Regression Modeling for Incidence of Diabetics

Amar Yahya Zebari

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
February 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 Applied Mathematics and Computer Sciences.

Prof. Dr. Nazim Mahmudov
Chair, Department of Applied Mathematics
and Computer Sciences

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 Sciences.

Asst. Prof. Dr. Mehmet Ali Tut
Supervisor

Examining Committee

1. Assoc. Prof. Dr. Rashad Aliyev
2. Asst. Prof. Dr. Huseyin Etikan
3. Asst. Prof. Dr. Mehmet Ali Tut

ABSTRACT

Biostatistics is one of the important approaches for decision makers in the health sciences for mathematical modeling and predictions. The choosing of a topic of diabetes to be applied in this study due to the importance of finding a cure of this disease, which is the incidence rates increased in the last years. The reason of this increased and the types are investigated by the researchers, to illustrate how some variables as weight and age its effects on diabetes.

The study was conducted on a sample of 1385 patients with diabetes, randomly selected from the community data diabetics in the Diabetics Center province of Duhok/ Kurdistan Region of Iraq, of 10,083 patients with diabetes, and applies the theories of linear regression on this data to create a mathematical equation helps us to anticipate future injury rates. The results are then compared with the results of statistical study on the Greek Cypriot patients less than 15 years of age, to clarify the differences and to clarify the effects.

The use of the Statistical Software Packages for Social Sciences (SPSS) in this study is to obtain more accurate results and reduce the time and voltage. This study basically is the application of linear regression modeling to cases of diabetic patients. Chapter one included a brief about diabetes and its type and reference to some other statistical researches conducted on diabetic's data. Chapter two is about the theories and concepts that can use it in application this study and obtained requires results. The third chapter is the application of these statistical theories on diabetic's data that there is a belief that have an effect on incidence this disease like weight and family

genetic history, and analyze the results graphically and illustrations using the Statistical Package for Social Sciences which is referred as SPSS, then modeling a mathematical regression equation for these data. The results showed several statistics about the Duhok data. Several differences in terms of means between males and females were listed. Duhok data and its statistics were compared with a data related with Cyprus region.

A regression function was also constructed for predicting diabetes for some next time periods. An exponential model fitted the current Duhok data.

Keywords: Biostatistics, Statistical Analysis, Diabetes, Regression Analysis, Mathematical Modeling, Statistical Software Packages for Social Sciences (SPSS).

ÖZ

Biyoistatistik, sağlık bilimleriyle ilgili biyolojik veri analizi ve modellemesinde kullanılabilen önemli bir daldır. Bu çalışmada Irak'ın Duhok bölgesi için diyabetik hastalarla ilgili bilgilerin analizi yapılarak özellikle erkek ve kadın hastalar arasındaki değişik statistiki ilişkilerin tesbit edilmesine çalışılmıştır. Ayrıca Kıbrıs'daki bir durum analizindeki verilerle de Dukok verileri arasında bir karşılaştırma yapılmıştır. Ayrıca Duhok bölgesindeki diyabetli hasta sayısının ilerleyen zaman dilimlerindeki değişimin kestirilebilmesi için regresyon analizi de yapılarak sözkonusu verilerin en iyi exponansiyel modelle modellenebildiği ortaya konmuştur.

Anahtar Kelimeler: Biyoistatistik, İstatistiksel Veri Analizi, Diyabet, Matematiksel Modelleme, Regresyon Analizi.

DEDICATED

To my Lovely Family

To my Brothers and Sisters

ACKNOWLEDGMENT

Above all, I thank God who helped me and allowed me to complete my graduate studies in this prestigious university.

In particular, I would like to thank also my supervisor Prof. Dr. Mehmet Ali Tut, for the time and effort that he gave me, to help me accomplish this thesis.

Also, thanks to assistant Mr. Mani Mehraei, who gave me a lot of advice, assistance and encouragement to complete this thesis.

Thanks also for all distinguished professors in Applied Mathematics and Computer Sciences faculty in Eastern Mediterranean University (EMU).

Last but not least, I would like to thank my mother and my big brother who always supported me and encouraged me by all means to complete my graduate studies, thank you for your prayers, thanks to my brothers and sisters, and all my friends.

Finally, I would like to thank all people in Cyprus, Thank you for the wonderful times we spent with you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Genaral

The modeling is a general and precise technique used in multivariate analysis for methods as a special cease and simplifying the relationship between variables. One of the main modeling applications is a regression models that is an extension for a simple linear regression analysis which may be bounded by regression weights to be equal for each others, or to determine numeric values. The linear regression is a statistical measure, attempts to determine the relationship between the dependent variable that is almost referred by Y, and a number of other variables called independent variables and often denoted by X [1].

The linear regression model will help to predict future patient's data, and that can be benefit, for example, but not limited to, in medical studies. A linear regression modeling for diabetics data represent one of many studies that researchers can do it to help diabetics and specialists about diabetes to find out the real reason and effects that causes this disease and the extent of the disparity in the impact of those causes, thus finding a treatment and appropriate health awareness. The application of some statistical theories and modeling through the use of computer programs contribute effectively in reducing the time and great effort for such studies, which are usually with great data, and thus obtain more accurate results and ease of understanding of outcomes for researchers from non-specialists using shapes and graphs and curves in this software.

The large increase in the numbers of people with diabetes around the world, made us choose this topic for the research under study. Where that according to the regression theories, the linear regression equation modeling data for diabetics will help in an approximation determine the number of people with diabetes in the next years in the region under study, and study the effects of which are believed to be linked to incidence of diabetes, according to the available data.

This study is purely a statistical study, and will help medical researchers in their study about diabetes, its causes, affects.

It is important to note that there are two main types of diabetes:

- Type 1 diabetes, symbolized by T1D: " Also called Juvenile – Onset, usually caused by an autoimmune reaction, this type affect any age, but usually develops in children and young people, if people with Type 1 diabetes do not have access to insulin, they will die"[2].

- Type 2 diabetes, symbolized by T2D:" Also called "Non – Insulin Dependent Diabetes" or" Adult – Onset diabetes ". The diagnosis of Type 2 diabetes can occur at any age, and account for at least 90% of all cases of diabetes"[3]. For example; in the united states "Among the seventeen million diabetic, there is a rate of 90% to 95% of them, infected with Type 2 Diabetes"[4]. There are also several reasons for the acquisition of diabetes, have been classified in the diabetics data that we have into two types:

- The first is the acquisition of genetic diabetes, symbolized by (P) which means a positive diagnosis.

- The second type is the acquisition of the diabetes for other reasons, such as obesity.

In a brief definition of diabetes, Dr. Ranger Hanas says:

"It so is important to clarify that diabetes was not caused by anything that you or our family have done or failed to do."[5]

The sample data under study include this information about every patient:

- Patient's age.
- Gender
- Diabetes Type
- Weight.
- Length.
- Date of Diabetes.
- Acquisition.

Thus, we have seven variables under study, with the loss of few number of data classified within the registry errors, where the statistical theories deal with these missing values by one of four methods that is: [6]

1. Mean imputation.
2. Hot deck imputation.
3. Regression imputation.
4. Multiple imputations.

The number of people with diabetes may differ from one country to another, so, it is more accurate to analyze the diabetic data in two different countries, weather sampling or study the entire community if possible.

Due to the urgent need for such a study, many of which were conducted by health organizations or by personal research, some of which is still conducted periodically by the health authorities to determine the changes that occur to prepare people with diabetes.

In the research, which was conducted by Picker Institute Europe[7], Patients were divided according to the reliability of primary care trusts to three factors: Ethnic diversity, age distribution and deprivation, the study was conducted on the two cities Sheffield and Devon, respectively, with a sample size n=900, the data was collected at first by telephone and then by Emile, each patient were asked several questions regarding the extent of confidence in treatment and types of treatment used by categories of sex and age groups and ethnicity, And any treatments patients believe they respond better and the relationship among all of these with the price of treatment.

In other research that due by The College of Health Care Sciences at Nova Southeastern University[8], that is about assess whether the characteristics of some hospitals have significant impacts on the length of survival of patients admitted to non-federal hospitals with Type 2 diabetes and angina heart muscle using the comparison and analysis of data. Theories that are used linear regression analysis, the encrypted form, descriptive and binary variable. Sample was collected according to the international classification of disease, and the sample for this study was for inpatients at 2006. The sample of 2774 participants were selected from a database of inpatient at the national level and the cost of health care and benefit to the project in 2006, The criteria used in this study are:

- Admitted to non-federal hospitals.

- Diagnose patients inside the hospital to both diabetes and angina.
- The age of 20 - 84 years old.

1.2 Design of The Sudy

1.2.1 Methods

The design of this study includes the analysis of binary data for patients with diabetes in the governorate of Duhok/ Kurdistan region of Iraq, and modeling of the linear regression equation for this data, helps us to predict the number of diabetics in the future, in the region under study. Then, it is compared with the statistical analyzes of data analyzes of diabetics in Northern Cyprus.

Generally, we will use statistical methods and in particular we will use linear regression theory, that is: "answer questions about the dependence of a response variable on one or more predictors, including prediction of future values of a response, discovering which predictors are important in estimating the impact of changing a predictor or a treatment on the value of the response"[9].

When we have a large sample, it would be difficult to find the necessary analyzes and avoid errors resulting from manual solution. Therefore, it is better to use of Statistical Package for the Social Sciences, symbolized as SPSS " was released in it is first version in 1968 after being developed by Norman H. Nie, Dale H. Bent and C. Hadlai Hull"[10].

This program can be used in all statistical analyzes, and creating illustrative graphs, such as Histogram, Box Plot, Pie, Line, Error Bar, Area, Scatter plot, Drop line, with save time and effort and provide an accurate results.

1.2.2 Population in The Study

We have a total sample of (1,385) patients of diabetes, were randomly selected from the diabetes community roughly 10,083 diabetic in the Diabetes Center/ province of Duhok/ Kurdistan region of Iraq.

There are some missing values for some patients with diabetes data fall within registry errors, and statistical analyses of the data are taken errors occurring as a result of the data collection or sampling bias into consideration.

Observed in a community that has been selected sample of it, that the number of diabetic patients in a large increase, year after year, which calls for the case study.
On the other hand, the choice of another sample of diabetic patients from another community, and conducting statistical analyzes it. Then, a statistical comparison between samples, will lead to a great understanding of relationship between the disease and the exiting variables within patient data.

Chapter 2

Mathematical Background

2.1 Background

Regression modeling refers to a mathematical explanation of a process in terms of a set of associated variables. The value of a dependent variable based on the level of many independent variables. For Example; the yield of a certain production process may be depends on the pressure, throughput, temperature. The car's fuel efficiency may be depends on the weight of the car, specifications engine and body. Show product on the market depends on the price that the customer intends to pay. In all of these situations we are interested to get a "model" or a "law" for the relationship between the dependent variable (often referred as y), and the independent variables (referred as x).

Regression analysis deals with the modeling functional relationship between the response variable (dependent variable y) with one or more of the explanatory variables (independent variable x). [11].

2.2 General Model

In all example mentioned above, cases which have only one response variable is modeled as:

$$Y = \mu + \varepsilon \tag{2.1}$$

When Y is a response variable, ε is a random error and μ is the peremptory component and written as: [11]

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{2.2}$$

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where $x_1, x_2, \ldots, x_k$ are p explanatory variables, β_0 is a regression slope intercept, and $\beta_1, \beta_2, \ldots, \beta_p$ are p coefficient regression, assuming that the explanatory variables are measured without errors. In addition, the errors for all cases ε_i and ε_j are assumed independent.

The model in (2.1) referred as linear in the parameters. To illustrate the idea of linearity more, the following four models shall be observed: [11]

1. $\mu = \beta_0 + \beta_1 x$
2. $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ (2.3)
3. $\mu = \beta_0 + \beta_1 x_1 \beta_2 x_2^2$
4. $\mu = \beta_0 + \beta_1 x_1 \exp(\beta_2 x)$

Models 1 and 3 are linear in the parameters β_i because the derivation of linear regression equation $(\frac{\partial \mu}{\partial \beta i})$ that used to minimized error and to find the best fit line, do not depend on the parameters [12].

Model 4 is non-linear in the parameters because the derivation of the equation with respect to β_1 is:

$$\frac{\partial \mu}{\partial \beta_1} = \exp(\beta_2 x) \; and \; \frac{\partial \mu}{\partial \beta_2} = \beta_1 x \exp(\beta_2 x),$$

Depend on the parameters. In equation 1 and 2, the model can be extended in many ways.

1- Functional relationship perhaps is non-linear, and we consider a model as that in (2.3, 4) to clarify the non-linear pattern.

2- May be we suppose that $V(y) = \sigma^2(x)$ is a function of explanatory variables. Where $\sigma^2(x)$ is the population variance and we can calculate it by the following formula: $\sigma^2 = \frac{\Sigma(x_i - \bar{x})^2}{N}$

3- For different cases, response may not be independent [11].

We can understand the relationships by many different ways. Researcher can be start from a developed theory and use the data essentially for the estimation of unknown parameters and for testing if the theory is consistent with experiential information. Another way that is usually used in the social sciences, that is begin from the data and use and experiential modeling approach to derive a model that provides

reasonable description of the relationship. In some situation theory will suggest certain models. In other cases, theory may be incomplete or may not exist [11].

2.3 Simple Linear Regression Model

The model is referred as:

$Y = \mu + \varepsilon \qquad\qquad where\ \mu = \beta_{0+}\ \beta x$

It is usually referred to as simple linear regression model because there are only one predictor variable is involved. If we have n pairs of observation

$(x_i, y_i);\ i = 0, 1, 2, \dots, n.$

Then, these observations can characterize as:

$y_i = \beta_0 + \beta_1 x_i + \ \varepsilon_i\ ;\ i = 0, 1, 2, \dots, n.$

2.4 Important Assumptions

Standard analysis depends on the following assumption around regression variable x and random error $\varepsilon_i\ where\ i = 0, 1, 2, \dots, n.$

1- The regression variable be under the experimenter control, which can determine the values $x_1, x_2, \dots, x_n$. This means that $x_1, x_2, \dots, x_n$ can be taken as constants and they are not random variables.

2- $E(\varepsilon_i) = 0\ ; i = 1, 2, \dots, n. where\ E(\varepsilon_i)$ is the expected value of random error.

This implies to:

$\mu_{i=} E(y_i) = y_i = \beta_0 + \beta_1 x_i\ ;\ i = 0, 1, 2, \dots, n$ [13].

3- $V(\varepsilon_i) = \sigma^2$ is constant for all $i = 0, 1, 2, \dots, n.$

This implies to the variations $V(y_i) = \sigma^2$ are all the same and all observations have the same accuracy.

4- The differences between errors $\varepsilon_i \; and \; \varepsilon_j$ subsequently, responses differences $y_i \; and \; y_j$ are independent. This implies to:

$$Cov\left(\varepsilon_i, \varepsilon_j\right) = 0 \; for \; i \neq j.$$

Where $Cov\left(\varepsilon_i, \varepsilon_j\right)$ is an abbreviation for the covariance that characterizes the degree to which two different variables are linked in a linear way [14].

The model implies to the response variable observations y_i derived from probability distributions with

$$\mu_i = E(y_i) = y_i = \beta_0 + \beta_1 x_i$$

And fixed variance σ^2. In addition, any two observations y_i, y_j are independent for all $i \neq j$.

2.5 Estimation of Parameters

2.5.1 Maximum Likelihood Estimation

Maximum likelihood estimation chooses the estimates of the parameters so the likelihood estimation is maximum. Likelihood for the parameters $\beta_0, \beta_1, \sigma^2$, is the Joint Probability Density Function for $y_1, y_2, \dots, y_n$ seen as a parameters function.

Probability distribution for y must be determined if we want to use this approach. In addition to the assumptions that formed before, we will assume that ε_i has a normal distribution with mean=0 and variance $= \sigma^2$ and the dependent variable distributed with mean equal to $\beta_0 + \beta_1 x_i$ and variance equal to σ^2[11].

$$\varepsilon_i \sim N(0, \sigma^2) \; and \; y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

The probability density function for the $i'th$ response yi is:

$$P(y_i|\beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma 2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right] \quad (2.4)$$

The joint probability density function of $y_1 y_2, \ldots, y_n$ is $p(y_1 y_2, \ldots, y_n)$ is

$$p(y_1, y_2, \ldots, y_n|\beta_0 + \beta_1 x_i, \sigma^2) = (\frac{1}{\sqrt{2\pi}})^n \sigma^{-n} \exp[-\frac{1}{\sqrt{2\sigma 2}} \sum (y_i - \beta_0 - \beta_1 x_i)^2 \quad (2.5)$$

Treated these as a function of the parameters implies to the likelihood function $L(\beta_0, \beta_1, \sigma^2|y_1, y_2, \ldots \ldots, y_n)$ and its logarithm, that it is:

$$L(\beta_0, \beta_1, \sigma^2) = Ln\, L(\beta_0, \beta_1, \sigma^2) = k - n\, Ln\, \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \quad (2.6)$$

We have, $k = \left(-\frac{n}{2}\right) Ln\, 2\pi$ is a constant that does not depend on the parameters [11].

The Maximum Likelihood Estimator (MLE's) of $\beta_0, \beta_1, \sigma^2$ to maximize $L(\beta_0, \beta_1, \sigma^2)$.

Maximizing the log-likelihood $L(\beta_0,\beta_1,\sigma^2)$ with respect to $\beta_0\ and\ \beta_1$ is equivalent to minimizing $\sum_{i=1}^{n}(\beta_0-\beta_1-\beta_1x_i)^2$. The method of least squares is referred as the method of estimating $\beta_0\ and\ \beta_1$ and β_2 by minimizing

$S(\beta_0,\beta_1)=(\beta_0-\beta_1-\beta_1x_i)^2$[11].

2.5.2 Least Squares Estimation

The study shows that the maximum likelihood estimation with assumption normality distribution implies to the least squares estimation.

If we want to get the line

$\beta_0+\beta_1x_i$

That is the closest to the points (x_i,y_i). The errors

$\varepsilon_i=y_i-\mu_i=y_i-\beta_0-\beta_1x_i\ ;i=(1,2,\dots,n)$ must be less than as possible, one of the way to satisfy that is minimizing the function:

$$S(\beta_0,\beta_1)=\sum_{i=1}^{n}\varepsilon_i{}^2=\sum_{i=1}^{n}(y_i-\mu_i)^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2 \qquad (2.6)$$

With respect to β_0,β_1.

This way takes the squared distance as a measure of proximity. The squared error loss is the function established from the maximum likelihood procedure.

Taking derivatives with respect to β_0, β_1 and setting the derivatives to zero for minimizing the errors.

$\frac{\partial(\beta_0,\beta_1)}{\beta_0} = -2\sum(y_i - \beta_0 - \beta_1 x_i) = 0$, and

$\frac{\partial(\beta_0,\beta_1)}{\beta_1} = -2\sum(y_i - \beta_0 - \beta_1 x_i)x_i = 0$

Implies to the two below equations:

$$n\beta_0 + (\sum xi)\beta_1 = \sum y_i \quad (2.7)$$

$$(\sum x_i)\beta_0 + (\sum x_i^2)\beta_1 = \sum x_i y_i \quad (2.8)$$

These two equations are called as the normal equations. Assume that $\hat{\beta}_0 \; and \; \hat{\beta}_1$ refers to the solutions for β_0, β_1 in the equations (2.7) and (2.8)

We can see that:

$$\beta_1 = \frac{\sum x_i y_i - [(\sum x_i)(\sum y_i)/n]}{\sum x_i^2 - [(\sum x_i)^2/n]} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \quad (2.9)$$

$$where \; S_{xx} = \sum(x_i - \bar{x})^2$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, where \; \bar{y} = \frac{\sum y_i}{n} \; and \; \bar{x} = \frac{\sum x_i}{n} \quad (2.10)$$

They are referred as the least squares estimates (LSE's) of $\beta_0 \; and \; \beta_1$, respectively[11].

2.6 Properties of Least Squares Estimation

We can see that from least squares estimation equations, that the following properties are satisfied:

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

$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} = \sum c_i y_i$, $where\ c_i = \frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$ is constant and have the following properties:

1. $\sum c_i = 0$
2. $\sum c_i x_i = 1$
3. $\sum c_i^2 = \frac{\sum(x_i - \bar{x})}{S_{xx}} = \frac{1}{S_{xx}}$; when $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, [11].

2.7 Expected Values of Least Squares Estimates

Obviously, those equation below show that $E(\hat{\beta}_0) = \beta_1$ is an unbiased estimator of β_1. This leads to that when experimental is repeated for many times, the average of estimates of $\hat{\beta}_1$ compatible with the true value β_1 [11].

$$1.\ E(\hat{\beta}_1) = E(c_i y_i) = \sum c_i\ (\beta_0 + \beta_1 x_i) = \beta_0 + \sum c_i + \beta_1 \sum c_i x_i = 0 + \beta_1 * 1$$
$$= \beta_1$$

$$2.\ E(\hat{\beta}_0) = E(\bar{y} - \beta_1 \bar{x}) = E(\bar{y}) - \bar{x}E(\hat{\beta}_1) = E(\bar{y}) - \beta_1 \bar{x}.$$

For

$$E(\bar{y}) = E\left(\frac{\sum y_i}{n}\right) = \left(\frac{\sum \beta_0 + \beta_1 x_i}{n}\right) = \beta_0 + \beta_1 \bar{x}.$$

Hence

$$E(\hat{\beta}_0) = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0.$$

$\hat{\beta}_0$ is unbiased for β_0

3. The LSE's of $\mu_0 = \beta_0 + \beta_1 x_0$ is given by $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ and

$E(\hat{\mu}) = \beta_0 + \beta_1 x_0 = \mu_0$

Hence

$\hat{\mu}_0$ is unbiased for μ_0.

4. It is easy to show that S^2 is unbiased estimator for σ^2, this means that

$E(S^2) = \sigma^2$[11].

2.8 Estimation of the Population Variance σ^2

Minimization of the likelihood function $L(\beta_0, \beta_1, \sigma^2)$ in equation (2.3) with respect

to σ^2 implies to the MLE

$$\sigma^2 = \frac{S(\hat{\beta}_0, \hat{\beta}_1)}{n}, S(\hat{\beta}_0, \hat{\beta}_1) = \sum \left[y_i - (\hat{\beta}_0, \hat{\beta}_1 x_i)^2 \right] = \sum e_i^2 \quad (2.11)$$

$\sum e_i^2$ is the residual sum of squares. The LSE of σ^2 is little different,

$$S^2 = \frac{S(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = MSE \quad (2.12)$$

2.9 Variance of Least Squares Estimation

Among all linear unbiased estimates The smallest value among all variance that will be found here is of β_j, the one with the smallest variance is the least square

estimation, [15].

1. $V(\hat{\beta}_1) = V(\sum c_i y_i) = \sum c_i^2 V(y_i) = \sum c_i^2 \sigma^2$, then, $V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

2. To calculate $V(\hat{\beta}_0)$, we will do as follows:

Let $k_i = \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}}$, then

$\hat{\beta}_0 = \sum k_i y_i$. So,

$V(\hat{\beta}_0) = \sum k_i^2 \sigma^2 = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}]$

3. $V(\hat{\mu}_0)$ will be:

$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} - \hat{\beta}_1 \bar{x} + \beta_1 x_0 = \bar{y} + \hat{\beta}_1(x_0 - \bar{x}) = \sum[\frac{y_i}{n} + (x_0 - \bar{x})\frac{(x_i - \bar{x})y_i}{S_{XX}}]$

$$= \sum[\frac{1}{n} + (x_0 - \bar{x})^2 \frac{(x_i - \bar{x})y_i}{S_{xx}}]y_i$$

$V(\hat{\mu}_0) = \sum_{i=1}^{n} d_i^2 \sigma^2 = \sigma^2[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]$ [11].

2.10 Inferences about the Regression Parameters

The uncertainties in the estimates can be shown out by confidence intervals, and for the researcher may need to make assumption about the distribution of errors.

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$

We assumed that the errors are normally distributed with mean equal to zero and variance equal to σ^2[11].

Chapter 3

Mathematical Modeling of Diabetics Incidence Rates

3.1 Statistical Analysis for Duhok Diabetics Data

In this chapter we will calculate some important descriptive statistics for (1385 diabetes patients) that we choose them randomly from Diabetes Center/ Duhok/ Kurdistan Region/ Republic of Iraq [15], using the Statistical Package for the Social Science (SPSS).

3.1.1 Descriptive: (weight, length and age)

SPSS Steps

Analyze → descriptive statistics → descriptive → drag (weight, length, age) to (variables) box → click option → choose (mean, std. deviation, min, max, variance, range) → continue → ok.

Table 1. Descriptive statistics table for age, weight and length.

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Age	1378	84	1927	2011	1962.04	12.812	164.156
Weight	1365	150	11	161	78.17	16.931	286.654
Length	1359	179	15	194	157.49	11.005	121.115
Valid N (listwise)	1347						

In the Descriptive Statistics table, we found (valid number, range, minimum values, maximum values, mean, standard deviation and variance) for every variable in our

study, we can see in the N column that we have a few missed values for every variables, the total number of missed values in our sample is equal to 38.

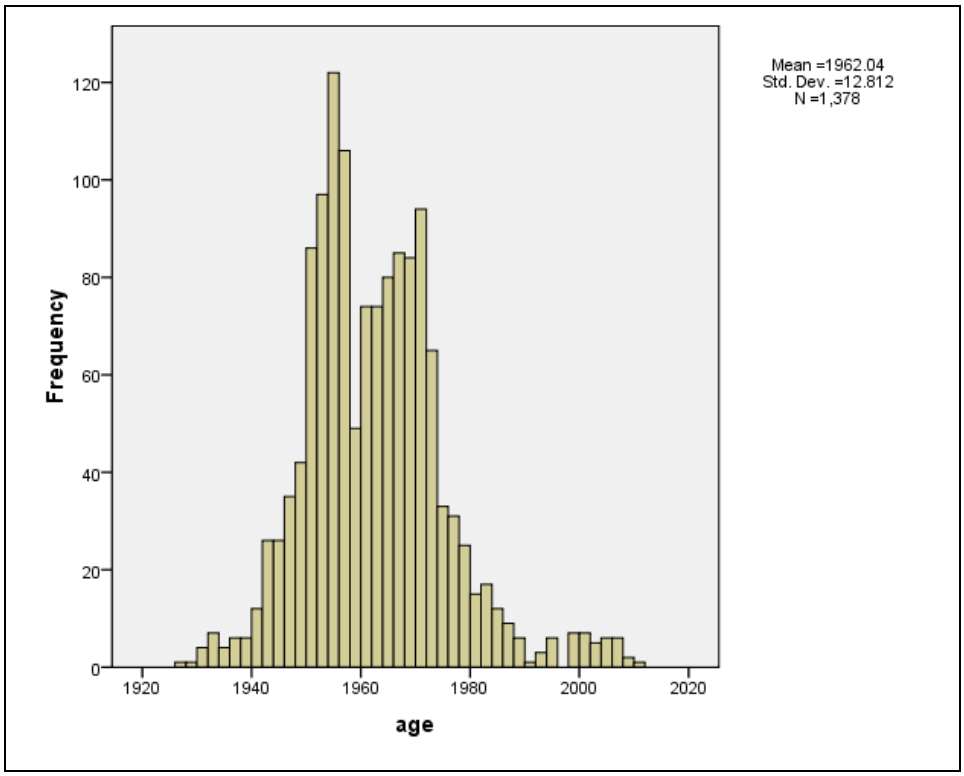

Figure 1. Histogram for Age

SPSS Steps

Graphs → legacy dialog→ interactive → histogram → drag (age binned) to the x-axis → click ok.

If we draw a curve in the above histogram we can note that the data does not follow the normal distribution, and we will see that more clearly later in this study by the test of normality. Histogram graph shows that the highest levels of the frequencies are between 1954-1956 years of birth.

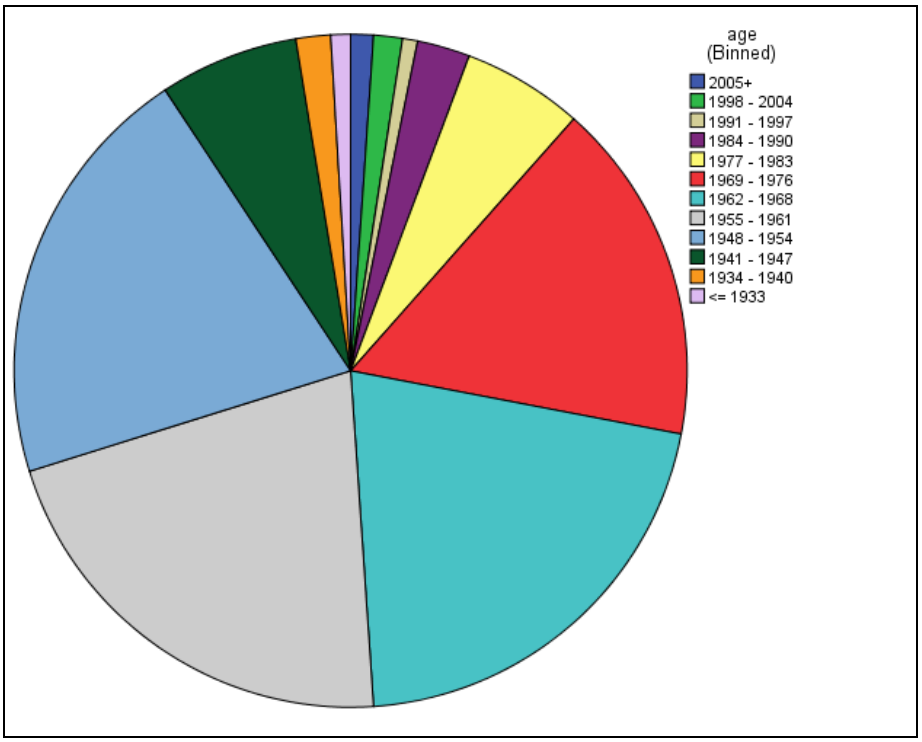

Figure 2. Pie Graph for Age

SPSS Steps

Graphs → legacy dialog → pie → appoint (summaries for groups of cases) → define → drag (patients age [binned]) to (define slices by) → click ok.

In the pie graph, using an (age binned) variables, it is clear that there is a great affinity between three categories (1948-1954, 1955-1961 and 1962-1968), or we can say that generally: those people aged in 2013 between (45-65) years old. Followed by red part of the pie figure, that represents the age group (1969-1976). Thus, we can say that the group in which the fewest number of patients that is painted by light brown color which is representative the group (1991-1997), or in another words: the people aged in 2013 between 16-22 years old. Overview, we can say that the number of diabetic patients in a significant decline in the place under study.

3.1.2 The quantile - quantile or Q-Q plot

The Q-Q plot (that called quantile - quantile) is a graphic tool used to show us if we assume the right distribution for our data. In general, this graphic tools works by computing the expected value for every data on the distribution. If the data follow the distribution, then the points should approximately fall on a straight line in Q-Q plot.

Figure 3. Normal Q-Q Plot for Age

Then, in the bellow Q-Q plot that represent our sample, we can see many points does not fall approximately on a straight line, and this means that the data does not follow our assumed distribution –The Normal Distribution-.

3.1.3 Normal Distribution Test

Now, we will tests if our data follows the normal distribution, with confidence =95%.

H_0: the data follows the normal distribution.

Ha: the data does not follow the normal distribution.

SPSS Steps

Analyze → Descriptive Statistics → Explore.

Drag (Patients age) to (Dependent List) box.

Click (Statistics) → put (✓) on (Descriptive).

Determine (confidence Interval for mean).

In our example we assumed the confidence interval is 95%.

Click ok.

Explore

Table 2. case processing summary table

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age	1378	99.5%	7	.5%	1385	100.0%

We have 1378 valid number of patients, and that is up by 99.5 from our sample, and 0.5% missed values.

Table 3. Descriptive statistics table

			Statistic	Std. Error
Age	Mean		1962.04	.345
	95% Confidence Interval for Mean	Lower Bound	1961.36	
		Upper Bound	1962.72	
	5% Trimmed Mean		1961.49	
	Median		1961.00	
	Variance		164.156	
	Std. Deviation		12.812	
	Minimum		1927	
	Maximum		2011	
	Range		84	
	Interquartile Range		17	
	Skewness		.721	.066
	Kurtosis		1.359	.132

We can see that in table 3, the age mean is 1962, and we are sure with confidence 95% that the sample mean lies between 1961.36 to 1962.72 and the median for our sample is 1961, the first birth year is 1927, and the last birth year is 2011. The range is equal to 84.

Table 4. Tests of Normality table

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	Df	Sig.
Age	.075	1378	.000	.965	1378	.000

a. Lilliefors Significance Correction

It is clear that from both Kolmogrov-Smirnov and Shapiro-wilk tests that significance value=0 < confidence interval $\propto$ = 0.05. So, we will reject null

hypothesis H_0 and this means that there is a significant difference and this reassurance our note above in Q-Q plot.

3.2 Frequencies[(gender, diabetes type, date of diabetes, acquisition and age (binned)]

SPSS Steps

Analyze → descriptive statistics → frequencies → drag (gender, diabetes type, date of diabetes, weight, length, acquisition,) → click statistics → choose the measurements that we want to measure → continue → ok.

Frequencies

Table 5. Diabetes type statistics table

N	Valid	1385
	Missing	0

Table 6. Diabetes type frequency table

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		24	1.7	1.7	1.7
	T1	72	5.2	5.2	6.9
	T2	1289	93.1	93.1	100.0
	Total	1385	100.0	100.0	

In the diabetes type we have sample size = 1385, and we have 24 missed values which represent 1.7% of the total sample number, then the total number of valid values is 1361. There are 72 person suffer from Type1 diabetes and this is represent 5.2%, and 1289 person suffer from Type2 diabetes and this is represent 93.1%. It is clearly that Type2 of diabetes is the most widespread.

3.3 Age Groups

SPSS Steps

Transform → visual binning → drag (age) to (variable to bin) → continue → write name in (binned variable) → click (make cut point) → first cut point location = the cut point value of the first class → click make labels → click ok.

Number of cut point $= 1 + 3.322 \log(1378) \approx 12.$

width $= \frac{range}{k} = \frac{2011-1927}{12} = 7.$

Table 7. Age group information

N	Valid	1378
	Missing	7
Mean		7.42
Median		8.00
Mode		8
Range		11
Minimum		1
Maximum		12

Table 8. Age group statistics

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2005+	15	1.1	1.1	1.1
	1998 – 2004	19	1.4	1.4	2.5
	1991 – 1997	10	.7	.7	3.2
	1984 – 1990	35	2.5	2.5	5.7
	1977 – 1983	80	5.8	5.8	11.5
	1969 – 1976	227	16.4	16.5	28.0
	1962 – 1968	288	20.8	20.9	48.9
	1955 – 1961	293	21.2	21.3	70.2
	1948 – 1954	283	20.4	20.5	90.7
	1941 – 1947	92	6.6	6.7	97.4
	1934 – 1940	23	1.7	1.7	99.1
	<= 1933	13	.9	.9	100.0
	Total	1378	99.5	100.0	
Missing	System	7	.5		
Total		1385	100.0		

In the age (Binned) table, we found that the most ages affected with diabetes are bounded in the following groups: 1948-1954, 1955-1961, 1962-1968 and 1969-1976, with 283, 293, 288, 227 respectively. There is a small increase in the group 1955-1961. This means that those people whose aged between 65 to 37 years old in 2013. They are 1091 patients and they are represents 79.2% of our sample, contained in the mentioned groups.

Table 9. Gender Frequency Table

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	F	868	62.7	62.7	62.7
	M	517	37.3	37.3	100.0
	Total	1385	100.0	100.0	

It is clear that the number of females with diabetes is greater than the number of males in our random sample, there are 868 female with diabetes and this represent 62.7 % from all the sample, and 517 male with diabetes and they are represent 37.3% in the same sample.

Table 10. Weight group statistics table

N	Valid	1365
	Missing	20
Mean		5.92
Std. Error of Mean		.036
Median		6.00
Mode		6
Std. Deviation		1.328
Range		11

$$K = 1 + 3.322 \log(n)$$

$$K = 1 + 3.322 \log(1365) = 11.414 \approx 11$$

When

K: is suggests number of class

n: is the valid total of weight in this data.

$$width = \frac{maximum\ value - minimum\ vale}{k} = \frac{161 - 11}{11} = 13.636 \approx 13$$

Table 11. Weight group statistics table

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< 21	10	.7	.7	.7
	21 – 33	17	1.2	1.2	2.0
	34 – 46	11	.8	.8	2.8
	47 – 59	94	6.8	6.9	9.7
	60 – 72	344	24.8	25.2	34.9
	73 – 85	489	35.3	35.8	70.7
	86 – 98	268	19.4	19.6	90.3
	99 – 111	98	7.1	7.2	97.5
	112 – 124	24	1.7	1.8	99.3
	125 – 137	7	.5	.5	99.8
	138 – 150	2	.1	.1	99.9
	151+	1	.1	.1	100.0
	Total	1365	98.6	100.0	
Missing	System	20	1.4		
Total		1385	100.0		

We have 20 missed weights information in our sample size caused by registry errors. We note that there is no big numbers of patients at the weights 11 kg to 46 kg where there are vary between an one patient to two patients. Then, patients' number began increasing progressively for weights at 47 kg to 53 kg. Then there are noticeable increase for patients weights from 54 kg to 93 kg and gradually begins to escalate gradually with varying degrees until it reaches the number to 43 patients at 70 kg, then reported the highest number of patients at both 81 kg and 83 kg where the number of patients reaches there to 84 patients. Then, patients' number began decreasing and varying progressively for weights at 94 kg until reaches to one patients at weights 137 kg to 161 kg. As we see in the weight histogram.

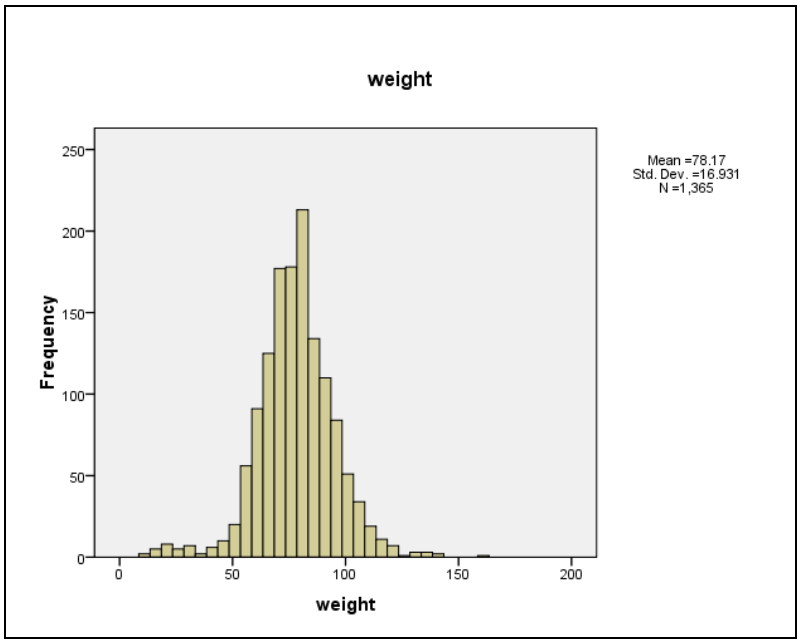

Figure 4. Weight Histogram

Table 12. Length group statistics table

N	Valid	1359
	Missing	26
Mode		15

We see in table 12 that valid number of patients lengths equal to 1359 value, with 26 missed values.

Table 13. Length binned table

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< 24	1	.1	.1	.1
	84 – 93	1	.1	.1	.1
	94 – 103	1	.1	.1	.2
	104 – 113	5	.4	.4	.6
	114 – 123	6	.4	.4	1.0
	124 – 133	6	.4	.4	1.5
	134 – 143	40	2.9	2.9	4.4
	144 – 153	425	30.7	31.3	35.7
	154 – 163	484	34.9	35.6	71.3
	164 – 173	318	23.0	23.4	94.7
	174 – 183	70	5.1	5.2	99.9
	184+	2	.1	.1	100.0
	Total	1359	98.1	100.0	
Missing	System	26	1.9		
Total		1385	100.0		

In the above statistical lengths table, there are 26 missed value for lengths, and this missed data may have been caused by registry errors.

Also, we can see that there are approximate equal numbers of patients beginning from 15 cm to 133 cm where the numbers is varying between one to six patients. Those numbers are progressively varies escalate at 143 cm to 173 cm where the number is varying among 6 patients at 140 cm then peaking 153 cm with 70 patients and then varies decreasing again at 177 cm to 194 cm with 10 patients to 1 patients.

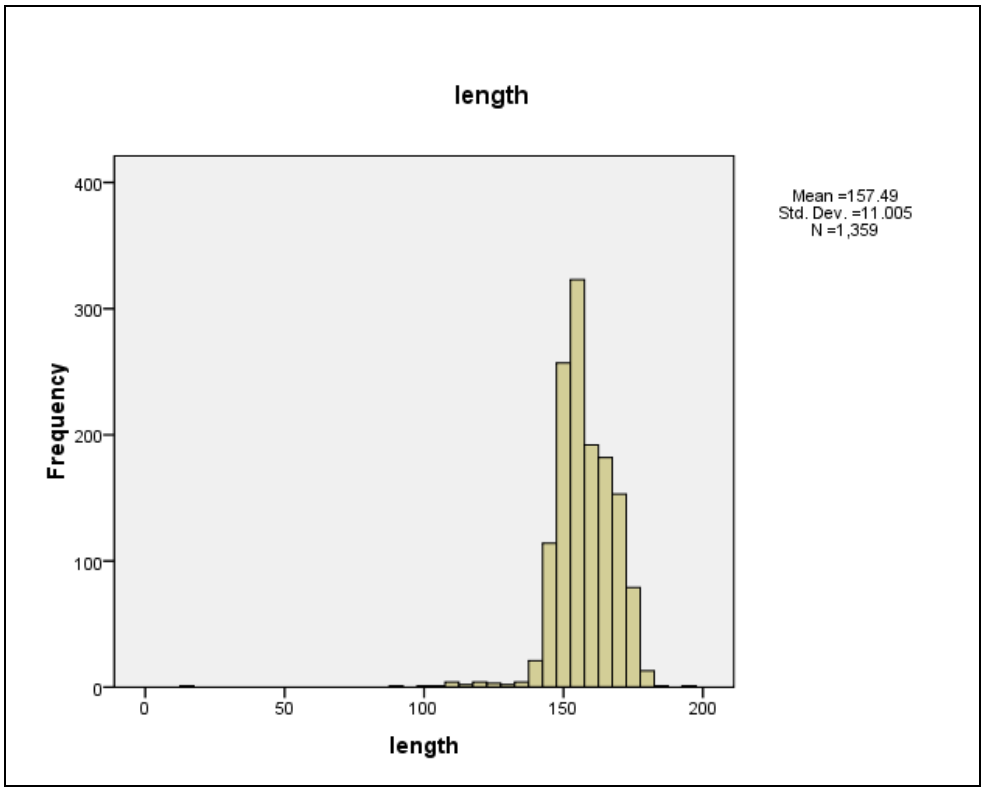

Figure 5. Length Histogram

Table 14. Date of diabetes information

N	Valid	1300
	Missing	85

In the date of diabetes information table; we can see that we have 1300 valid number of patients our sample, and 85 missed number of patients.

Table 15. Date of diabetes frequency table

	Frequency	Percent	Valid Percent	Cumulative Percent
	85	6.1	6.1	6.1
1980	1	.1	.1	6.2
1982	2	.1	.1	6.4
1986	1	.1	.1	6.4
1987	2	.1	.1	6.6
1988	1	.1	.1	6.6
1989	1	.1	.1	6.7
1990	1	.1	.1	6.8
1991	6	.4	.4	7.2
1992	3	.2	.2	7.4
1993	10	.7	.7	8.2
1994	2	.1	.1	8.3
1995	2	.1	.1	8.4
1996	5	.4	.4	8.8
1997	17	1.2	1.2	10.0
1998	27	1.9	1.9	12.0
1999	7	.5	.5	12.5
2000	37	2.7	2.7	15.2
2001	5	.4	.4	15.5
2002	42	3.0	3.0	18.6
2003	68	4.9	4.9	23.5
2004	10	.7	.7	24.2
2005	45	3.2	3.2	27.4
2006	34	2.5	2.5	29.9
2007	52	3.8	3.8	33.6
2008	84	6.1	6.1	39.7
2009	89	6.4	6.4	46.1
2010	110	7.9	7.9	54.1
2011	150	10.8	10.8	64.9
2012	384	27.7	27.7	92.6
2013	102	7.4	7.4	100.0

In the date of diabetes table, we have no missed data in our sample. Since 1980 to 1990, number of patients escalates between 1 to 2 patients. Patients number began escalate increasing from 1991 to the current year 2013. We note that in recent years, diabetics patients is increasing continuously, beginning in 2007 with 52 patients, 2008 with 84 patients, 2009 with 89 patients, 2010 with 150 patients, and the number of patients peaked in 2012 with 384 patients.

$$k = 1 + 3.322\log(1300) = 5.9 \approx 6$$

$$width = \frac{2013 - 1980}{6} = 5.5 \approx 6$$

Table 16. Date of diabetes group table

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< 1988	6	.4	.5	.5
	1988 – 1992	12	.9	.9	1.4
	1993 – 1997	36	2.6	2.8	4.2
	1998 – 2002	118	8.5	9.1	13.2
	2003 – 2007	209	15.1	16.1	29.3
	2008+	919	66.4	70.7	100.0
	Total	1300	93.9	100.0	
Missing	System	85	6.1		
Total		1385	100.0		

Table 17. Acquisition types statistics table

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		84	6.1	6.1	6.1
	N	489	35.3	35.3	41.4
	P	812	58.7	58.	100.0
	Total	1385	100.0	100.0	

Type acquire the disease table show us if the diabetes patients acquired genetically diabetes or by other causes like obesity, beta-cell damage, pancreatic damage or other causes. P means positive, this means that the acquisition of a genetic disease. N means negative, this means that the acquisition of other causes. We see in the above table that diabetes patients number with positive diagnosis is 812, that is 58.7%, and patients with diabetes number with negative diagnosis is 489 patients, that is 35.3%. Now we can say that the genetic cause is the stronger cause making people have the diabetes.

We can separate variables (males and females) as shown below.

Steps:

Data → Split Files → drag (gender) to (Groups Based On) → Choose (organize output by groups) → ok.

Now, to calculate descriptive statistics for two variables (like males and females) we will follow this steps:

Analyze → descriptive statistics → crosstabs → drag (gender) to ($Row_{(s)}$ box) → drag other variables to ($column_{(s)}$) box → ok.

3.4 Correlations

In this part, we try to find the correlation coefficient between our variables, to illustrate what is the importance of each variable in this study.

3.4.1 Gender Correlation

For both of them (male and female), we will try to find the correlation coefficient each separately.

3.4.1.1 Gender = Female

Now, the correlation coefficient for female with age, weight, length and date of diabetes variables will be shown in table 18.

Table 18. Correlations table for females with age, weight, length and date of diabetes

		Age	Weight	Length	Dateofdiabetes
Age	Pearson Correlation	1	-.007-	-.076-*	.278**
	Sig. (2-tailed)		.834	.027	.000
	N	863	850	845	811
Weight	Pearson Correlation	-.007-	1	.304**	.156**
	Sig. (2-tailed)	.834		.000	.000
	N	850	855	847	805
Length	Pearson Correlation	-.076-*	.304**	1	-.028-
	Sig. (2-tailed)	.027	.000		.431
	N	845	847	850	800
Dateofdiabetes	Pearson Correlation	.278**	.156**	-.028-	1
	Sig. (2-tailed)	.000	.000	.431	
	N	811	805	800	816

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

a. gender = 1

First, we want to mention that the total number of females with diabetes in our sample is 868. We note that the age variable have a very high positive linear relationship with female equal to +1. We can see also that we have n= 863 females, this means that there are 5 missed values.

The correlation coefficient between females and their weights is -0.007, this means that almost there is no linear relationship between them. We have n= 850, this means that there are 18 missed values.

The correlation coefficient between female and their lengths equal to -0.076 and it is means that almost there is no relationship.

For the linear relationship between female and date of diabetes acquire, the correlation coefficient is 0.278 and it is a weak positive relationship.

3.4.1.2 Correlation Gender = Male

The correlation coefficient for male with their ages, weights, lengths and date of diabetes variables will be shown in this part.

Table 19. Correlation table for male with age, weight, length and date of diabetes

		Age	weight	length	Dateofdiabetes
Age	Pearson Correlation	1	-.323-**	-.309-**	.212**
	Sig. (2-tailed)		.000	.000	.000
	N	515	508	507	482
Weight	Pearson Correlation	-.323-**	1	.618**	.025
	Sig. (2-tailed)	.000		.000	.589
	N	508	510	507	477
Length	Pearson Correlation	-.309-**	.618**	1	-.055-
	Sig. (2-tailed)	.000	.000		.234
	N	507	507	509	476
Dateofdiabetes	Pearson Correlation	.212**	.025	-.055-	1
	Sig. (2-tailed)	.000	.589	.234	
	N	482	477	476	484

**. Correlation is significant at the 0.01 level (2-tailed). a. gender = 2

From table 19, that shows us the correlation coefficient between male and their ages, weight, length and date of diabetes. At first we should mention that the total number of males with diabetes in our sample is 517.

The correlation coefficient between male and their ages is equal to +1, and this means that there is strong linear relationship between males with diabetes and their ages. We have n= 515 patients, and this means that there are just two missed values.

The correlation coefficient between males with diabetes and their weights is -0.323, and it is weak negative linear relationship, n=508, so we have 9 missed values.

Also, between males with diabetes and their lengths, the correlation coefficient is -0.309, and it is weak negative linear relationship, and we have 10 missed values.

The correlation coefficient between male with diabetes and their acquire diabetes is 0.212, and it is also weak positive linear relationship, n=482, so we have 35 missed values.

3.4.2 Correlation Coefficient for Diabetes Type

We try to understand the relationship between both types of diabetes (diabetes type1, diabetes type2) with their ages, weights, lengths and date of diabetes acquire.

Diabetestype = Type1

First, we will find the correlation coefficient for patients with diabetes typ1 and their ages, weights, lengths and date of diabetes acquire. We also want to mention that the total number of diabetics with type1 is 72 and with type2 is 1289 patients in our sample.

Table 20. Correlation table for diabetes type with age, weight, length and date of diabetes

		Age	weight	length	Dateofdiabetes
Age	Pearson Correlation	1	-.803-**	-.633-**	.521**
	Sig. (2-tailed)		.000	.000	.000
	N	72	71	69	68
Weight	Pearson Correlation	-.803-**	1	.739**	-.404-**
	Sig. (2-tailed)	.000		.000	.001
	N	71	71	68	67
Length	Pearson Correlation	-.633-**	.739**	1	-.234-
	Sig. (2-tailed)	.000	.000		.061
	N	69	68	69	65
Dateofdiabetes	Pearson Correlation	.521**	-.404-**	-.234-	1
	Sig. (2-tailed)	.000	.001	.061	
	N	68	67	65	68

**. Correlation is significant at the 0.01 level (2-tailed).
a. diabetestype = 1

Patients with diabetes type1 have a very strong positive relationship with their ages, that is the correlation coefficient equal to +1, we have n=72 and this means that there is no missed values.

The correlation coefficient between patients with diabetes and their weights is -0.803 it is strong negatively linear relationship, we have only one missed value.

Also, the correlation coefficient between patients with diabetes type1 and their lengths is -0.633, this is a moderate negative linear relationship, and we have 3 missed values.

There is a moderate positive relationship equal to 0.521 between patients with diabetes and their date of diabetes acquires, we have 4 missed values.

Diabetes Type = Type2

In this step, we will try to illustrate the relationship between patients with diabetes type2 and their ages, weights, lengths and date of diabetes. We should mention that the total number of patients with diabetes type2 in our sample is 1289.

Table 21. Correlation table for diabetes type2 with age, weight, length and date of diabetes

		age	weight	length	Dateofdiabetes
Age	Pearson Correlation	1	.215**	.149**	.255**
	Sig. (2-tailed)		.000	.000	.000
	N	1282	1263	1259	1209
Weight	Pearson Correlation	.215**	1	.281**	.182**
	Sig. (2-tailed)	.000		.000	.000
	N	1263	1270	1262	1199
Length	Pearson Correlation	.149**	.281**	1	.042
	Sig. (2-tailed)	.000	.000		.144
	N	1259	1262	1266	1195
Dateofdiabetes	Pearson Correlation	.255**	.182**	.042	1
	Sig. (2-tailed)	.000	.000	.144	
	N	1209	1199	1195	1216

**. Correlation is significant at the 0.01 level (2-tailed).
a. diabetestype = 2

We see that there is a strong positive linear relationship equal to +1 between patients with diabetes type2 and their ages, we have 7 missed values.

The correlation coefficient between patients with type2 diabetes equal to 0.215 and it is weak positive linear relationship, and there are 26 missed values.

For patients with type2 diabetes and their lengths, there is also positive weak relationship equal to 0.149, and there are 30 missed values.

The correlation coefficient between patients with diabetes type2 and their date of diabetes acquire is 0.255, it is weak positive linear relationship and we have 80 missed values.

3.4.2 Crosstabs

Table 22. Case processing summary table for gender with all other variables

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
gender * age	1378	99.5%	7	.5%	1385	100.0%
gender * diabetestype	1385	100.0%	0	.0%	1385	100.0%
gender * weight	1365	98.6%	20	1.4%	1385	100.0%
gender * length	1359	98.1%	26	1.9%	1385	100.0%
gender*dateofdiabete	1385	100.0%	0	.0%	1385	100.0%
gender * Acquisition	1385	100.0%	0	.0%	1385	100.0%

Cross tabs helps us to obtain all information about single variable values related to another variables, one by one.

In the table22, we can see that valid and missed number of patients for every variable related with gender variable.

Table 23. Gender with diabetes type cross table

		Diabetestype			
			T1	T2	Total
gender	F	13	29	826	869
	M	11	43	463	517
Total		24	72	1289	1385

We can see from the above table that the number of males with diabetes type1 is more than female number, but for type2 of diabetes, the number of females is more than number of male.

Table 24. Gender with acquisition cross table

		Acquisition			Total
			N	P	
gender	F	51	278	539	868
	M	33	211	273	517
Total		84	489	812	1385

As we mentioned before, there are two types of diabetes acquisition, positive and negative. In table 24, we see that for both positive and negative results, the number of female patients is more than number of male patients.

3.5 Comparison between Duhok and Cyprus Patients with Type1 Diabetes Less Than Fifteen Years Old

We will make a comparison analysis between Duhok females with diabetes type1 who are less than 15 years old, and Cyprus females with diabetes type1 who are less than 15 years old. [16]

3.5.1 T-Test

Now, we will apply T-test on our sample, and we used T-test because we do not know the standard deviation of the population.

3.5.1.1 Paired Sample Statistics for Duhok females and Duhok males

Test hypothesis mean to see that if Duhok female mean equal or not to Duhok male mean.

$H_0: \mu_{DF} = \mu_{DM}$

$H_0: \mu_{DF} \neq \mu_{DM}$

Where;

μ_{DF}: The mean of Duhok female

μ_{DM}: The mean of Duhok male.

Table 25. Paired samples statistics table

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Dfemale	4.25	20	2.900	.648
	Dmale	5.10	20	2.269	.507

Table 26. Paired samples correlations table

		N	Correlation	Sig.
Pair 1	Dfemale & Dmale	20	.716	.000

There is a high positive correlation between Duhok female and Duhok male equal to 0.716 and significant value approximately equal to 0.000 < 0.05.

Table 27. Paired samples test table

	Paired Differences					T	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% CI				
				Lower	Upper			
Dfemale – Dmale	-.850-	2.033	.455	-1.802-	.102	-1.870-	19	.077

We cannot reject H_0, this means that we accept that mean of D_f and mean of D_m are equal with 95% confidence.

3.5.1.2 T-Test One Sample T-Test for Duhok females and Cyprus females

$H_0: \mu_{DF} = 9.1$

$H_1: \mu_{DF} \neq 9.1$

Where;

9.1: Is the mean of Cyprus female.

Table 28. One-Sample Statistics table

	N	Mean	Std. Deviation	Std. Error Mean
Dfemale	20	4.25	2.900	.648

Table 29. One-Sample Test table

	Test Value = 9.1					
	T	Df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Dfemale	-7.480-	19	.000	-4.850-	-6.21-	-3.49-

P-value approximately =0.000 < 0.05. We reject H_0 with 95% confidence. It means that the mean of D_F is different from the mean of $C_F = 9.1$.

3.5.2 One Way

We use one-way test because we have nominal variable (gender), we want to know if the female mean for both Cyprus and Duhok are equals or not, so we will use ANOVA table that is used for testing two or more means and it is considered as extension for T-test.

3.5.2.1 ANOVA for Duhok Female with Cyprus Female

Table 30 tests the following hypothesis using ANOVA table.

$\sigma_{DF}^2 = \sigma_{CF}^2$

$\sigma_{DF}^2 \neq \sigma_{CF}^2$

Where;

σ_{DF}^2 is the Duhok female population

σ_{CF}^2 is Cyprus female population

Table 30. ANOVA table for Duhok female

	Sum of Squares	d.f	Mean Square	F	Sig.
Between Groups	116.550	9	12.950	2.998	.051
Within Groups	43.200	10	4.320		
Total	159.750	19			

We see that P-value is 0.051 > 0.050. So we will accept H_0 that is the variance of Duhok females equal to the variance of Cyprus females.

3.6 Mathematical Modeling of Diabetics Incidence Rates

3.6.1 Outlier Data

To obtain the more fitting regression model, we used this statistical technique to neglect the anomalous data; we will try first to excluding the anomalous values as shown in the table 36.

Table 31. Case processing summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
DateOfDiabetes	29	100.0%	0	0.0%	29	100.0%

Table 32. Extreme Values

		Case Number	Value
DateOfDiabets	Highest 1	29	2012
	2	28	2011
	3	27	2010
	4	26	2009
	5	25	2008
	Lowest 1	1	1980
	2	2	1982
	3	3	1986
	4	4	1987
	5	5	1988

We can see the first five years starting at 1980 to 1988 have only one patient, so we will excluding those years from our mathematical modeling.

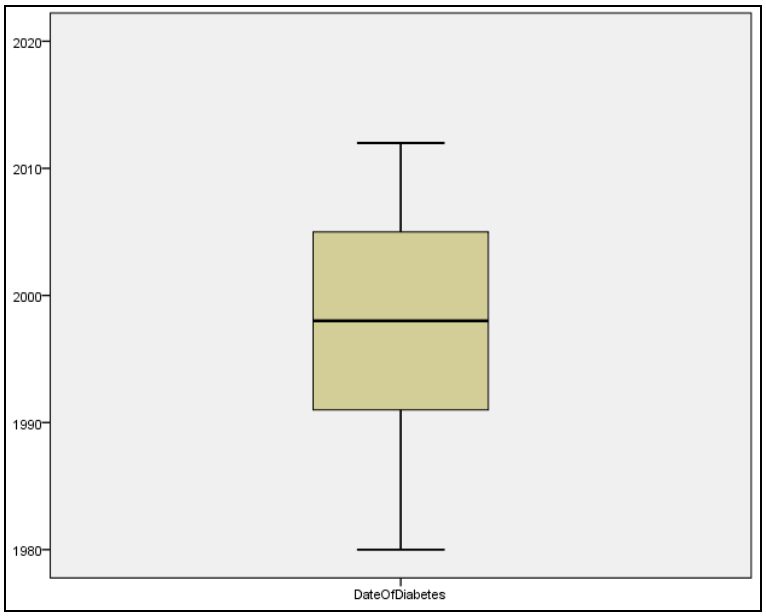

Figure 6. Date of Diabetes Outliers

Curve Fit for Linear, Logarithmic, Inverse and Exponential Equations

In this part, we will try to find the fit mathematical model regression of diabetes incidence in Duhok/ Kurdistan region of Iraq, using the Statistical Package for the Social Science (SPSS).We assume that the dependent variable is the number of patients and the independent variable is the date of diabetes.

Table 33. Model description table

Model Name		MOD_1
Dependent Variable	1	NumberOfPatients
Equation	1	Linear
	2	Logarithmic
	3	Inverse
	4	Exponential[a]
Independent Variable		Yearrank
Constant		Included
Variable Whose Values Label Observations in Plots		Unspecified

a. The model requires all non-missing values to be positive.

We can see in table36 that we have one dependent variable and one independent variable, also we tried in this step to find four equation models and which one among them is the best equation model.

Table 34. Cases statistics

	N
Total Cases	21
Excluded Cases[a]	0
Forecasted Cases	0
Newly Created Cases	0

Total cases in our sample is 21 cases, there is no excluded or forecasted or newly case.

Table 35. Variable processing summary

		Variables	
		Dependent	Independent
		NumberOfPatients	yearrank
Number of Positive Values		23	23
Number of Zeros		0	0
Number of Negative Values		0	0
Number of Missing Values	User-Missing	0	0
	System-Missing	0	0

In the table 38, we can see that there are 21 positive values for both dependent and independent variables, and there is no zero or negative or missing values.

Table 36. Model Summary and parameter estimates table

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.469	18.581	1	21	.000	-48.518-	8.355
Logarithmic	.270	7.782	1	21	.011	-65.016-	52.035
Inverse	.097	2.256	1	21	.148	71.393	-121.051-
Exponential	.783	75.687	1	21	.000	1.597	.206

The independent variable is yearrank.

From R square column, we see that the first three models (Linear, Logarithmic and Inverse) have a normal positive correlation approximately equal to 0.41, but the exponential model has a more high positive correlation approximately equal to 0.83.

3.6.2 Curve Fit for Linear Equation

Now we will find each equation separately, first we will start with linear equation model.

Table 37. Descriptive Linear equation table

Model Name		MOD_2
Dependent Variable	1	NumberOfPatients
Equation	1	Linear
Independent Variable		Yearrank
Constant		Included
Variable Whose Values Label Observations in Plots		Unspecified

There is one dependent variable that is the number of patients and one independent variable that is the date of diabetes.

Table 38. Linear Model Summary and Parameter Estimates table

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.469	18.581	1	21	.000	-48.518-	8.355

The independent variable is year rank.

In the table 41, we note that there is a positive correlation coefficient equal to 0.409. Significant value approximately to 0.000 and it is less than 0.05 this means that there is a significant linear correlation. Our linear equation model is:

$$\hat{Y} = -48.518 + 8.355(x)$$

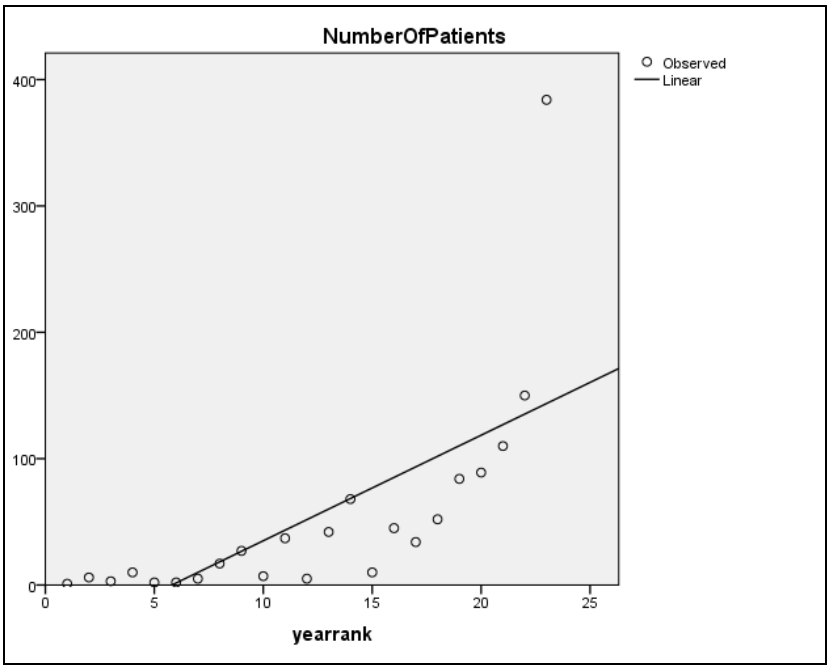

Figure 7. Linear Equation Plots

3.6.3 Curve Fit Logarithmic

Table 39. Logarithmic model summary and parameters estimates table

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Logarithmic	.270	7.782	1	21	.011	-65.016-	52.035

The independent variable is year rank.

Also, in the Logarithmic curve fit we have positive correlation coefficient equal to 0.408. Also, we have significant value approximately equal to 0.011 and it is less than 0.05, this means that there is a significant logarithmic correlation. Our logarithmic equation model is:

$$\hat{Y} = -65.016 + 52.035\ Log(x)$$

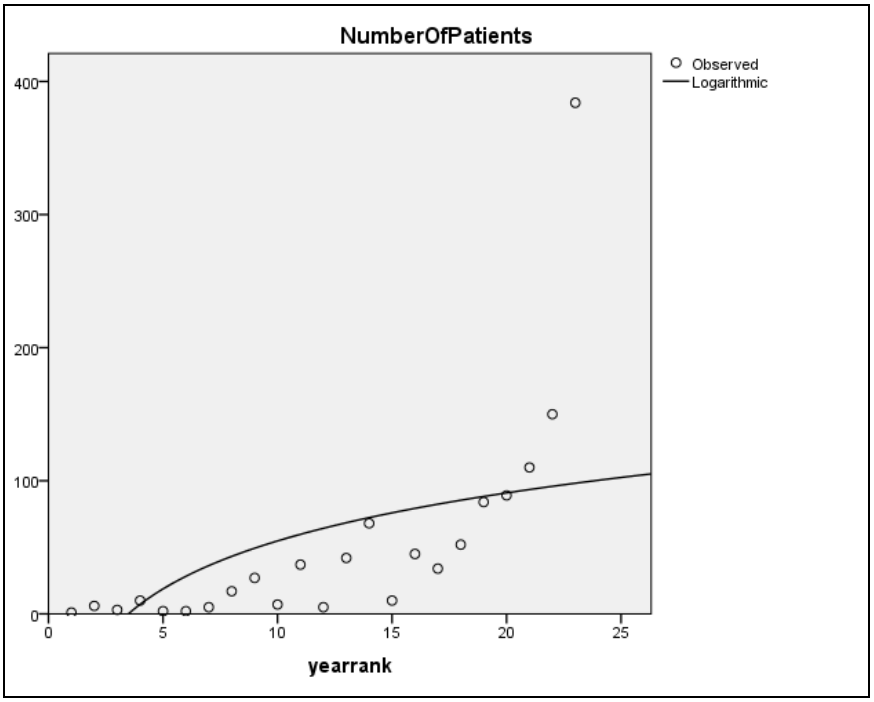

Figure 8. Logarithmic Equation Plots

3.6.4 Curve Fit for Inverse Equation

Table 40. Inverse model summary and parameters estimates table

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Inverse	.097	2.256	1	21	.148	71.393	-121.051-

The independent variable is year rank.

In table43 correlation coefficient are positive and equal to 0.406. Significant value is approximately equal to 0.148 and more than 0.05 this means that there is no significant Inverse regression. Our Inverse equation model is:

$$\hat{Y} = 71.393 + \frac{-121.051}{x}$$

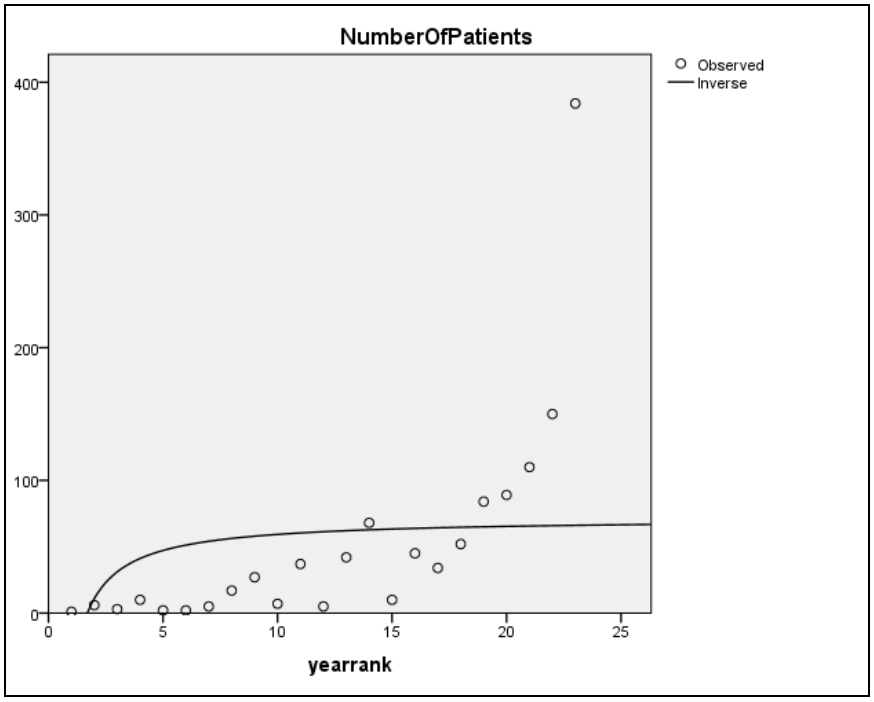

Figure 9. Inverse Equation Plots

3.6.5 Curve Fit for Exponential Equation

Table 41. Inverse model summary and parameters estimates table

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Exponential	.783	75.687	1	21	.000	1.597	.206

The independent variable is yearrank.

Dependent Variable: NumberOfPatients.

We note that from table 43 that there is a high positive correlation coefficient equal to 0.826 and it is the highest one among other three equations models. Significant value is approximately equal to 0.000 and it is less than 0.05 this means that there is a significant exponential regression. The exponential equation model is the best equation among other three equations (linear, logarithmic and Inverse).

$$\hat{Y} = 1.597 * e^{0.206\,(x)}$$

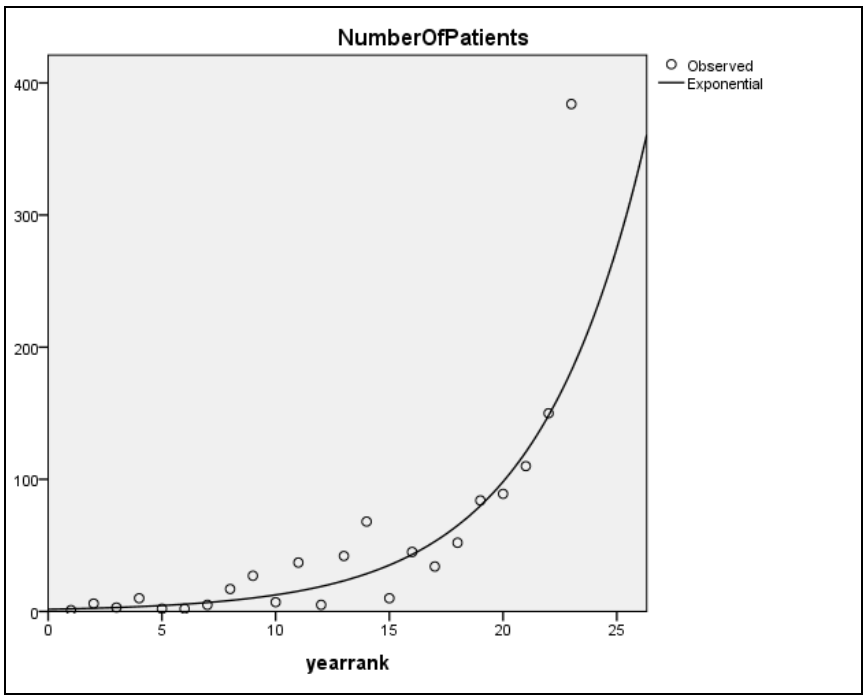

Figure 10. Exponential Equation Plots

Table 42. Expected number of patients and error between expecting and original values

Date	Year	Num	Fit-1	ERR-1
1990	1	1	1.96192	-.96192-
1991	2	6	2.41063	3.58937
1992	3	3	2.96197	.03803
1993	4	10	3.63940	6.36060
1994	5	2	4.47176	-2.47176-
1995	6	2	5.49449	-3.49449-
1996	7	5	6.75114	-1.75114-
1997	8	17	8.29518	8.70482
1998	9	27	10.19237	16.80763
1999	10	7	12.52346	-5.52346-
2000	11	37	15.38769	21.61231
2001	12	5	18.90700	-13.90700-
2002	13	42	23.23121	18.76879
2003	14	68	28.54441	39.45559
2004	15	10	35.07278	-25.07278-
2005	16	45	43.09426	1.90574
2006	17	34	52.95032	-18.95032-
2007	18	52	65.06056	-13.06056-
2008	19	84	79.94052	4.05948
2009	20	89	98.22367	-9.22367-
2010	21	110	120.68834	-10.68834-
2011	22	150	148.29090	1.70910
2012	23	384	182.20642	201.79358

Where;

Date: means the real date of diabetes

Year: refers to the rank for every year

Fit-1: the expected number for patients for every year

ERR-1: the expected error between real number of patients and expected number.

Exponential Curve Fit

We want to test wither $\beta_1 = \beta_2 = 0$ or not, so we will use ANOVA table to proof the following hypothesis testing.

$H_0 = \beta_1 = \beta_2 = 0$

$H_1 = \beta_1 \neq \beta_2 \neq 0$

Table 43. ANOVA table for parameter

	Sum of Squares	d.f.	Mean Square	F	Sig.
Regression	42.930	1	42.930	75.687	.000
Residual	11.911	21	.567		
Total	54.842	22			

The independent variable is yearrank.

We can see that the significant value P-value is approximately = 0.000 $<$ 0.05 so we will reject H_0. This means that our regression model is significant.

Chapter 4

CONCLUSION

- The more people with diabetes in our sample are those aged between 45 to 65 years old.
- Number of female with diabetes is more than number of male in the sample, the proportion of females in the sample reaches to 62.7%.
- Most number of people with diabetes is those people whose weights among 73-85 kg with proportion 35.8%.
- Most number of people with diabetes are those people whose lengths among 154-163 cm with proportion 30.7%.
- In the recent years the number of people in a continuous increase, where the highest rates of infection in the period 2008-2013 by up to 66.4% of the sample size.
- People with genetic diabetes represents 58.7% of our sample, and people with other diabetes causes by up to 41.3%.
- The number of male with diabetes type1 by up to 0.59 is more than the number of female. At the same time, the number of female by up to 0.64 with diabetes type2 is more than the number of male.
- In both, genetic diabetes cause and other diabetes cause, the number of females are more than the number of males.
- We note that from correlation coefficient analysis parts, that age variable has strong positive linear relationship with both types of diabetes, weights, lengths and date of diabetes variables equal to +1.

- In the correlation coefficient part, we also see that diabetes type1 have strong negative relationship with weight equal to -0.803.
- Regression modelling for number of patients in Duhok per year with 95% confidence interval is:

  $$\hat{Y} = 1.597 * e^{0.206\,(x)}$$

- The mean for both Duhok male and Duhok female aged less than 15 years old with diabetes type1 are equal to each other.
- Comparison of Cypriot female with diabetes type1 patients who less than 15 years old and Duhok female with the same ages and type of diabetes, with confidence interval equal to 90%, we see that the mean of both of them are equal to each other, but because the lack of enough information about males it is not possible arguments about males.
- Maybe the reason of increasing in the number of patients with diabetes in Duhok city in recent years is because the disturbed changes in the region economy starting from 1980 to 1988 through the Iraq-Iran war, and the economy siege from 1990 to 2003. Then, détente this economic crisis after 2003 and speeding citizens. All of these reasons may be affects on the high increasing the number of people with diabetes at recent years.
- We recommend researchers in their future studies to take into consideration the effects of economic changes, climate changes and social situation of people with diabetes.
- We recommend that there be more statistical studies for row data of people with diabetes, and the comparisons that shows the differences between the number of diabetics in different regions, to determine the causes of acquiring

the disease or the causes that leads to an increasing in the incidence of the diabetes.

REFERENCES

[1] Regression Models for Categorical and Limited Dependent Variables, J. Scott Long.

[2] International Diabetes Federation, www.idf.org/types-diabetes.

[3] Same last source.

[4] The American Diabetes Association ADA www.diabetes.org

[5] Type 1 Diabetes in Children Adolescents and Young Adults, Third Edition, D. Ranger Hanas, Page 328.

[6] Statistical Analysis of Clinical Data on a Packet Calculator part 2, Ton J. Cleophas, Aeilko H. Zwinderman, p7**.**

[7] Diabetes Patients with Experience Project**,** Jason Boyed, Judy Suopway, www.pickereurope.org

[8] Comparative Statistical Analysis of Inpatients with Diabetic Myocardial, Priscilla O.Okunji Phd, Afrooz Afghani PhD. www.IJAHSP.nova.edu

[9] Applied Linear Regression, Third Eddition, Sanford Weisberg, Page 8.

[10] n.wikipedia.org/wiki/Spss.

[11] Bovas Abrahamas, Johannes Ledolter, Introduction to Regression Modeling, p2.

[12] Mosteller F. and J.W. Tukey, Data Analysis and regression, p 588, 1977.

[13] William Mendenhall, Regression Analysis, seventh edition, p188.

[14] Sorin Draghici, Statistical and Data Analysis for Microarrays using R and Bioconductor, second edition, p 219.

[15] http://www.duhokhealth.org/en/node/419

[16] http://www.ncbi.nlm.nih.gov/pubmed/22450348