresyon modelleri için açıklayıcı bilgiler, ayrica basit ve çoklu doğrusal regresyon modelleri ve linear lojistik regresyon modeller için temel tanımlar içermektedir. Aynı zamanda bu tezin amacı kredi sıralaması için basit, çoklu doğrusal regresyon ve linear lojistik modellemeleri kullanıp, uygun model bulmak ve bulunan model fonksiyonlarını istatistiksel yöntemlerle analiz etmektir.

Anahtar Kelimeler: Veri madenciliği, doğrusal regresyon, lojistik regresyon.

This thesis is dedicated to my Father, my Mother, my brother Ammar, and my sisters Lujain and Sewar.

Also I dedicate my thesis to the soul of my Grandfather.

And I would like to dedicate my thesis to my professors and to my country Jordan.

ACKNOWLEDGMENT

At first and foremost, thanks to our merciful God, for his assistance in my success reaching such sophisticated level of education, through my journey at the EMU. This thesis indicates a vast milestone, as a unique opportunity and experience in my life. As in these 2 years, I endured several hardships, in educational and other aspects. However living abroad in such an opportunity, was advantageous, since it assisted me in overcoming the faced obstacles. I would like to address my special thanks and sincerest gratitude, to everyone collaborated with me to fulfill this academic task. Especially I would like to thank my supervisor Assist. Prof. Dr. Ersin Kuset Bodur for her patience, wisdom, and knowledge attribution in aiding me through the accomplishment of my Master Thesis. In addition to that, I would like to acknowledge my special thanks to dozens of remarkable individuals, I encountered through my academic voyage, for their contribution in the accomplishment of my master thesis. Moreover I like to address my thanks to various computing techniques and programs, as they played an essential role in my data processing, easing the vast process of data management. Not neglecting the vital role of my parents, brother, sisters, and other family members which supported me through my studies in both moral and financial aspects. In addition to that my thanks will not be capable, of expressing my sincerest gratitude to my friends in Jordan, and those I met in Cyprus, they were friends and family at the same time.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Data mining is the calculation stage of the knowledge discovery in KDD process. The aim of data mining is to find out information within huge data and transform the information to build useful patterns for future use of technology, or science. Data mining is also known as analytic level of knowledge discovery in databases, (KDD); it finds appropriate information by examining the data.

Basically, KDD has five main stages: selection, pre-processing, transformation, data mining and evaluation of data set. On the other hand, data mining consists of three stages; in the first stage, the data is selected; cleaning, transforming are some benefits to apply in this stage. In stage two, the best method is selected since there are different techniques of data mining; the choice of model depends on the performance of data. After that, the model has been used to predict and explain the result of the unknown data in stage three, [1].

Mainly, we would like to emphasize data mining techniques in four categories: clustering, classification, association and regression methods. Clustering is known as unsupervised learning, a set of objects are given but classes are not predefined, the objects are partitioned into subclasses or groups, such that elements in a class have a common set of properties. Similarity between elements of same class is higher than objects of different classes.

Clusters can be represented using different algorithms the most used algorithm is *k-means algorithm, and also, C-means clustering, hierarchical clustering and HAC algorithm* can be used to define the clusters, 0.

In association rule frequent patterns, associations, correlations, or casual structures among set of items or objects are explored in transactional databases, relational databases or other information repository. Examples of algorithms which can be used to show association rules are *frequent pattern growth and Apriori algorithm*, 0.

Classification is called supervised learning; a set of objects is given with classes. It is a kind of predictive modelling; a training set is created, containing a set of attribute with the relevant outcome. An algorithm is used to find relationship between the attributes, that would make it possible to predict outcome, and then, the algorithm is given a data set not seen before, which contains the same set of attributes, except for the prediction attribute-not yet known. Examples of classification algorithms are *ID3 algorithm and C4.5 algorithm*, [1], [2].

Regression finds relationships between independent variables and dependent variables. Instead of predicting classes, we predict real-valued fields. Regression can be shown using linear regression, multiple regression, logistic regression or quadratic regression.

In the 1700s, Bayes` theorem, and in the 1800s regression analysis have been used to find useful information within data set. Later on, different techniques such as neural network, cluster analysis, genetic algorithms or decision trees have been applied.

The data which is used to mine information consists of a lot of observations called vectors. Sometimes the relationships between the dependent and independent variables of the vectors can be explained easily but sometimes it is more difficult to define the relationships of those variables. One of the tools that investigate the relationships between the variables is the regression analysis. Regression is a process in order to examine associations among the variables within the data set. Regression analysis uses statistical tools to figure out the relationship between dependent variable and independent variables, [2].

Regression models are simple linear or multiple linear or non-linear such that the linear model is one of the methodology to discriminate the relationships between dependent and independent variables. The simple linear regression has the form of $y = f(x) = b_0 + b_1 x$. Hence the multiple linear model having more than one independent variables has the form of $y = f(x_{i1}, x_{i2}, ..., x_{in}) = b_0 + b_1 x_{i1} + ... + b_n x_{in}$ where $i = 1, 2, ..., m$. Defining the regression model, our aim is to predict the new observations. There are different types of regression model such as linear, logistic, non-linear, log transformations.

In many research studies the logistic regression has been used for instance logistic regression can be applied very successfully in business and genetic applications to model the existing data. We realize that logistic regression can be applied when the dependent variable is binary, 0 or 1. In linear logistic regressions, the dependent variables can be categorical or continuous or interval variables.

Logistic regression is defined using the logit transformation of the dependent variable using S-shape curve, [4]. Actually logistic regression is defined by means of the logit transformation when the graph between the independent variable and $\Pr(Y=1|X=x)$ gives the S-shape curves. The logit transformation is $ln\left(\frac{\Pr(Y=1|X=x)}{1-\Pr(Y=1|X=x)}\right)$ and we use the logit function as dependent variable by the way the result of the transformation will be explained by linear function to construct the regression, [5].

First of all credit scoring is built in the 1960s, but widely credit scoring popularity increased in the 1975s when credit cards business area had become popular. Credit scoring is a tool which is used to evaluate the risk aspect of credit applications. In many applications the decision of credit scoring problems is based on the statistical tools, [16].

Different techniques such as discriminant analysis, probit analysis, logistic regression, linear programing, decision trees, neural networks and genetic algorithms have been offered to build the credit scoring problems.

Lately, lots of published papers focused on characteristics of regression models, for example, in [7], the author took the data from financial institution; the model is

developed using logistic regression, neural networks and genetic algorithms after that the capacities of these three models are measured and the aim of the paper is to search the effect of similarity metrics on performance of the used system, [16].

Additionally Euclidean distance, Manhattan distance and weights were used to construct linear and multivariable regression models. In [8], in order to evaluate credit scoring for small campanies logistic regression and multicriteria decision making are used and the author combined both methods to figure out efficient strategy for high capability. The performance of scoring models is discussed for credit scoring problem, and the goal was to develop the model for credit scoring, in [9].

In addition, the known classification algorithms such as logistic regression, discriminant analysis, k-nearest neighbor, neural networks and decision trees have been proposed for credit scoring data sets and also advanced classifiers are compared to increase the performance of credit scoring in this study, [10].

This thesis consists of five chapters, as well as these five chapters of this study are ordered as follows. Chapter 1 covers the introduction part. Fundamental definitions and principles of regression models and briefly logistic regression model descriptions are presented in Chapter 2. Simple linear and multiple linear regression problems of credit scoring are solved in Chapter 3. Finally, the credit scoring example using logistic regression is analyzed in Chapter 4.

Chapter 2

REVIEW OF REGRESSION MODELS and USEFUL DEFINITIONS

2.1 Review of Simple Linear Regression

We draw the scatter graph of all points of the data set to understand briefly the nature of the correlation between x and y . We may use this graph to supervise the relationship between two variables x and y , or between y and $\hat{y}$ or to discuss the quality of the model regression. Sometimes we try to understand the correlation between the variables x and y using the scatter graph.

To improve the performance of data mining techniques we may transform the data for the best discussions or results, so the measure values can be scaled to a range $[-1,1]$ by normalizing the data. We use the correlation perspective when scatter plot or especially covariance results are not sufficient to interpret the behavior of entire data determining whether the model is either linear or non-linear.

Different definitions can be used to measure the direction and the strength of the variables in the data. In this work, for the given data to talk about the relationship between variables and to define the direction or the strength of the variables, we will use two definitions; those two definitions are known as the covariance and the correlation coefficient.

When the scatter graph is not sufficient in the discussion, mostly it is not; we will have sufficient information about the direction, and the strength of the variables and the performance of the model by discussing the values of correlation and covariance.

Suppose there are n observations of the data, and also suppose $\bar{x} = \frac{1}{n}(\sum_{i=1}^{n} x_i)$ and $\bar{y} = \frac{1}{n}(\sum_{i=1}^{n} y_i)$, these values are called mean of x and y, respectively.

1. If $(x_i - \bar{x})(y_i - \bar{y})$ is positive, most of the points are in the first and third quadrants.
2. If $(x_i - \bar{x})(y_i - \bar{y})$ is negative, then many points of the data should be in the second and fourth quadrants.

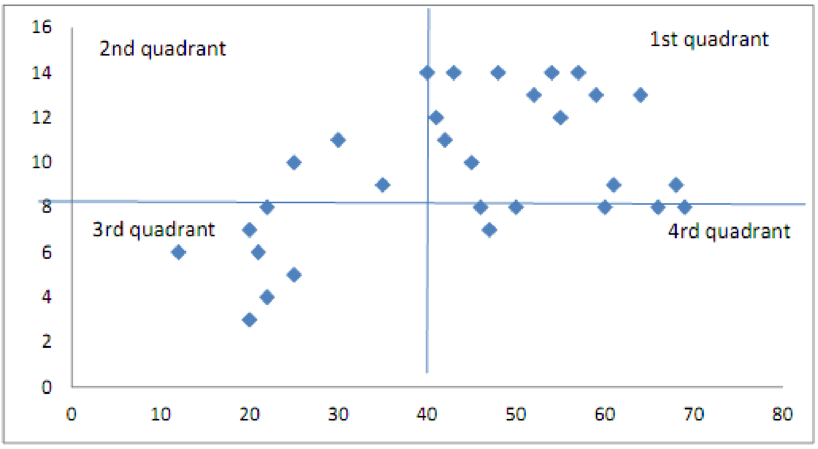

Figure 2.1: Graph for correlation

The following graph in figure 2.1 is the scatter plot of the data indicating the slope of the function that is mentioned in the above case. The covariance of the variables x and y is defined as

$$\text{Cov}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}\ . \tag{1}$$

In equation (1) n is the number of the observations. Meanwhile equation (1) indicates the direction of the linear function of x and y, by the way we may explain the sign of the slope of the line according to the result of equation 1.

There are the following two cases for covariance:

1. If $Cov(x,y)>0$, the graph moves from the left to the right, this means there is positive relationship between x and y.
2. If $Cov(x,y)<0$, the graph moves from the from the right to the left, so there is negative relationship between x and y.

But, sometimes we do not get appropriate information using the covariance between x and y. At that time also we discuss the correlation between x and y that is defined as

$$\text{Cor}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=0}^{n}(\text{y}_i-\bar{y})^2}}\ . \tag{2}$$

The measure values of entire data are scaled to a range [-1,1] by normalizing equation (1). Equation (2) is the covariance of the normalize variables of x and y where $-1\le Cor(x,y)\le 1$.

1. If $Cor(\mathrm{x},\mathrm{y})$ is around 1, that means there is strong positive linear relationship between x and y.
2. If $Cor(\mathrm{x},\mathrm{y})$ is around -1, then there is strong negative linear relationship between x and y.
3. If $Cor(\mathrm{x},\mathrm{y})$ is around 0, then there is non-linear relationship between x and y.

Consider the data in table 2.1 in the xy-plane. We assume that there exits n observations in data set.

Table 2.1: Data set with n observations

x_i	x_1	x_2	$\cdots$	x_n
y_i	y_1	y_2	$\cdots$	y_n

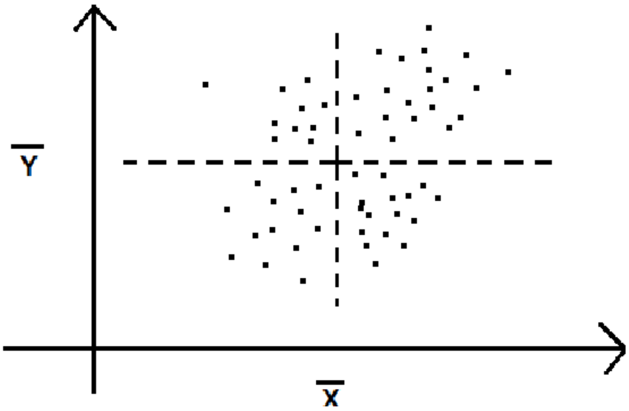

Figure 2.2: Graph of data set

where $i=1,2,...,n$.

And, also we suppose that the scatter plot of the dataset is linear. We write a linear function that fits the given data in terms of the independent variable x and dependent variable y as $y = \alpha + \beta x + \varepsilon$, where α and β are the y-intercept and the slope of the linear function, respectively and ε is an error. Equation (1) can be applied to the linear equation $y_i = \alpha + \beta x_i + \varepsilon_i$, where $i = 1, 2, ..., n$ assuming that the scatter graph of data that looks like figure 2.2.

In our calculations, we prefer to apply the standard least squares method to estimate the values of the parameters α and β in order to construct the linear function. Constructing the function, $y_i = f(x_i) = \alpha + \beta x_i$, our aim is to minimize the sum of the squares of the vertical distances from each of the point of data to the function $y_i = f(x_i) = \alpha + \beta x_i$, called regression line.

$$\text{Sum of the squares} = S = \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2 = \sum_{i=1}^{n} \varepsilon_i^2 \, .$$

If $S = 0$, the sum of the squares is minimized, the regression line fits data perfectly. Using the regression equation, we write

$$\varepsilon_i = y_i - \alpha - \beta x_i, \; i = 1, 2, ..., n$$

$$S = \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2 = \sum_{i=1}^{n} \varepsilon_i^2 \tag{3}$$

$$= \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \, .$$

The minimum value(s) of equation (3) lies at the critical points, so we take the first partial derivatives of $S = S(\alpha, \beta)$ with respect to the unknowns α and β.

Consequently, the values of α and β are derived as

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4}$$

and

$$\alpha = \bar{y} - \beta.\bar{x} \tag{5}$$

where the point $(\bar{x}, \bar{y})$ lies on the function. More useful descriptions to equations 4 and 5 will be presented in the Appendix.

The function, $y_i = f(x_i) = \alpha + \beta x_i$, called the simple linear regression line passing through the center of gravity of the dataset, its graph is shown in figure 2.3.

2.2 Interpreting Simple Linear Regression Model

The quality of the model is considered when the data is fitted by the model, because it is not guarantee of the regression model to be useful. There are various ways to discuss the quality of the model in the literature.

We would like to rank these ways as follows:

1. Using the assumptions
2. By the scatter plot: if $Cor(\mathrm{x},\mathrm{y})$ is around 1 or -1, there exists strong linear (positive or negative) relationship between x and y where $-1 \le Cor(\mathrm{x},\mathrm{y}) \le 1$.
3. Examine the scatter plot between y and the expected value of y, which is $\hat{y}$, the set of points should be closer, this means we calculate $Cor(\mathrm{y},\hat{\mathrm{y}})$ where $0 \le Cor(\mathrm{y},\hat{\mathrm{y}}) \le 1$.
4. Using the square of the correlation coefficient.

In our calculations, also R square test has been used in order to test the quality of the regression model, [11].

R **-squared test:** this test measures the correlation between the variables x and y of regression. Let (x_k, y_k) be any point of the data, $i = 1,...,n$.

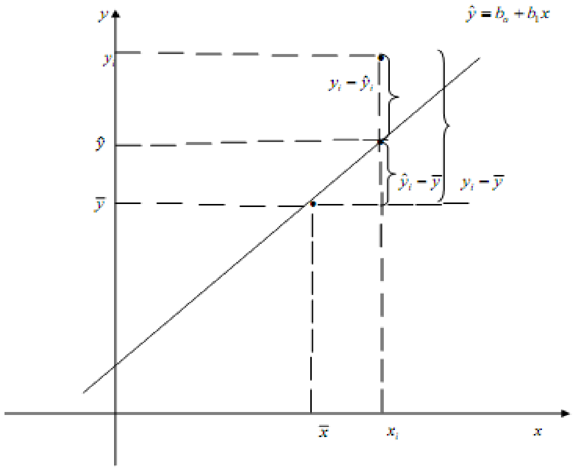

Figure 2.3: SSR, SSE, SST

We may write the following sums considering figure 2.3.

$$SST = \sum_{i=1}^{n} (\mathrm{y}_i - \bar{\mathrm{y}})^2 = \text{the sum of squares deviations of } y \text{ from } \bar{y}$$

$$SSR = \sum_{i=1}^{n} (\hat{\mathrm{y}}_i - \bar{\mathrm{y}})^2 = \text{the sum of squares of regression}$$

$$SSE = \sum_{i=1}^{n} (\mathrm{y}_i - \hat{\mathrm{y}})^2 = \text{the sum of square errors.}$$

In figure 2.3 the distance from y to $\hat{y}$ can be defined using the following calculation. Let $y_i = y_i$. Then we write it as $y_i = \hat{y}_i + (y_i - \hat{y}_i)$, then subtract $\bar{y}_i$ from each side to get $y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$, and take square of two sides, but cancel the term from that equation, $2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)$, this proof is given by Draper and Smith, for more information see [11].

$$\sum_{i=1}^{n}(\mathrm{y}_i-\bar{\mathrm{y}})^2=\sum_{i=1}^{n}(\hat{\mathrm{y}}_i-\bar{\mathrm{y}})^2+\sum_{i=1}^{n}(\mathrm{y}_i-\hat{\mathrm{y}})^2 .$$

This equation can be written in this form $SST = SSR + SSE$.

$$R^2=\frac{\text{explained variation}}{\text{total variation}}=\frac{SSR}{SST}=\frac{\sum_{i=1}^{n}(\hat{\mathrm{y}}_i-\bar{\mathrm{y}})^2}{\sum_{i=1}^{n}(\mathrm{y}_i-\bar{\mathrm{y}})^2}$$

$$=\frac{SST-SSE}{SST}=1-\frac{SSE}{SST}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(\mathrm{y}_i-\hat{\mathrm{y}})^2}{\sum_{i=1}^{n}(\mathrm{y}_i-\bar{\mathrm{y}})^2}$$

where $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x$ and $i=1,\dots,n$.

The value of R^2 is used to measure the fit of the regression model where $-1\le R^2\le 1$. When the value of R^2 is equal to 1 this means the regression model is perfect. Then this implies that the value of $r=\sqrt{R^2}$ is between 0 and 1.

Analysis of Variance Table: ANOVAs table for simple linear regression can be seen in table 2.2. In ANOVA table, degree of freedom, df, sum of squares, SS, mean square, MS, F-ratio and P-value are presented. In table 2.2, m and n mean the number of predictor variables and the number of observations, respectively.

Definition: The goodness of the data is calculated by R^2 which is equal to $R^2 = \dfrac{SSR}{SST}$.

Definition: The Adjusted R^2 should be less than R^2 and it is equal to

$$R^2 = 1 - \frac{(1 - R^2)(n \text{ - } m)}{n - m - 1}.$$

Table 2.2: ANOVA table with one independent variable

Source of Variation	**Degree of freedom**	**Sum of square**	**Mean Square**	**F**
Regression	m	$= \sum_{i=1}^{n} (\widehat{y} - \overline{y})^2$	$MSR = \dfrac{SSR}{m}$	$F = \dfrac{MSR}{MSE}$
Error	$n - m - 1$	$= \sum_{i=1}^{n} (y - \overline{y})^2$	$MSE = \dfrac{SSE}{n - m - 1}$	
Total	$n - 1$	$= \sum_{i=1}^{n} (\widehat{y} - \overline{y})^2$		

Definition: Estimation of Standard Error. The standard deviation of the variation of n observations according to the regression model is calculated by

$$S_e = \sqrt{\frac{SSE}{n - m - 1}}$$

where m is the number of the predictor (independent) variables.

Definition: The standard deviation of th**e** slope for regression model is defined by

$$S_{b_1} = \frac{S_e}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \frac{\sqrt{\frac{SSE}{n-m-1}}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} .$$

Definition: Confidence Interval for slope. A confidence interval for the slope, b_1, is defined as:

$$b_1 \mp t(n-2, \frac{1}{2}\alpha).S_{b_1} .$$

In this definition, the standard error is calculated by $S_{b_1} = \frac{S_e}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$, and n is the number of the observations. t is obtained from the t-distribution table and $\frac{\alpha}{2} = \frac{1 - \text{percentage of confidence}}{2}$.

Similarly, the confidence interval for b_0 is calculated by

$$b_0 \mp t(n-2, \frac{1}{2}\alpha).S_{b_0} .$$

The standard error of b_0 is calculated by $S_{b_0} = \frac{S_e.\sqrt{\sum_{i=1}^{n} x_i{}^2 .}}{\sqrt{n}.\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$, and n is the number of the observations, t is obtained from the t-distribution table.

2.3 Review of Simple Linear Regression Model Using Matrix Form

Let x be an input and y be output variables. And, suppose there are n observations in data set such that (x_i, y_i) represents the $i-$th observation. As it is defined in section 2.1 the linear model is given as $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1,...,n$. The matrix form of the linear function $y_i = \alpha + \beta x_i + \varepsilon_i$ according to n observations can be written as the matrix form:

$$y_i = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \alpha_0 + \beta_1 x_1 \\ \vdots \\ \alpha_0 + \beta_1 x_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{6}$$

By writing above equation in the matrix form we get

$$\hat{y}_i = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} . \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{7}$$

where $b = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$, $\hat{y} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix}$, and $x = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ are $2x1$, $nx1$ and $nx2$ matrices, respectively.

If the matrix x is mxn and b is in R^n, every least square solution of the simple linear (or multiple linear) systems $y = x.b$ satisfies the equation $x^T x.b = x^T .y$, and it has a unique solution $b = (\mathrm{x^T x})^{-1}.\mathrm{x^T y}$, [11], [17].

The equation (2) can be written in the following matrix form

$$\underbrace{\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}}_{\Downarrow} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

$$\begin{pmatrix} 1+1+\cdots+1 & x_1 + x_2 + \ldots + x_n \\ x_1 + x_2 + \ldots + x_n & x_1^2 + x_2^2 + \ldots + x_n^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} y_1 + y_2 + \ldots + y_n \\ x_1 y_1 + x_2 y_2 + \ldots + x_n y_n \end{pmatrix}. \tag{8}$$

By the way matrix operations can be used to put equation (3) into the product form

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \tag{9}$$

and

$$x^T x.b = x^T .y\,.$$

Now our aim is to find the coefficients using the matrix algebra, in other words our aim is to find b for this purpose the matrix $x^T .x$ must be invertible. For the invertibility of the matrix $x^T x$, we show that $\det(x^T x) \neq 0$.

First, in order to calculate the determinant of $x^T .x$, we decompose the matrix $x^T .x$ as:

$$x^T x = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} 1+\cdots+1 & x_1+\cdots+x_n \\ x_1+\cdots+x_n & x_1^2+\cdots+x_n^2 \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}. \tag{10}$$

Then the determinant of 2x2 matrix is calculated by

$$\det(x^T x) = n\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i . \sum_{i=1}^n x_i .$$

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

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

$$= n\left(\sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + n.\overline{x}.\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

$$= n\left(\sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + n.\overline{x}.\overline{x}\right)$$

$$= n\left(\sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \overline{x}.\overline{x}\right)$$

$$= n\sum_{i=1}^{n} (x_i^2 - 2\overline{x}x_i + \left(\overline{x}\right)^2)$$

$$= n.\sum_{i=1}^{n} (x_i - \overline{x})^2 \neq 0 \qquad (11)$$

then

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 \neq 0 \text{ since } n \neq 0 .$$

If $\det(x^T x)$ is different from zero, the matrix $x^T .x$ is invertible and there is a unique solution to equation (3). The unique solution to equation (3) is obtained by multiply both sides of that equation by $(x^T x)^{-1}$ from the left side $(x^T x)^{-1}.x^T x.\mathrm{b} = (x^T x)^{-1} x^T .y$ where

$$\left(x^T x\right)^{-1} x^T x. = \mathrm{I}_{2x2} ,$$

then $b = (x^T x)^{-1} x^T .y$.

The inverse of the matrix $x^T x$ is equal to

$$\left(x^T x\right)^{-1} = \frac{1}{\det(x^T x)} Adj(x^T x) \; .$$

And this equation becomes

$$\left(x^T x\right)^{-1} = \frac{1}{\mathrm{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}. \qquad (12)$$

This inverse matrix in equation (6) is derived from the equations (4) and (5), then we get

$$= \begin{pmatrix} \dfrac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2} & -\dfrac{\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ -\dfrac{\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n}(x_i - \overline{x})^2} & \dfrac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \end{pmatrix}.$$

This equation is substituted into equation (4) to get the coefficients of the linear regression model, [11].

$$b = \begin{pmatrix} \dfrac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2} & \dfrac{-\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ \dfrac{-\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n}(x_i - \overline{x})^2} & \dfrac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$= (A^T A)^{-1} A^T y.$$

Definition 2.3.1 For the matrix form, the variance of the matrix b is

$$V(b) = V\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}\right) = \begin{pmatrix} V(\alpha) & Cov(\alpha,\beta) \\ Cov(\alpha,\beta) & V(\alpha) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 \sum_{i=1}^{n} x_i^2 & \dfrac{-\bar{x}\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ \dfrac{-\bar{x}\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} & \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \end{pmatrix}.$$

Definition 2.3.2 Calculation of covariance between the estimated coefficients α and β in the matrix form is

$$Cov(b_0, \mathrm{b}_1) = -\bar{x}\left(\frac{n}{\sum_{i=1}^{n} x_i^{\,2}} \right)^{\frac{1}{2}} .$$

Definition 2.3.3 Calculation of variance of $\hat{y}$ in the matrix form is

$$V(\hat{\mathrm{y}}_0) = \begin{pmatrix} 1 & x_0 \end{pmatrix} . \begin{pmatrix} V(\alpha) & Cov(\alpha,\beta) \\ Cov(\alpha,\beta) & V(\alpha) \end{pmatrix} . \begin{pmatrix} 1 \\ x_0 \end{pmatrix}$$

$$= x_0^T.(x^T x)^{-1}.\sigma^2 x_0$$

$$= x_0^T.(x^T x)^{-1}.x_0\sigma^2$$

Where $\sigma^2 = s^2$.

Table 2.3: ANOVA table for simple linear regression in matrix form

Source of Variation	**Degree of freedom**	**Sum of squares**	**Mean Square**	***F***
Regression	m	$b^T x^T y - n\bar{y}^2$	$MSR = \dfrac{SSR}{m}$	$F = \dfrac{MSR}{MSE}$
Error	$n-m-1$	$y^T y - b^T x^T y$	$MSE = \dfrac{SSE}{n-m-1}$	
Total	$n-1$	$y^T y - n\bar{y}^2$		

Now, we would like to introduce more useful definitions to evaluate confidence intervals of regression function in matrix form. Let C_{jj} be the j-th diagonal entry in the inverse matrix $(X^T X)^{-1}$.

Standard error of b_j is identical to $S_e(b_j) = \sqrt{MSE.c_{jj}}$.

100% $1-\alpha$ the confidence interval for b_j is evaluated by $b_j \pm t.\sqrt{MSE.c_{jj}}$.

2.4 Review of Multiple Linear Regression Model

We use the matrix algebra, $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \varepsilon_i$, $i = 1,...,n$, to define the multiple regression models with m independent variables and a single dependent variable y that is

$$y_i = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \alpha + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_n x_{1m} \\ \vdots \\ \alpha + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_n x_{nm} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The analysis of multiple regressions is similar to the matrix form of the linear regression model.

In Chapter 3, we will also use the matrix algebra to form the regression function and also the definitions will be used that are presented in section 2.3.

Also, matrix algebra will be used to discuss the correlation between the variables, to form ANOVA table, to interpret the coefficients of multiple regression models, [12].

2.5 Review of Logistic Regression Model

In this section we will present logistic regression model and in chapter 4 we will introduce the example of the logistic regression model. In order to introduce the logistic regression model we suppose that the variable x is an independent variable and y is a dependent variable but y should be binary variable, 0 or 1.

Let $\pi = \Pr(Y = 1 | X = x)$. The relationship between the probability π and input variable x can be represented by the logistic function. The graph of π with respect to x is the S -shape curve that is non-linear, [5].

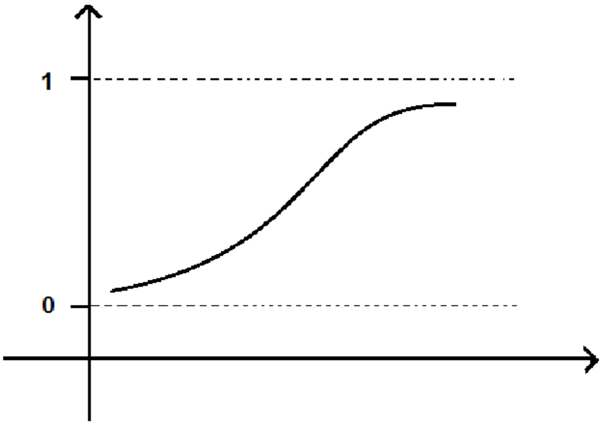

Figure 2.4: S-shape curve

The regression model is formed using the S-shape curve:

$$\pi = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{13}$$

Using the equation (1), we may write

$$1-\pi = 1-\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{14}$$

$$=\frac{1}{1+e^{\beta_0+\beta_1 x}}$$

Dividing equation (1) by equation (2), we obtain

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 x} \tag{15}$$

And, taking logarithm of both sides to the base e of equation (3) we find

$$ln\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1 x\,.$$

The expression $ln\left(\frac{\pi}{1-\pi}\right)$ is called the logit transformation.

The logit transformation is used for the logistic regression to determine whether the model fits the data or not. Also, the ratio $\frac{\pi}{1-\pi}$ is known as the Odds ratio where $\pi = Pr\,(Y=1|X=x)$ and $1-\pi=\Pr\,(Y=0|X=x)$.

Usually, this function

$$L(\mathrm{x})=ln\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1 x \qquad (16)$$

is used to fit the data when the dependent variable is binary variable (categorical variable) with one or more independent variables (categorical or interval).

In general, in order to discover the coefficients the maximum likelihood is used. We may display equation (4) by $\ln\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k$ with k independent variables, $x_1, x_2, ..., x_k$, since the dependent variable in data is the binary output, 0 or 1, [1].

To find the estimated coefficients the maximum likelihood estimation is used instead of usual way i.e. the least square estimation.

The likelihood function can be tested for significance of the model that is defined as, [14],

$$\ell(\beta|x)=\prod_{i=1}^{n}[\pi(\mathrm{x}_i)]^{y_i}[1-\pi(x_i)]^{1-y_i}.$$

Then, the coefficients of the logistic regression function may be calculated by taking the partial derivatives of the log likelihood function which is equal to

$$ln(\ell(\beta|x))=\sum_{i=1}^{n}[y_i\ln(\pi(\mathrm{x}_i))+(1-\mathrm{y}_i)\ln(1-\pi(x_i))].$$

Analysis of Logistic Regression Model:

In order to analyze the logistic regression model the value of deviance can be calculated which is equal to $D=-2\ln\sum_{i=1}^{n}[y_i\ln(\frac{\pi_i}{y_i})+(1-y_i)\ln(\frac{1-\pi_i}{1-y_i})]$. To decide whether the independent variable is significant or not, the value of G is obtained and it is equal to

$$G=2\{\sum_{i=1}^{n}[y_i\ln(\pi_i)+(1-y_i)\ln(1-\pi_i)]-$$

$$[\sum y_i.\ln(\sum y_i)+\sum(1-y_i).\ln(\sum(1-y_i)-\mathrm{nln(n)}]\}.$$

Sometimes Wald test is used to determine whether the independent variable is significant or not, we assume that $b_1=0$, the ratio equals to

$$Z_{Wald}=\left(\frac{b_1}{Se(b_1)}\right)^2$$

such that $Se(b_1)$ is the standard error. For the logistic regression model $100(1-\alpha)\%$ confidence interval is formed as $b_0\mp z.Se(b_0)$ and $b_1\mp z.Se(b_1)$ where z is the z-critical value.

If the confidence interval does not consist of one, we assume that $b_1 \neq 0$, this means the corresponding coefficient is significant. Most of the time confidence intervals are more powerful than hypothesis tests, [3].

Chapter 3

MODELLING SCORING DATA USING LINEAR and MULTIPLE LINEAR REGRESIONS

3.1 Simple Linear Regression Model for Credit Scoring Data

In this chapter our aim is to discuss three different data sets by using simple linear and multiple linear regression models. For this purpose we will use the definitions to interpret the regression functions in which they have been presented in previous sections.

Problem 1:

The credit data set consists of two variables; the independent variable x represents the net income and the dependent variable y represents loan amount of each customer. Our aim is to describe the simple linear regression model between two variables. We use the least square estimate in order to find the values of coefficients of regression model. In the first part of solution, we will construct the simple regression model, and in the second part we will analyze the regression model using the statistical tools. Table 3.1 is the scatter plot of data set with 100 observations.

Table 3.1: Data set of problem 1

Number of observations	x =net income	y = loan amount
1.	1073	3000
2.	893	3000
3.	664	6000
⋮	⋮	⋮
98.	1089	9500
99.	1987	10000
100.	461	4000

The scatter graph of credit scoring data is presented in Figure 3.1.

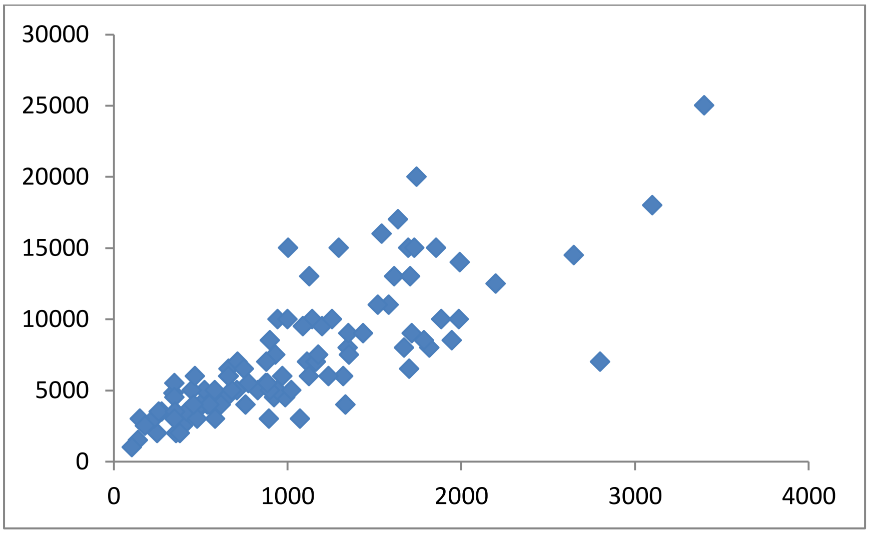

Figure 3.1: Scatter graph of problem 1

The coefficients of simple linear regression function have been evaluated using the results in table 3.2:

$$b_1 = \frac{\sum_{i=1}^{100} x_i - \bar{x} \quad y_i - \bar{y}}{\sum_{i=1}^{100} x_i - \bar{x} \;^2} = \frac{227792790}{42616949} = 5.345121961$$

and

$$b_0 = 7313.000 + (1045.7) \times (5.345121961) = 1723.605966 .$$

And the linear regression model becomes $\widehat{y} = 1723.605966 + 5.345121961x$.

Table 3.2: Calculations of coefficients

Number of observations	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$x_i - \bar{x} \;^2$	$x_i - \bar{x} \quad y_i - \bar{y}$
1.	1073	3000	27	-4318	745.29	-117881.4
2.	893	3000	-153	-4318	23317.29	659358.6
⋮	⋮	⋮	⋮	⋮	⋮	⋮
98.	1089	9500	43	2182	1874.89	94480.6
99.	1987	10000	941	2681	886045.69	2524566.6
100.	461	4000	-585	-3318	341874.09	1940034.6
SUM:	1045.7	7313	0	0	42616949.00	227792790

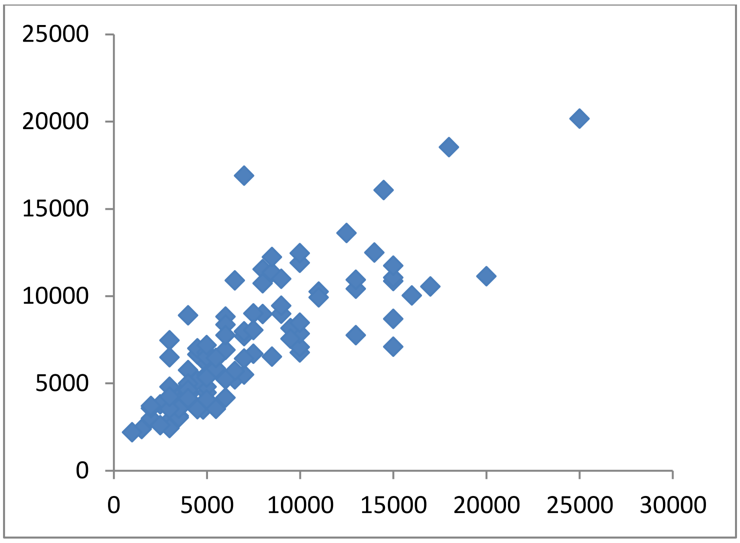

Figure 3.2: Scatter graph of y w.r to $\hat{y}$

Analysis of Regression Model The regression model of the data is $\hat{y} = 1723.605966 + 5.345121961x$. Now, we will discuss the reliability of the model that is obtained by statistical calculations. By the way table 3.3 is obtained calculating the values of SSR, SSE and SST for further investigations.

Table 3.3: Sum of squares

$\hat{y}$	SSR	SSE	SST
7458.92183	19858.96221	19881983.89	18645124
6496.799877	674369.6417	122227609.38	18645124
$\vdots$	$\vdots$	$\vdots$	$\vdots$
7544.443782	51276.78619	3824200.124	4761124
12344.3633	25264328.05	5496039.294	7193124
4187.70719	9798733.076	35233.98919	11009124
	$\sum_{i=1}^{n=100} (\hat{y}_i - \bar{y})^2$ $= 1217582744$	$\sum_{i=1}^{n=100} (y_i - \hat{y})^2$ $= 760712855.7$	$\sum_{i=1}^{n=100} (y_i - \bar{y})^2$ $= 1978295600$

Quality of Coefficients: ANOVA TABLE (Analysis of Variance):

We prefer to discuss significant of the regression line using the variance. With this objective the ANOVA table has been created evaluating the following concepts. All the results of the calculations can be seen in table 3.4.

$$MSR = \frac{1217582744}{1} = 1217582744,\ MSE = \frac{760712855.7}{98} = 7762376.078.$$

Table 3.4: Anova table for simple linear regression

Source	Degree of freedom	Sum of Squares (SS)	Mean of Squares (MS)	Significant F
Regression	1	1217582744	1217582744	156.8569639
Error	98	760712855.7	7762376.078	
Total	99	1978295600		

The value of R^2 is $R^2 = \frac{SSR}{SST} = \frac{1217582744}{1978295600} = 0.615470582$, then

$r = \sqrt{0.615470582} = 0.784519332 \approx 0.78$, and this result shows the strong positive correlation between the dependent and independent variables, [12].

Assessment of Standard Error: The value of standard error is

$$S_e = \sqrt{\frac{SSE}{n-m-1}} = \sqrt{\frac{760712855.7}{98}} = \sqrt{7762376.078} = 2786.104104,$$

the standard deviation of y is $\sqrt{1978295600/99} = 4470.210715$, the prediction error is reduced from 4470.210715 to 2718.104104.

Adjusted R-squared: $=1-\dfrac{(1-R^2)(n-1)}{n-m-1}$

$$=1-\frac{(1-0.615470582)\times 99}{98}=0.611546813.$$

And the result of adjusted R-squared is less than the value of R-squared.

Table 3.5: Regression statistics

Regression Statistics	
Multiple R	$r=0.784519332$
R **-Square**	$R^2=0.615470582$
Adjusted R **-squared**	0.611546813
Standard error	$S_e=2786.104104$
The number of observations	$n=100$

Confidence Intervals: in order to obtain the confidence intervals, first of all the coefficients errors are evaluated then using them we get the confidence intervals for both coefficients as follows:

Confidence Interval for b_1 **:**

$$S_{b_1}=\frac{S_e}{\sqrt{\sum_{i=1}^{n}(\mathrm{X}_i-\bar{X})^2}}=\frac{\sqrt{\dfrac{SSE}{n-m-1}}}{\sqrt{\sum_{i=1}^{n}(\mathrm{X}_i-\bar{X})^2}}=\frac{2786.104104}{\sqrt{42616949}}=0.426782067.$$

Confidence interval for slope : $b_1 \pm t_{\frac{\alpha}{2},n-2}.S_{b_1}$

Upper limit = $5.345121961+(1.984)\times(0.426782067)=6.191857582$

Lower limit = $5.345121961-(1.984)\times(0.426782067)=4.49838634$

95% confidence level is $4.49838634 < b_1 < 6.191857582$.

The slope of the regression line is between those two limits, for every additional increase of x the value of y will also increase between 4.49838634 and 6.191857582 with 95% confidence.

Confidence Interval for b_0:

$$S_{b_0}=\frac{S_e}{\sqrt{\sum_{i=1}^{n}(\mathrm{X}_i-\bar{X})^2}}\times\sqrt{\frac{\sum_{i=1}^{n}X_i^2}{n}}=\frac{\sqrt{\frac{SSE}{n-m-1}}}{\sqrt{\sum_{i=1}^{n}(\mathrm{X}_i-\bar{X})^2}}\times\sqrt{\frac{\sum_{i=1}^{n}X_i^2}{n}}$$

$$=(0.426782067)\times(1232.74409)$$

$$=528.7635294.$$

Confidence interval for intercept : $b_0 \pm t_{\frac{\alpha}{2},n-2}.S_{b_0}$

Upper limit = $1723.605966+(1.984)\times(528.7635294)=2772.672808$

Lower limit = $1723.605966-(1.984)\times(528.7635294)=674.5391233$

For 95% confidence level, $597.9969373 < b_0 < 2635.084977$.

Test: for $H_0: b_1=\beta_{10}$, $H_1: b_1 \neq \beta_{10}$. Assume that $H_0:\beta_{10}=0$. $t=\frac{b_1-\beta_{10}}{Se(b_1)}$

Then $t=\frac{b_1-0}{Se(b_1)}$, and $t=\frac{b_1-0}{Se(b_1)}=\frac{5.345121961}{0.426782067}=12.5242422$

Since $|t|=12.5242422 > f(98,0.05)=1.984$, reject $\beta_{10}=0$.

And also for $H_0: b_0=\beta_{10}$, $H_0: b_0 \neq \beta_{10}$

Assume that $H_0 : \beta_{10} = 0$. $t = \dfrac{b_1 - \beta_{10}}{Se(b_1)}$

Then $t = \dfrac{b_1 - 0}{Se(b_1)}$, $t = \dfrac{b_0 - 0}{Se(b_0)} = \dfrac{1723.605966}{528.7635294} = 3.25969147$

Since $|t| = 3.25969147 > f(98, 0.05) = 1.984$, reject $\beta_{10} = 0$.

Table 3.6: Confidence intervals

	Coefficients	**Standard Error**	***t*-statistics**	***P*-value**	**95% Lower**	**95% Upper**
b_0	1723.605966	528.7635294	3.25969147	0.0021735	674.5391233	2772.672808
b_1	5.345121961	0.426782067	12.5242422	3.03×10^{-23}	4.49838634	6.191857582

Table 3.7: Actual and predicted amounts

Observations	x = Age	y = Actual Amount	$\widehat{y}$ = Predicted Amounts
1.	1073	3000	7466.847508
2.	893	3000	6485.435412
3.	664	6000	5236.861135
⋮	⋮	⋮	⋮
98.	1089	9500	7554.084138
99.	1987	10000	12450.24004
100.	461	4000	4130.046383

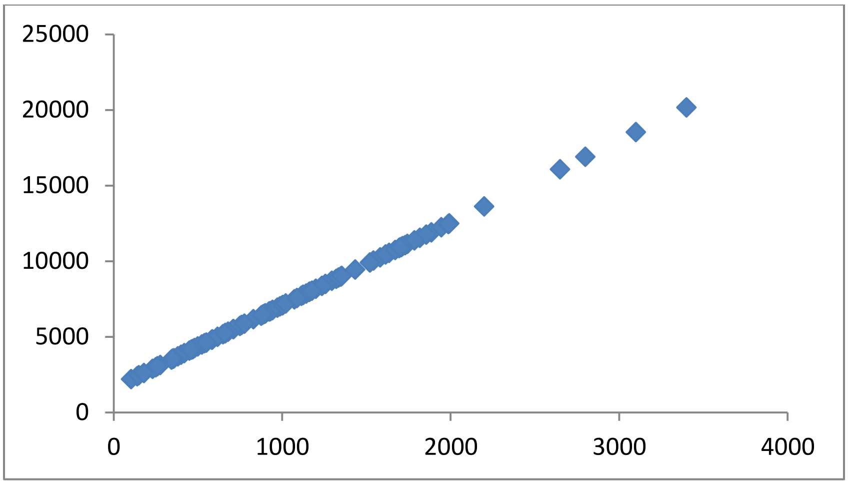

Figure 3.3: Graph of x w.r.t. to $\hat{y}$

Table 3.8 Confidence intervals of problem 1

Observations	y = **Actual Amount**	$\hat{y}$ = **Predicted Amount**	**Error**	**Lower bound %95**	**Upper bound %95**
1.	3000	7466.847	-4458.922	7318.3404	7599.024
2.	3000	6485.435	-3496.800	6203.8164	6789.384
3.	6000	5236.861	727.233	4785.8942	5759.342
⋮	⋮	⋮	⋮	⋮	⋮
98.	9500	7554.084	1955.556	7417.4092	7670.992
99.	10000	12450.240	-2344.363	12977.6456	11710.2
100.	4000	4130.04	-187.707	3528.9588	4846.248

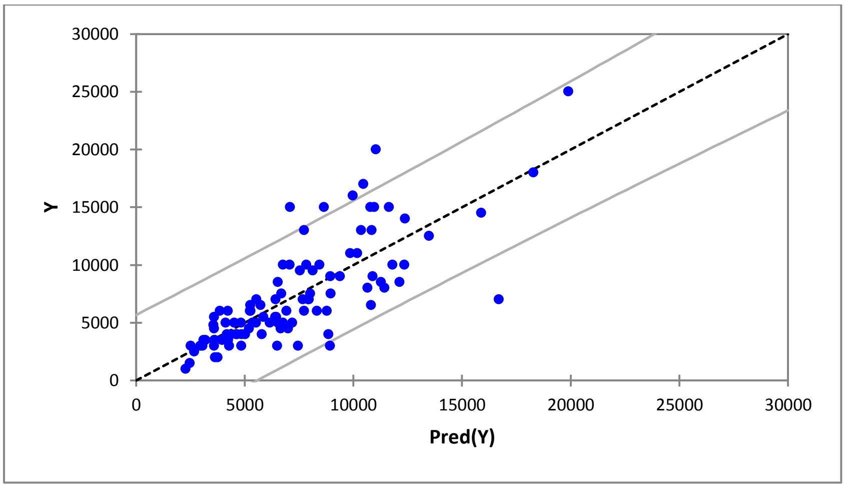

Figure 3.4: Plot of confidence intervals of problem 1

The distribution of independent variables is uniform within the range, but the errors are not homogeneous, we can say that figure 3.5 is the expected distribution of errors for linear model. The linear model describes the functional relationship between the independent and dependent variables in problem 1.

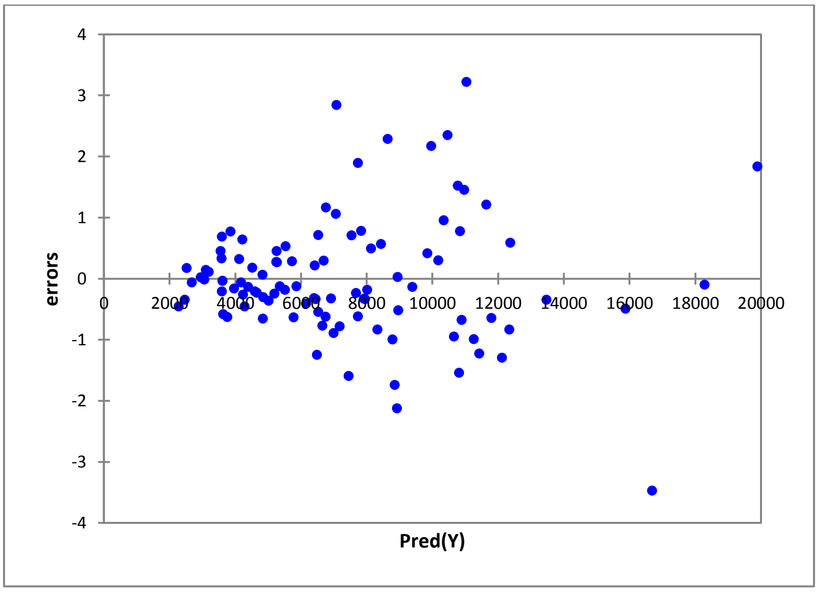

Figure 3.5: Plot of errors of problem 1

3.2 Linear Multiple Regression Model for Credit Scoring Data

Problem 2:

In this section the multiple linear regression models will be considered of the credit scoring problem with four independent and one dependent variables of the form;

$$y_i = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

where x_{i1} and x_{i2} are the independent variables $i = 1, 2, ..., n$. And, it is known that $b = (x^T x)^{-1} x^T . y$ and it can be written in the matrix form

$$b = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \left(\begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{pmatrix} \times \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{pmatrix} \right)^{-1} \times \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{pmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

In the credit scoring data set, there are n=100 observations with four independent variables and one dependent variable. The data set has been shown in table 3.9.

The matrix form was used to evaluate the coefficients of the multiple linear regression function. In data set, the independent variables x_1, x_2, x_3 and x_4 represent net income in dollars, age, last employment period and loan maturity, respectively. As well as the dependent variable is y representing the loan amount in dollars.

Table 3.9: Data set of problem 2

x_1 **:net income/$**	x_2 **: age/year**	x_3 **:last employment/years**	x_4 **:loan maturity/years**	y **:loan amount/$**
1073	29	3	36	3000
893	32	4	36	3000
664	25	2	36	6000
⋮	⋮	⋮	⋮	⋮

Table 3.9: (continued)

1987	39	6	30	10000
461	27	3	36	4000

There are $n = 100$ observations in data set. The data set in table 3.9 is converted into the matrix form as follows:

In order to calculate estimated coefficients using matrix algebra the necessary matrices are calculated such as $x^T.x$, $(\mathrm{x}^T.x)^{-1}$ and $(\mathrm{x}^T.x)^{-1}.x^T y$.

$$x = \begin{pmatrix} 1 & 1073 & 29 & 3 & 36 \\ 1 & 893 & 32 & 4 & 36 \\ 1 & 664 & 25 & 2 & 36 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 461 & 27 & 3 & 36 \end{pmatrix}_{100x5} ,$$

$$y = \begin{pmatrix} 3000 \\ 3000 \\ 6000 \\ \vdots \\ 4000 \end{pmatrix}_{100x1}$$

$$x^T.\mathrm{x} = \begin{pmatrix} 100 & 104570 & 3439 & 486 & 2814 \\ 104570 & 151965798 & 3997613 & 615934 & 2805540 \\ 3439 & 3997613 & 126725 & 18426 & 95886 \\ 486 & 615934 & 18426 & 3001.5 & 13104 \\ 2814 & 2805540 & 95886 & 13104 & 85284 \end{pmatrix}$$

$$(\mathrm{x}^T.x)^{-1} = \begin{pmatrix} 0.31802378 & 4.02024\times10^{-6} & -0.004766 & -0.000198 & -0.005236 \\ 4.0202\times10^{-6} & 4.87138\times10^{-8} & -1.51\times10^{-6} & -3.69\times10^{-6} & 5.29\times10^{-7} \\ -0.0047663 & -1.50958\times10^{-6} & 0.00031 & -0.000618 & -4.68\times10^{-5} \\ -0.0001976 & -3.69001\times10^{-6} & -0.000618 & 0.004022 & 0.000204 \\ -0.0052364 & 5.29053\times10^{-7} & -4.68\times10^{-5} & 0.000204 & 0.000188 \end{pmatrix}$$

$$\mathrm{b} = (x^T.x)^{-1}.\mathrm{x}^T\ y = \begin{pmatrix} -4160.2477 \\ 4.92080541 \\ 52.3842917 \\ 180.293656 \\ 129.880544 \end{pmatrix}$$

Then, the linear multiple regression model becomes

$\widehat{y} = -4160.2477 + 4.92080541x_1 + 52.3842917x_2 + 180.293656x_3 + 129.880544x_4$.

Now, the next step is to examine linear multiple regression model using the following calculations that is obtained by above calculations.

$$SSR = b^T x^T y - n\overline{y}^2 = 6750513591 - 5355312400 = 1395201191$$

$$SSE = y^T y - b^T x^T y = 7346540000 - 6750513591 = 596026409$$

$$SST = y^T y - n\overline{y}^2 = 7346540000 - 5355312400 = 1991227600$$

Table 3.10 is related by ANOVAs table of problem 2.

Table 3.10: ANOVA table of problem 2 with four independent variables

Source	Degree of freedom (df)	Sum of Squares (SS)	Mean of Squares (MS)	Significant F
Regression	4	1395201191	348800298	55.5949
Error	95	596026409	6273962.2	
Total	99	1991227600		

$$R^2 = \frac{SSR}{SST} = 0.70067389$$

$$r = \sqrt{0.700673898} = 0.83706266.$$

Regression Statistics:

$$S_e = \sqrt{\frac{SSE}{n-m-1}} = \sqrt{\frac{59602409}{95}} = \sqrt{6273962} = 2504.787854.$$

Adjusted R-squared: $=1-\frac{(1-R^2)(n-1)}{n-m-1}$. The adjusted R-square calculation is

$$=1-\frac{(1-0.70067389)\times 99}{95} = 0.68807069.$$

Table 3.11: Regression statistics of problem 2

Regression Statistics	
Multiple R	$r = 0.83706266$
R-Square	$R^2 = 0.70067389$
Adjusted R-squared	0.68807069
Standard error	$S_e = 2504.78785$
The number of observations	$n = 100$

Confidence Interval for b_0: $C_{00} = 0.31802378$

$$S_e(b_0) = \sqrt{MSE.c_{00}} = \sqrt{6273962.2 \times 0.31802378} = 1412.539964$$

Confidence Interval for $b_0 = b_0 \pm t.\sqrt{MSE.c_{00}}$

$= -4160.2477 \pm (1.985) \times 1412.539964$

95% confidence interval is $-6964.139504 < b_0 < -1356.355846$.

Confidence Interval for b_1: $C_{11} = 4.87138 \times 10^{-8}$

$S_e(b_1) = \sqrt{MSE.c_{11}} = \sqrt{6273962.2 \times 4.87138 \times 10^{-8}} = 0.552836976$

Confidence Interval for b$_1$: $b_1 \pm t.\sqrt{MSE.c_{00}}$

$= 4.92080541 \pm (1.985) \times 0.552836976$

For 95% confidence level, $3.823424009 < b_1 < 6.018186805$.

Confidence Interval for b_2**:** $C_{22} = 0.00031$

$S_e(b_2) = \sqrt{MSE.c_{22}} = \sqrt{6273962.2 \times 0.00031} = 44.1061201$

Confidence Interval for b$_2$ $=b_2 \pm t.\sqrt{MSE.c_{22}}$

$= 52.3842917 \pm (1.985) \times 44.1061201$

For 95% confidence level, $-35.16635671 < b_2 < 139.9349401$.

Confidence Interval for b_3**:** $C_{33} = 0.004022$

$S_e(b_3) = \sqrt{MSE.c_{33}} = \sqrt{6273962.2 \times 0.004022} = 158.8578439$

Confidence Interval for b$_3$: $b_3 \pm t.\sqrt{MSE.c_{33}}$

$= 180.293656 \pm (1.985) \times 158.8578439$

For 95% confidence level, $-135.0391643 < b_3 < 495.6264759$

Confidence Interval for b_4**:** $C_{44} = 0.000188$

$S_e(b_4) = \sqrt{MSE.c_{44}} = \sqrt{6273962.2 \times 0.000188} = 34.36716$

Confidence Interval for b$_4$: $b_4 \pm t.\sqrt{MSE.c_{33}}$

$= 129.880544 \pm (1.985) \times 34.36716$

For 95% confidence level is

$61.6617384 < b_4 < 198.0993492$.

Table 3.12: Confidence intervals of problem 2

	Coefficients	**Standard Error**	***t*-statistics**	***P*-value**	**95% Lower**	**95% Upper**
b_0	-4160.247	1412.539	-2.945	0.004	-6964.13950	-1356.35584
b_1	4.9208054	0.552836	8.901	<0.0001	3.82342400	6.018186805
b_2	52.384291	44.10612	1.188	0.238	-35.1663567	139.9349401
b_3	180.29365	158.8578	1.135	0.259	-135.03916	495.6264759
b_4	129.88005	3436716	3.779	0.000	61.6617384	198.0993492

The confidence intervals for b_2, b_3 consist of zero and their P values are rather high, we say that this multiple regression model does not fit the data that means the regression linear model is not statistically significant even though the value of r is 0.83706266.

In figure 3.6, the plot of lower and upper values of problem 2 are presented, and we used XLSTAT version 2014.3.05 Excel application to sketch figures 3.6 and 3.7.

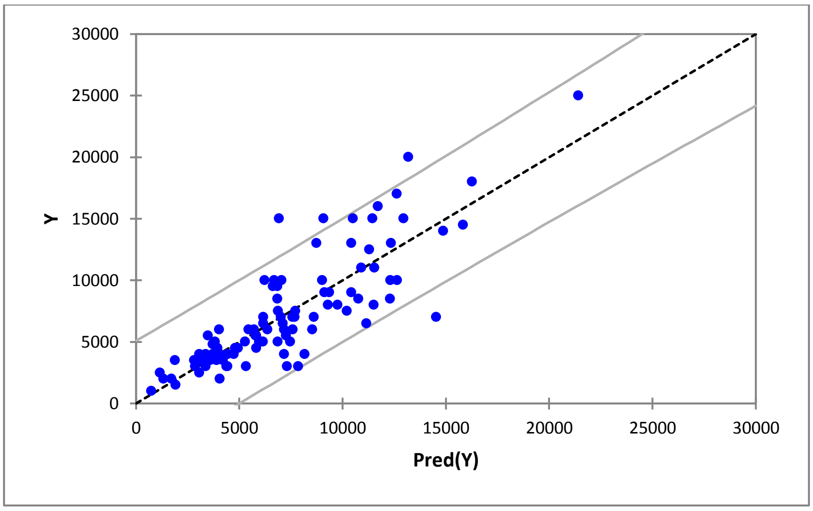

Figure 3.6: Plot of lower and upper values of problem 2

In figure 3.7, it can be seen that the errors are not homogeneous.

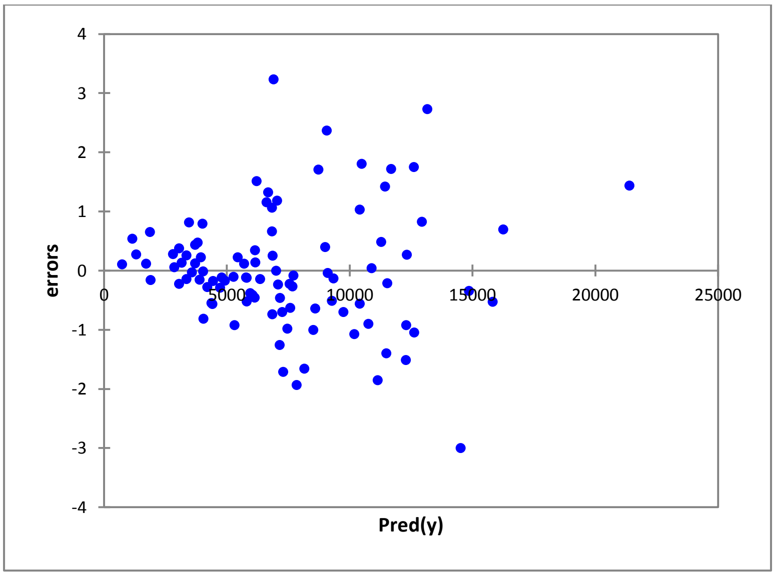

Figure 3.7: Plot of errors of problem 2

Problem 3: To obtain this new data set**,** the number of the independent variables is reduced to two in the previous data set to figure out the multiple regression models that fits the data. x_1 and x_4 represent net income in dollars and loan maturity in years, respectively.

Table 3.13: Data set of problem 3

x_1 : **net income/$**	x_4 : **loan maturity/year**	y : **loan amount/$**
1073	36	3000
893	36	3000
664	36	6000
⋮	⋮	⋮
1987	30	10000
461	36	4000

$$x = \begin{pmatrix} 1 & 1073 & 36 \\ 1 & 893 & 36 \\ 1 & 664 & 36 \\ \vdots & \vdots & \vdots \\ 1 & 461 & 36 \end{pmatrix}_{100x3} \qquad y = \begin{pmatrix} 3000 \\ 3000 \\ 6000 \\ \vdots \\ 4000 \end{pmatrix}_{100x1}$$

$$x^T . \mathrm{y} = \begin{pmatrix} 100 & 104570 & 2814 \\ 104570 & 151965798 & 2805540 \\ 2814 & 2805540 & 85284 \end{pmatrix}$$

$$(\mathrm{x}^T . \mathrm{x})^{-1} = \begin{pmatrix} 0.211087231 & -4.24464\times10^{-5} & -0.005568624 \\ -4.24464\times10^{-5} & 2.52932\times10^{-8} & 5.6849\times10^{-7} \\ -0.005568624 & 5.6849\times10^{-7} & 0.000176765 \end{pmatrix}.$$

Using above matrices the values of the estimated coefficients are evaluated that can be seen in the following matrix.

$$b = (\mathrm{x}^T \mathrm{x})^{-1}.\mathrm{x}^T \mathrm{y} = \begin{pmatrix} -2366.348613 \\ 5.858894908 \\ 126.4286499 \end{pmatrix}$$

$$\hat{y} = -2366.348613 + 5.858894908x_1 + 126.4286499x_4 .$$

The following table is the ANOVA table for problem 1, table 3.14 is created by the same formulas that are presented in table 2.3

Table 3.14: ANOVA table of problem 3 with two independent variables

Source	Degree of freedom	Sum of Squares (SS)	Mean of Squares (MS)	Significant F
Regression	2	1357320178	678660089	103.848
Error	97	633907422	6535128.061	
Total	99	1991227600		

To form table 3.15 the values of R^2, r, standard error, and the adjusted R-squared all are evaluated to analyze the significance of the model.

Table 3.15: Regression statistics of problem 3

Regression Statistics	
Multiple R	$r = 0.825620943$
R-Square	$R^2 = 0.681649942$
Adjusted R-squared	0.675086023
Standard error	$S_e = 2556.389654$
The number of observations	$n = 100$

Standard error for b_0 :

$$c_{00} = \sqrt{MSE \times \mathrm{c}_{00}} = \sqrt{6535128.061 \times 0.2110872} = 1174.513$$

Confidence intervals for b_0: $b_0 \mp t.\sqrt{MSE \times c_{00}}$

$= -2366.348613 \pm (1.985) \times 1174.513$

95% confidence level is $-4697.76 < b_0 < -34.9403$.

Similarly, the confidence intervals for other coefficients have been calculated, all are in table 3.15.

Table 3.16: Confidence intervals of problem 3

	Coefficients	**Standard error**	**t-statistics**	**P-value**	**95% Lower bound**	**95% Upper bound**
b_0	–2366.348613	1174.513	–2.015	0.047	–4697.76	–34.9403
b_1	5.858894908	0.407	14.411	< 0.0001	5.0551979	6.665811
b_4	126.4286499	33.998	3.720	0.000	58.97203	193.8853

Comparing problem 2 and problem 3 we say that the standard errors of the coefficients of problem 3 are less than the standard errors of the corresponding coefficients of problem 2. And, also the confidence intervals in this problem do not consists of zero. The linear regression model that is obtained in problem 3 fits the data better than problem 2. Also, we used XLSTAT version 2014.3.05 Excel application to sketch figures 3.8 and 3.9.

In the following figures 3.8 and 3.9 the relationship between the confidence intervals and errors can be seen.

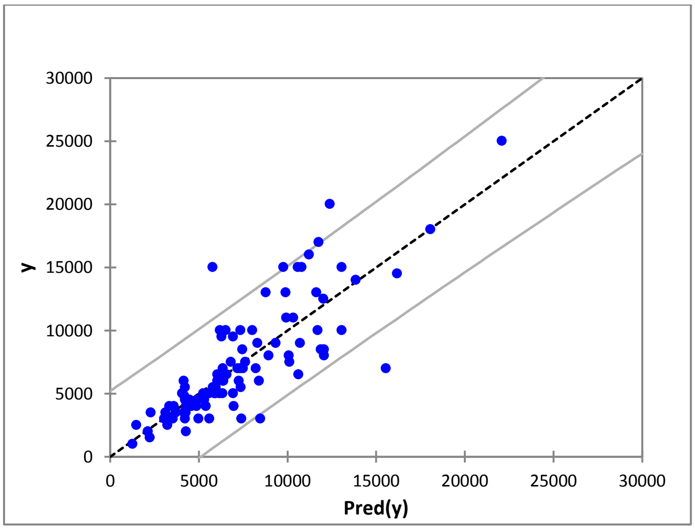

Figure 3.8: Plot of confidence intervals

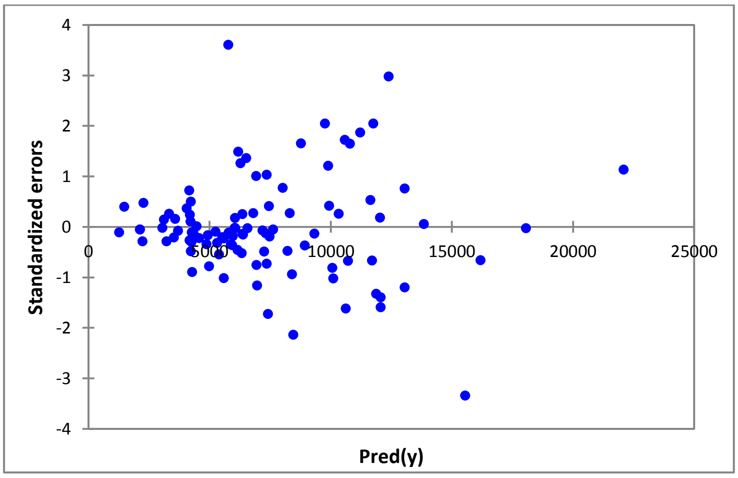

Figure 3.9: Plots of errors

Chapter 4

LOGISTIC REGRESSION PROBLEM

4.1 Credit Scoring Data

Problem 4: Given the data set which has two variables; the independent variable, x, represents the age of customers and the dependent variable, y, represents accepted or denied cases of applications of customers. Table 4.1 is the whole data set for problem 4. The scatter graph of data is shown in Figure 4.1, and then the data set is transformed into the new form that is given in table 4.2

Table 4.1: Data set of problem 4

Number of observations	Age	Accepted/Denied
1	20	Denied
2	21	Denied
3	22	Accepted
⋮	⋮	⋮
98	56	Accepted
99	56	Accepted

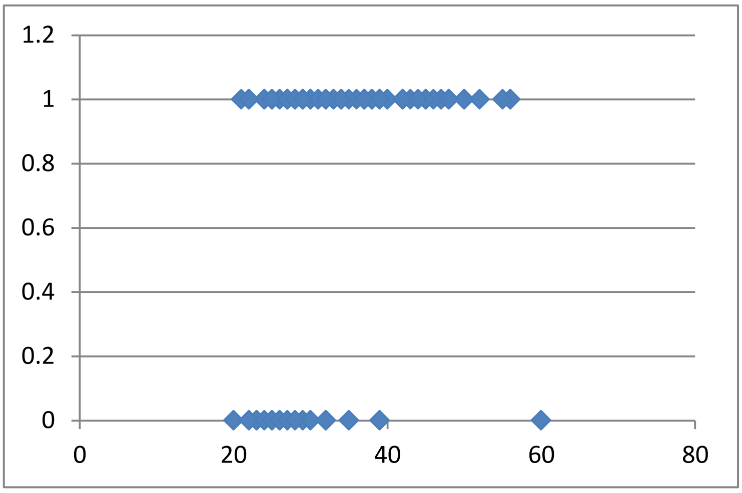

Figure 4.1: Scatter graph of data

Table 4.2: Partitions of data

		Total	*Accept*
1	20-24	12	5
2	25-29	26	18
3	30-34	20	17
4	35-39	16	14
5	40-44	8	7
6	45-60	18	17

After that the midpoint of each interval is selected to define the independent variable x_{age} and the probabilities, odds and ln(Odds) are calculated using table 4.2, the corresponding results are presented in table 4.3.

In table 4.3, the Odds of each interval and logarithmic values are evaluated for every partition. Now let us show the calculation of the Odds for the first partition.

$$Odds = \frac{Pr(accepted)}{1 - Pr(accepted)} = \frac{0.416667}{1 - 0.416667} = 0.714287\,.$$

Table 4.3: Some results of problem 4

x_{age} **(mid-points)**	**Total**	**Accept**	**Probability**	**Odds**	**Ln(Odds)**
22	12	5	0.416667	0.714286	-0.33647
27	26	18	0.692308	2.25	0.81093
32	20	17	0.85	5.666667	1.734601
37	16	14	0.875	7	1.94591
42	8	7	0.875	7	1.94591
52.5	18	17	0.944444	17	2.833213

The graph of the dependent variable y with respect to the probability is S - shape curve, this relationship is presented in figure 4.2.

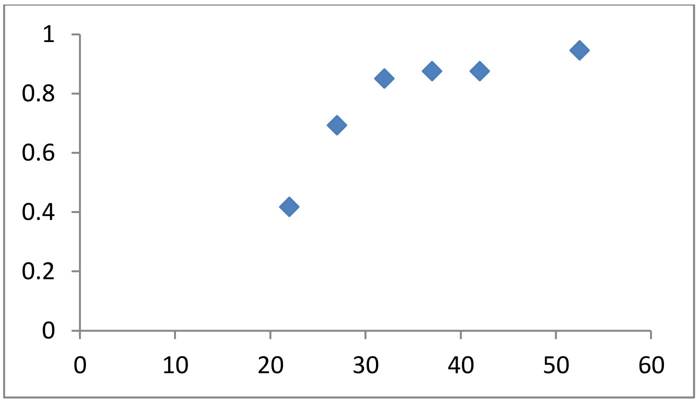

Figure 4.2: Plot of the proportions in each partition

And in order to define the logistic model we used the logit function which is $\ln(Odds) = b_0 + b_1 x$.

The least square estimation is applied to find the coefficients, then the logit function is obtained as $\hat{g}(x)$=logit$(\pi(x)) = 0.093 - 1.819 x_{age}$

Table 4.4: Predicted probability

Age	$\text{logit}(y) = 0.093 - 1.819 x_{age}$	**Predicted probability**
0-24	0.228	0.556754
25-29	0.693	0.666634
30-34	1.158	0.760969
35-39	1.623	0.835208
40-44	2.088	0.889731
45-60	3.0645	0.955404

The scatter graph of x_{age} and the predicted probability is shown in figure 4.3

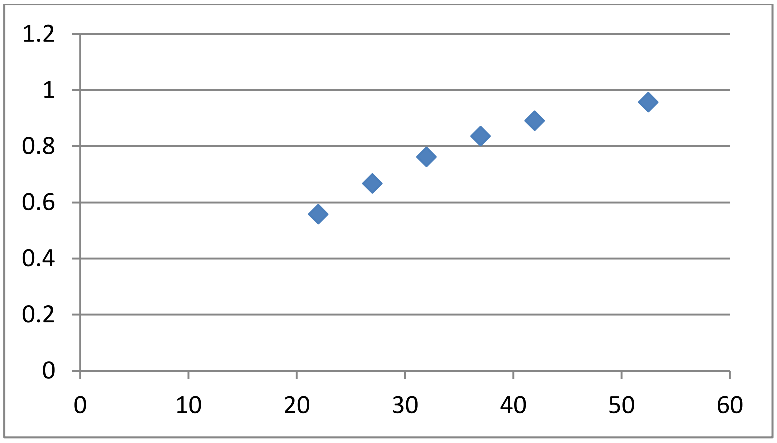

Figure 4.3: Scatter graph of predicted probability

The graph of the logit function is linear; it can be seen in figure 4.4

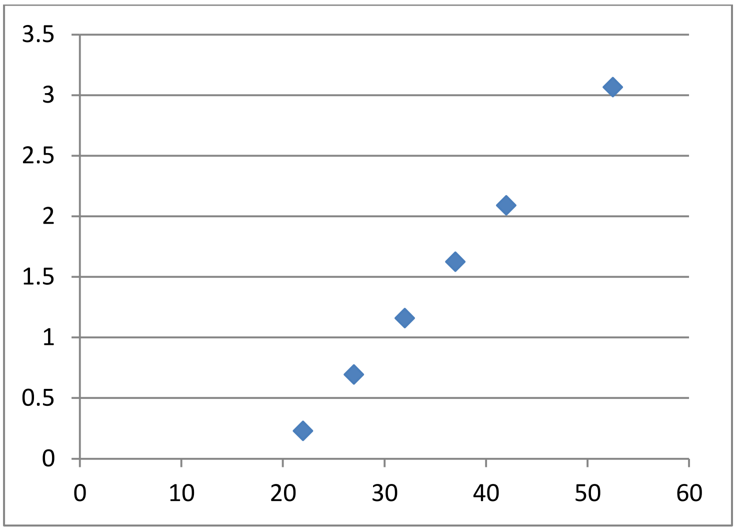

Figure 4.4: Plot of logit function

Usually the maximum likelihood estimation is used to estimate the coefficients of logistic regression model. The results in the following table 4.5 are created using SPSS. In this part we would like to present the model using the maximum likelihood estimation (using SPSS).

Significance of Logistic regression Model:

Table 4.5: Statistical results of logistic regression

	Coefficients	***Standard error***	Z_{Wald}	***df***	***Significant***	$\exp^{ceoff.}$
Intercept	-1.721	1.228	1.964	1	0.161	0.179
Slope	0.101	0.040	6.371	1	0.012	1.106

Then the logit function becomes $\widehat{g}(x) = -1.721 + 0.101x$ where the slope is 0.101 and the intercept is -1.721.

$$\widehat{\pi}(x_{age(i)}) = \frac{e^{-1.721+0.101x_{age(i)}}}{1+e^{-1.721+0.101x_{age(i)}}}$$

For example the probability of age 40 is $\widehat{\pi}_i(40) = \frac{e^{-1.721+0.101(40)}}{1+e^{-1.721+0.101(40)}} = 0.910438$

since $\widehat{g}(40) = -1.721 + 0.101(40)\text{=}10.1655$. The probability of accepting the credit application at age 40 is %91 and the probability of rejecting the credit application at age 40 is %9. The occurrence of acceptance of credit application at age 40, with %91, is positive.

Table 4.6: $\widehat{\pi}(x_i)$ function

Age	$\text{logit}(y) = 0.101 - 1.721x_{age}$	**Predicted probability**
0-24	0.228	0.622694
25-29	0.693	0.732237
30-34	1.158	0.819209
35-39	1.623	0.882467
40-44	2.088	0.925601
45-60	3.0645	0.97292

Significance of b_1**:** the statistic Z is $Z = \frac{0.101}{0.04} = 2.525$, this value is greater than 2, or the Wald statistic is $Z_{Wald} = (\frac{0.101}{0.04})^2 = 6.371$, as a result we may say that b_1 is not zero by hypothesis.

We define 95% confidence interval for the coefficient b_1 by using $b_1 \mp t.S_e(b_1)$, $0.101 \mp 1.96 \times 0.040$, then the upper limit is 0.1794 and the lower limit is 0.0242, 95% confidence interval does not consists of one, therefore b_1 is not zero, [5]. The logistic model that it is discussed in this section fits the data set.

Chapter 5

CONCLUSION

In this study simple linear, multiple linear and logistic regression models are studied in discussing the credit scoring data set by assigning different variables. We see that the simple linear regression model fits the data set very well in problem 1 because the value of r is almost one, and also confidence intervals in problem 1 do not consists of zero. Comparing problem 2 and problem 3, we say that the value of r in problem 2 is greater than the value of r in problem 3 but the confidence intervals contain zero in problem 2, consequently the multiple linear model in problem 3 describes better functional relationship between the dependent and independent variables than problem 2. The logistic regression is applied to fit the data with binary output variable, according to the logistic regression model the probability is poor to get credit from the bank when the age of customer is less than 25, and the probability is high to get credit when the age of applicant is greater than 35.

REFERENCES

[1] Michael J.A. Berry & Gordon J. Linoff. (2004). Data Mining Techniques. *Wiley Publishing, Inc*., 2nd Ed.

[2] Jiawei Hand & Micheline Kamber. (2006). Data Mining Concepts and Techniques. *Morgan Kaufman Publishers, Elsevier*, 2nd Ed.

[3] Agresti, Alan. (2002). Categorical Data Analysis. *John Wiley & Sons, Inc*. 2nd ed. ISBN: 0-471-36093-7, 165-356.

[4] Deborah Burr. (1998). On Errors-in-variables in Binary regression-Berkson Case. *Journal of the Amrican Statistical Association*. Vol. 83, 739-753.

[5] David W. Hosmer & Stanley Lemeshow. (1989). Applied Logistic Regression. *John Wiley and sons, Inc*. ISBN 0-471-61553-6.

[6] Kocenda, Evezen & Vojtek, Martin. (2009). Default predictors and credit scoring models for retail banking. Cesifo Working paper no.2862. Leibniz Institute for Economic Research at the University of Munih. 1-54.

[7] Maria Aparecida Gouvea & Eric Bacconi Goncalves. (2007). Credit Risk Analysis Applying Logistic Regression, Neural Networks and Genetic Algorithms Models. *POMS 18th Annual Conference.*Texas, USA, 2-49.

[8] Yanwen Dong (2007). A case based reasoning system for customer credit scoring: comparative study of similar measures. In proceedings of 51st Annual Meeting of the International Society for the Systems. Tokyo, Japan, 5-10.

[9] Natasa Sarlija, Kristina Soric, Silvija Vlah & Visnja Vojvodic Resenzweig. (2009). Logistic Regression and Multicriteria Decision making in Credit Scoring. *On Operational Research. 10th International Symposium on Operational Research SOR'09*, 174-185.

[10] Baesens, B., Gestel, T. Van, Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society,* 54, 627-635.

[11] Draper, Norman R. & Smith, Harry. (1998). Applied Regression Analysis. *John Wiley & Sons, Inc*. 3rd Ed., 15-46.

[12] Chatterjee, Samprit & Hadi, Ali S. (2006) Regression Analysis by Example. *A John Wiley & Sons, Inc., Publication*, QA278.2.C5, 4th Ed., 21-84.

[13] Tuffery, Stephane. (2011). Data Mining and Statistics for Decision Making. John Wiley & Sons, ltd. Publication, ISBN: 978-0-470-68829-9, 437-491.

[14] Kantardzic, Mehmet. (2011). Data mining Concepts, Models and Algorithms. *John Wiley & Sons, Inc*. 2nd Ed. ISBN: 978-0-470-89045-5, 140-167.

[15] Michael J.A. Berry, Gordon S. Linoff. (2004). Data Mining Techniques. *Wiley Publishing, Inc*. Chapter 9, 287-319.

[16] Abdou, H & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intelligent Systems in Accounting, Finance & Management, 18 (2-3), 59-88.

[17] Leon, Steven J. (2006). Linear Algebra with Applications. *Pearson International Edition*, 7th Ed., 234-475.

APPENDIX

Appendix: Least Square Estimate

Consider the following data with n observations in table A.1.

Table A.1: Data set

x_i	x_1	x_2	$\cdots$	x_n
y_i	y_1	y_2	$\cdots$	y_n

And also we suppose that the scatter plot of the dataset is linear. We write a linear model for the independent variable x and dependent variable y as

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

where β_0 and β_1 are the y-intercept and the slope, respectively and ε is an error. According to table A.1, equation (1) can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{2}$$

where $i = 1, 2, ..., n$. We need to estimate parameters β_0 and β_1 using the least squares method. By finding the line our aim is to minimize the sum of the squares of the vertical distances from each of the point to the regression line.

$$S = \varepsilon_1^2 + \varepsilon_2^2 + \; .... + \varepsilon_n^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

If S = 0, the sum of the squares is minimized, that means the regression line is perfect. From equation (2), we write

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x \, , \; i = 1, 2, ..., n$$

$$S = \varepsilon_1^2 + \varepsilon_2^2 + \; .... + \varepsilon_n^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The first partial derivatives of S are calculated with respect to β_0 and β_1, then we have

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \, . \, x_i.$$

Now, assume that $\frac{\partial S}{\partial \beta_0} = 0$ and $\frac{\partial S}{\partial \beta_1} = 0$,

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{3}$$

(4)

$$\sum_{i=1}^{n} (x_i y_i - \beta_0 x_i - \beta_1 \, x_i - \beta_1 x_i^2) = 0.$$

From equation (3) we get

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} \beta_1 x_i = 0$$

$$\sum_{i=1}^{n} y_i - n\beta_0 - \sum_{i=1}^{n} \beta_1 x_i = 0$$

this implies that

$$\beta_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \beta_1 . \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\beta_0 = \overline{y_\iota} - \beta_1 . \overline{x_\iota}. \tag{5}$$

From equation (4), we write

$$\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0. \tag{6}$$

Substitute equation (5) into equation (6) to find the following equation

$$\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n} \sum_{i=1}^{n} x_i\right).\sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0.$$

Simplifying the last equation, the value of the slope is obtained doing the following steps:

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i.\sum_{i=1}^{n} y_i}{-\frac{1}{n}\sum_{i=1}^{n} x_i.\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\,(\frac{1}{n}\sum_{i=1}^{n} x_i).(\frac{1}{n}\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i).(\frac{1}{n}\sum_{i=1}^{n} x_i)}$$

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

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\,\bar{y} - \bar{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} x_i + \bar{x}\sum_{i=1}^{n} x_i}$$

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

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

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

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

By putting the value of the slope into the equation (5), the value of y-intercept is defined. The values of the estimated parameters β_0 and β_1 minimize the sum of the square, in other words S approaches 0.

$\hat{y} = \beta_0 + \beta_1 x_i$ is the linear function defined by least squares where $\widehat{y_i}$ represents the predicted value of y_i for a given x_i .

If we substitute the value of β_0 into the predicted line then we obtain

$$\widehat{y_i} = \beta_0 + \beta_1 x_i$$

$$\widehat{y_i} = \overline{y_i} - \beta_1 \overline{x_i} + \beta_1 x_i$$

$$\widehat{y_i} = \overline{y_i} + \beta_1 (x_i - \overline{x_i}).$$

Of course the point $(\bar{x} , \bar{y})$ lies on regression line that means the regression line $\widehat{y_i} = \beta_0 + \beta_1 x_i$ contains the center of gravity of the dataset, [11].