Arab American University

Faculty of Graduate Studies

Multiple Regression: The Main Causes of Car Accidents in Jenin City As a Case Study

By

Fatima Dirar Hasan Ibrahim

Supervisor

Dr. Saleh Afaneh

Co-Supervisor

Dr. Mahmoud Almanasra

This thesis was submitted in partial fulfillment of the requirements for the degree of master in Applied Mathematics

July/2021

Thesis Approval

Multiple Regression: The Main Causes of Car Accidents in Jenin City As a Case Study

By

Fatima Dirar Hasan Ibrahim

This thesis was defended successfully on 1, July 2021 and approved by

Committee Member	Signature
(Supervisor) Dr. Saleh Sdiq Afaneh	.....
(Co-Supervisor) Dr. Mahmoud Farah Almanasra	.....
(Internal Examiner) Dr. Ammar Abdallah Qarariyah	.....
(External Examiner) Dr. Inad Mohammad Nawajah	.....

Declaration

The work in this thesis, unless otherwise referenced, is the researcher's own work and has not been submitted elsewhere for any other degree or qualification

Student's Name: Fatima Dirar Hasan Ibrahim

Signature: Fa

Date: 7.3.2022

Dedication

To my soulmate, my heartbeat, Aatika

Acknowledgements

In the first place, I present my gratitude to Allah for giving me the strength and patience to finish this thesis.

A lot of thanks for my supervisor Dr. Saleh Afaneh for the huge support He gave me and for his useful advice.I do not forget to thank my co-supervisor Dr. Mahmod Al Manasra for following me up step by step and encouraging me all the time. A lot of thanks for my teachers who light the road in front of me and make it possible to be crossed.

Many thanks for my parents, especially my mother who held my hand through the way, my father who helped me to draw my dream, my brothers who stood beside me during this hard journey.

I am also grateful for my friends who were always by my side.

Abstract
Multiple Regression: The Main Causes of Car Accidents in Jenin City As a Case Study

By

Fatima Dirar Hasan Ibrahim

Regression Analysis is a statistical tool that is used to study the effect of some independent variables over one or more dependent variable. In this thesis, we define simple and multiple linear regression model in order to apply regression on a case concerned with car accidents (which cause body injuries of any type) in Jenin city in 2018. The results appear after analyzing the data using different methods such as standard, hierarchical, stepwise multiple regression can be used to reduce the number of injuries of a car accident which can be useful for the traffic circle.

Contents

List of Tables

List of Figures

Introduction

The importance of math appears in the way it deals with daily life issues, especially that one concerned with human life and health. One of the most important topics in math is regression analysis. Regression can be used to study and analyze some real data in order to explain and extract the factors and the results of an event.

Transportation is an essential part of our daily life. However, traffic accidents have rapidly risen in the recent decades causing harm to human beings and properties. In this research, we will study the factors that may influence the occurrence of a traffic accident in the city of Jenin during 2018. For that purpose, real data from police reports mainly was collected concerned with gender, age, license of the driver and other factors which may have an effect on the number of injuries. The main variables in the data were used to analyze the effect of accidents on the number of injuries using the Statistical Package for the Social Sciences (SPSS).

1. Background and Related Work

According to Gorroochurn (2016), the word regression has first appeared when Galton was studying the laws of heredity. Regression as we know today wasn't the way that Galton first recognized. Firstly, he used the symbol "r" to define the coefficient of reversion which is a genetic process.

He didn't first reach to the concept of regression effect, the reversion that he had noticed (according to him) was undirectional process on offspring. Later on, he found that something other than the undirectional genetic process was affecting. He realized that the reversion was also occurring on parents from their own offspring, after that exactly, Galton decided to change the word reversion into regression so he could prove his hypothesis through the mathematical analysis. Hamilton Dickon found that it is not a genetic reversion but statistical phenomenon that could be either directions.

According to Dhakal (2018), The method of least square (which is the first form of regression) was invented by Legndre and Gauss in

1805.

Fanberg thought that probability is not related to statistic but is a mathematical method. Later on, Gauss laplace-synthesis thought that there is a connection between the normal error theory and the curve fitting method of least square and the analysis of data using linear models. Then the method of least square evolutes to be regression, so the first appearance returns to the German mathematical Karl Gauss.

When Gauss first invented regression he thought his invention was insignificant, he thought it's too simple to be used so he didn't record his invention, after many years Legndre published his invention of his method.

We can finally say that Gauss has suggested the least square method which supports legendres, Legendre was the first who made the invention public, he developed the original then used the least square but Gauss is mostly the owner of this invention and Gauss went far beyond legendre in the technical developments and conceptual and

linked the method to probability also he developed algorithm for estimates.

After Galton, Udny Yule and Karl Pearson developed regression into more statistical concept, according to them the joint distribution of response and explanatory variables is Gaussian. On the other hand, Fisher thought that his assumption is weak, between (1922-1925) he thought that the conditional distribution of the response variables is Gaussian but the joint distribution is not necessarily Gaussian (which is closer to Gauss formulation).

Between(1950-1960) the economist had to use electro mechanical desk (calculators) to calculate regression, before that it sometimes took more than 24 hours to do one task.

The researches concerned with regression are continuous, a new method has been developed for Robust Regression which has a correlated response such as time series and growth curves also deals with different types of predictors where they may be curves or images or graphs or other objects also regression involves missing data

, non-parametric regression, Bayesian methods for regression with predictor variables more than the observations.

According to Ma (2016), traffic cars crash have brought a serious damage to society by paying high expenses to medical health and affecting the productivity, according to Al-jabri (2015) there are many factors that affect the occurrence of an accident that may be related to human behavior, road geometries, traffic weather or more than one reason among and the understanding of those reasons is important to improve safety on roads.

According to Al-Sherhi (2017), non-motorist are more likely to make a car crush on roads, in his study he used the logistic step wise procedure to build a statistical predictor variables.

According to Saha and Basu (2017), in the last years there were a real work to suggest a statistical model to predict car crash without considering the variability in the car mix, this means that they need to improve a statistical model to predict car crash with accuracy as much as possible. According to Munyon (2017), there are

different variables that affect the fatality analysis reporting system, in his study he focused on finding the most affecting variable on the number of fatalities also considering the survivor/fatality status as the response factor using multinomial logistic regression and poisson regression and concluding an alternative modeling strategies comparing them with each other, according to Ahmad (2017) his research aim was discovering the main variables through some variables that cause death from road traffic and the effect of each one of them and in order to determine the effect of the explanatory variables and he used the maximum likelihood.

The aims of this thesis:

1-Study the effect of some chosen variables that are supposed to have an effect over the dependent variable.

2-Find the correlation coefficient "r"between the variables (to measure the strength of the relation between the independent and dependent variables).

3-Study the properties of the regression parameters.

4- Study a case concerned with car accidents and aplying a different methods such as stepwise and hierarchical multiple regression.

5-Build a model that predicts the number of injuries using the different predictors.

2. OutLine of the Thesis

First of all, we have the introduction which is the gate that enters us to the main topic of the research and summarizes the reasons for choosing this topic and how does regression serve us in the real life and the history of regression and the appearance of the word "regression" and some concepts related to regression like the least square method and some famous scientists who work to develop that domain. Also, some recent researches that concerned with regression or the case we study (car accidents) in other countries are discussed.

There are three main chapters in this research

Chapter 1 introduces linear regression, we define regression and simple linear regression model and the importance of regression

models, then we define the parameters of the regression model and their properties then we define the least square method and the maximum likelihood method in order to derive the regression parameters using these methods. Also, the definitions and the concepts that may be used in that research and any other mathematical laws that may be used to analyze the data such like testing the hypothesis, touching some topics concerned with the parameters (the biases) and the collinearity problems.

Chapter 2 introduces the multiple regression in the notation of matrices, defining the model and deriving the equation and the parameters also deriving measrments of precision of the multiple regression parameters, we discuss differnt types of regression such as hierarchical and stepwise regression.

Chapter 3 shows the practical side of this research from where the source of the data that we count on for the analysis results. Also, we define the variables that are included in the data and express the different types of those variables, then we show the outcomes of the

data analysis including the tables and figures and the properties of the different variables like the mean and the variance and the results of the test of significance arranged in tables.

Finally, we summarize the whole work and main conclusions that came out and suggest a future work that may be related to this thesis.

Chapter 1
Regression

1.1 Regression

According to Golberg (2014), Regression analysis is a collection of statistical techniques that lead to conclude an inferences about a relationship among interrelated variables.

It seems that sir Francis Galton was the first who introduced the word "regression". He found that the average heights of children don't tend toward the parents heights but rather toward the average compared to the parents.

According to Neter (1989), a regression model means building a mathematical expression that let us describe and predict the behavior of a variable of interest .For example, (the number of accidents occur in a unit of time, the price of gold, the number of injuries in a car accident) this kind of variables are called the dependent variables denoted by y, those variables may be affected by other variables for example the time, the demand on gold, the season, another impor-

tance for the modeling is describing the change of the mean $E(y)$ according to other variables (the variance doesn't affected by the change in other variables).

Those variables which provide an information that help us describe the behavior of the dependent variable called the predictors or the independent variables denoted by x and supposed to be constants. Also, the model is affected by unknown constants called parameters which controls the behavior of the model, those variables can be extracted from the data.

The type of the model that enters the parameters as a simple coefficients on the independent variables is called a linear model.

Another type is called nonlinear model "nonlinear in parameters", a nonlinear model can be linearized using a suitable transformation on the dependent variable y, those that can't be linearized are called intrinsically linear models.

1.2 Simple Linear Regression Model

According to wackerly (2008), the simple regression is the basic sort where we have one independent variable. Also, the model is linear which can be expressed as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, ..., n \ (1.1)$$

Here, y_i is the dependent variable in the i^{th} trail with two parameters β_0 and β_1. x_i is a known independent variable in the i^{th} trail (ϵ_i) is a random error term with mean

$$E(\epsilon_i) = 0$$

$$var(\epsilon_i) = \sigma^2$$

The regression model is simple because there is one independent variable, we also say that the model is linear in :

1-The Parameters because no parameter appear in an exponential or divided or multiplied by another a parameter.

2-The independent variable because the variable appears in the first

power. In regression, we call the parameters β_0 and β_1 as a regression coefficients. β_1 represents the slope of the regression line, it means the change in the mean of the probability distribution of y per unit increase in x. The parameter β_0 is the y intercept of the regression line (when the value of $x = 0$), β_0 gives the mean probability distribution of y at $x = 0$ (it has a particular meaning only when $x = 0$).

According to Neter (1989), The value of the parameters in the regression model are usually unknown so we estimate them from the data. There are two types of data for the regression analysis, experimental and non-experimental data, following we will discuss both types.

1- Non-experimental or observal data which are obtained without controlling the independent variables. For example, if we are interested in studying the relation between the salary paid for employees (x) and the productivity (y).

The data used for regression analysis are obtained from personnel

record, we call them observal data since they are not controlled. Regression analysis usually based on observal data since it is not possible to conduct controlled experimentation.

2-Experimental Data: we conduct a controlled experiment to provide data from which the regression parameters can be estimated. For example, a company need to study the relation between the period (in hours) that an employee works in a company and the productivity during that period, we call the data an experimental because we can control the work period.

1.2.1 Point Estimation of Mean Response

According to Neter (1989), if we have a sample estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ parameters in the regression function

$$E(y) = \beta_0 + \beta_1 x \ (1.2)$$

Then the regression function can be estimated by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \ (1.3)$$

y is the value of the estimated regression function at x level of in-

dependent variables.

$E(y)$ is the mean response (the mean of the probability distribution of y corresponding to x level of independence.

$\hat{y}$ is a point estimator of the mean response when the level of independent variable is x.

We can define the residual as the difference between the observed value y_i and the corresponding fitted value $\hat{y}_i$ denoted by e_i, such that:

$$e_i = y_i\text{- } \hat{y}_i$$

$$= y_i \text{ - } \hat{\beta}_0 - \hat{\beta}_1 x_i$$

However, the model error term ϵ_i can be expressed as

$$\epsilon_i = y_i - E\left(y_i\right)$$

1.2.2 Basic Assumptions of the Linear Regression Model

According to Kutner (2005), the simple linear regression model can be expressed as

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

where ϵ_i are normally distributed with mean zero and equal variance σ^2. Here are the assumptions of the linear regression model

1- The error terms are normally distributed with mean zero and variance σ^2.

2-the error terms are uncorrelated in the regression model because of normality assumption, that means that the error term in one trail has no effect on the error term for any other trail.

3-The value for the y_i (for the i th trail) exceed or falls short the value of the function of the error term ϵ_i.

4-The error terms are assumed to have the same variance (constant variance σ^2) there for, the y_i terms have the same variance.

$$Var(y_i) = var(\beta_0 + \beta_1 x_i + \epsilon_i) = var(\epsilon_i) = \sigma^2$$

1.3 Methods for Estimating Regression parameters

1.3.1 The Method of Least Square Estimation

Danielsson (2014) mentioned that The least square estimation is a method to approximate the solution of the system of equations. We can apply linear regression in the analysis that include statistics or large sets of data.

Assumptions :

In linear regression models, we have several assumptions concerned with the response and predictor variables and their relationship. In order to get a trustful model, those assumptions should be guaranteed (even though the model could be good without those assumptions), we can classify them as follow

1-Weak exogeneity: the input values could be treated as a fixed values not random.

2-Linearity: This means that the response variables are linear with respect to the coefficients β of the predictor variables.

3- Constant variance: That means that the error in the response variables must have the same variance (regardless the size of the predictor variables).

4-Independence of errors: The errors of the different response variables are uncorrelated.

5- Kutner (2005) mentioned Lack of Multicollinearity: In many cases, with nonexperimental data concerned with economics, business and biological sciences the predictors or the explanatory variables are usually correlated with them selves and with other independent variables that are related to the dependent variable but they are not included in the model. If the predictors are correlated among them selves, then the intercorrelation or (multicollinearity) is said to be exist (if the correlation is very high, we shall reserve the last term).

According to Rawlings (1998), the simple linear model has two parameters β_0 and β_1 which can be estimated from the data. If there were no random error in y_i, we can use any two points to solve

explicitly for the values of the parameters. However, each pair of observed data points give a different results. (All estimates would be identical only if the observed data fell exactly on the straight line).

This method is used to find a solution (the best one) using some criterion. The Least Square estimation procedure uses the criterion where the Least Squares solution must give the smallest possible sum of squared deviations of the criterion observed y_i from the estimates of their true means provided by the solution.

If $\hat{\beta}_0$ and $\hat{\beta}_1$ are numerical estimates of the parameters β_0 and β_1, respectively

Also

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

is the estimated mean of y for each x_i, $i = 1, \ldots, n$.

$\hat{y}_i$ is obtained by substituting the estimates for the parameters in the functional form of the model relating $E(y_i)$ to x_i.

The least squares principle chooses the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squares of the residuals (for any set of sample data) which can be found as followed

$$\sum (y_i - \hat{y}_i)^2 = SSR = \sum e_i^2$$

Where,

$e_i = y_i - \hat{y}_i$ is the observed residual for the i th observation. The estimators for β_0 and β_1 are found by using calculus to find the values that minimize SSR.

First of all, we derive SSR with respect to β_0 and β_1 in turn, then we set the derivative equal to zero. Using $\hat{\beta}_0$ and $\hat{\beta}_1$ to denote the values of β_0 and β_1 respectively (this values minimize the sum square of the residuals)

$$SSR = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

1-Derive with respect to $\hat{\beta}_0$ then set the derivative equal zero we get

$$-2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

2-Derive with respect to $\hat{\beta}_1$ then set the derivative equal zero we get

$$-2\sum x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Divide by -2 for both equations to get

$$\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

By expanding we get

$$\sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0$$

$$\sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$$

The following equations are called the normal equations

$$\sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i$$

$$\sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2$$

Then we can solve the normal equations to get $\hat{\beta}_0$ and $\hat{\beta}_1$ which are the point estimators of β_0 and β_1

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \frac{1}{n}(\sum y_i - \hat{\beta}_1 \sum x_i)$$

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

1.3.2 Bias and the Mean Square of the Point Estimators

According to Wackerly (2008), if we have a parameter called θ and we intend to specify a point estimator for the parameter called $\hat{\theta}$, then we prefer to have a mean or expected value of the distribution of estimate that equals the estimated parameter where $E(\hat{\theta}) = \theta$. If the point estimator satisfies this property, then it is unbiased estimator.

Definition: Let $\hat{\theta}$ be a point estimator of a parameter θ, then $\hat{\theta}$ is unbiased estimator if $\mathrm{E}(\hat{\theta}) = \theta$, if $\mathrm{E}(\hat{\theta}) \neq \theta$ then $\hat{\theta}$ is biased.

Definition: The bias of a point estimator θ is given by

$$B(\hat{\theta}) = E(\hat{\beta}_0) - \theta$$

Theorem: $\hat{\beta}_1$ is an unbiased estimator of β_1.

proof:

$$E(\hat{\beta}_1) = E(\tfrac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2})$$

$$= \frac{\sum(x_i-\bar{x})E(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$

$$=\frac{\sum(x_i-\bar{x})[E(y_i)-E(\Sigma^{y_i}/n)]}{\sum(x_i-\bar{x})^2}$$

$$=\frac{\sum(x_i-\bar{x})\left(\beta_0+\beta_1x_i-\frac{1}{n}\Sigma(\beta_0+\beta_1x_i)\right)}{\sum(x_i-\bar{x})^2}$$

$$=\frac{\sum(x_i-\bar{x})\left(\beta_0+\beta_1x_i-\frac{1}{n}(n\beta_0+\beta_1\Sigma x_i)\right)}{\Sigma(x_i-\bar{x})^2}$$

$$=\frac{\sum(x_i-\bar{x})(\beta_0+\beta_1x_i-\beta_o-\beta_1\bar{x})}{\sum(x_i-\bar{x})^2}$$

$$=\frac{\sum(x_i-\bar{x})\beta_1(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}$$

$$=\frac{\beta_1\Sigma(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}$$

$$=\beta_1$$

Theorem: $\hat{\beta}_0$ is an unbiased estimator of β_0.

proof:

$$E\left(\hat{\beta}_0\right)=E\left(\bar{y}-\hat{\beta}_1\bar{x}\right)$$

$$=E\left(\frac{\sum y_i}{n}\right)-\beta_1\bar{x}$$

$$=\sum\left(E\left(\frac{y_i}{n}\right)\right)-\beta_1\bar{x}$$

$$=\frac{1}{n}(\sum\beta_0+\beta_1x_i)-\beta_1\bar{x}$$

$$=\frac{1}{n}(n\beta_0+\beta_1\sum x_i)-\beta_1\bar{x}$$

$$=\beta_0+\beta_1\bar{x}-\beta_1\bar{x}$$

$$=\beta_0$$

Kutner (2005) mentioned a famous theorem called the Gauss Markov Theorem states

Theorem: Under the condition of regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance among all unbiased linear estimators.

1.3.3 Covariance

Blitzstein (2015) mentioned that the mean and variance provide a single number that describe the distribution of a single random variable. However, the covariance is a single number that summarizes the joint distribution of two random variables. The covariance measures a tendency of two to go up or down together. A positive covariance means that when x goes up y also do, a negative covariance means that when x goes up y goes down.

Definition: the covariance between random variables x and y is

$$cov(x, y) = E[(x - E(x))(y - E(y))]$$

The covariance have those properties:

1. $cov(x, x) = var(x)$

2. $(x, y) = cov(y, x)$

3.$cov(x, c) = 0, cisaconstant$

4.$cov(ax, y) = acov(x, y)$

5.$cov(x + y, z) = cov(x, z) + cov(y, z)$

6.$coc(x + y, z + w) = cov(x, z) + cov(x, w) + cov(y, z) + cov(y, w)$

7.$var(x + y) = var(x) + var(y) + 2cov(x, y)$

1.3.4 Precision of Estimation

According to Rawlings (1989) we can compute random variables from y_i such as $\bar{y}, \hat{y}, e, \hat{\beta}_0, \hat{\beta}_1$.

We call the standard errors and variance measures of precision because using them we can judge the reliability of the estimate.

y_i and the residuals are linear functions of $\hat{y}_i$.

Assume that $U = \Sigma a_i y_i$ (U is arbitrary linear function and a is con-

stant).

To find the variance

$$var\,(U) = \sum a_i^2 var\,(y_i) + \Sigma\Sigma_{i\neq j} a_i a_j cov(y_i, y_j)$$

If the random variables are independent, then the covariance will equal zero (then the double summation will vanish), if the variances of the random variables are equal then

$$var(U) = \sum (a_i^2)\,\sigma^2$$

For a sample of size (n), when y_i has a variance σ^2 and zero covariance then

$$var(\bar{y}) = var(\frac{\sum y_i}{n}\Big)$$

$$= \frac{1}{n^2}\Sigma var\,(y_i)$$

$$= \frac{1}{n^2}\sigma^2 n$$

$$var(\bar{y}) = \frac{\sigma^2}{n}$$

For $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\Sigma x_i y_i}{\sum x_i^2}$$

$$= \left(\frac{x_1}{\Sigma x_i^2}\right) y_1 + \left(\frac{x_2}{\Sigma x_i^2}\right) y_2 + ... + \left(\frac{x_n}{\Sigma x_i^2}\right) y_n$$

Since $\frac{\sum x_i}{\sum x_i^2}$ is constant

$$var(\hat{\beta}_1) = \left(\frac{x_1}{\Sigma x_i^2}\right)^2 \sigma^2 + \left(\frac{x_2}{\Sigma x_i^2}\right)^2 \sigma^2 + ... + \left(\frac{x_n}{\Sigma x_i^2}\right)^2 \sigma^2$$

$$= \frac{\Sigma x_i^2}{\left(\Sigma x_i^2\right)^2} \sigma^2$$

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma x_i^2}$$

To find the variance of $\hat{\beta}_0$

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

$$var(\hat{\beta}_0) = var\left(\bar{y}\right) + \left(-\bar{x}\right)^2 var(\hat{\beta}_1) + 2(-x) cov(\bar{y}, \hat{\beta}_1)$$

Where $\bar{y}$ and $\hat{\beta}_1$ are random variables but we know that

$$var(\bar{y}) = \frac{\sigma^2}{n}$$

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma x_i^2}$$

Now, we need to find the cov($\bar{y}$,$\hat{\beta}_1$)

Suppose that U and W are two functions of coefficients such that

$U = \sum a_i y_i$ and $W = \sum d_i y_i$

Then

$$cov(U, W) = a_i d_i var\left(y_i\right) + \Sigma\Sigma_{i \neq j} a_i d_j cov(y_i, y_j)$$

Overall n(n-1) combinations, also y_i are independent then the summation is zero

So

$$cov(U, W) = \sum a_i d_i \text{var}(y_i)$$

Assume that $U = \bar{y}$ and $W = \hat{\beta}_1$

We get,

$$a_i = \frac{1}{n} \text{ and } d_i = \frac{x_i}{\Sigma x_i^2}$$

Then

$$cov(\bar{y},\hat{\beta}_1) = \Sigma(\frac{1}{n}) \left(\frac{x_i}{\sum x_j^2} \right) var\,(y_i)$$
$$= \left(\frac{1}{n}\right) \left(\frac{\Sigma x_i}{\sum x_j^2} \right) \sigma^2 = 0$$

since $\Sigma x_i = 0$

Then,

$$(\hat{\beta}_0) = var(\bar{y}) + (\bar{x})^2 \text{ var}(\hat{\beta}_1)$$

$$= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum x_i^2}$$

$$= \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2} \right) \sigma^2$$

But

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

$$= \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

$$var(\hat{y}_i) = \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma x_j^2} \right] \sigma^2$$

That variance takes its minimum value when $x_i = \bar{x}$, increases as the value of x_i increase.

$\hat{y}_i$ is used as an estimate of $y_i = \beta_0 + \beta_1 x_i$ when $x_i = \bar{x}$.

Assume that we need to predict a future value y_0 at a specific value $x = x_0$ where we assume $\epsilon_0 \sim N(0, \sigma^2)$, so we use $y_0 = \beta_0 + \beta_1 x_0$ as an estimate of $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

As small as the difference between the predicted value $\hat{y}_0$ and the future value y_0 as success the prediction error.

The average square error of the difference is called Mean Square Error $E(\hat{y}_0 - y_0)^2$. If $E(\hat{y}_0 - y_0)^2 = 0$, then the variance equals the mean square error.

Where

$$var(\hat{y}_{pred}) = var(\hat{y}_0 - y_0)$$

$$=var(\hat{y}_0) + var(y_0) + cov(\hat{y}_0, y_0)$$

$$var(\hat{y}_0) + \sigma^2 = \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2}\right] \sigma^2$$

We can get the estimated variance by replacing σ^2 with s^2 in the previous equation.

Now, we show that $\hat{\beta}_1$ have minimum variance among all unbiased linear estimators as presented in Kutner (2005):

As known

$$\sigma^2(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

$$\text{Let } k_i = \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}$$

$$\text{Then, } \sigma^2(\hat{\beta}_1) = \sigma^2 \sum(k_i)^2$$

$$\text{Let } \hat{\beta}_1 = \sum c_i y_i$$

$$\sigma^2(\hat{\beta}_1) = \sum c_i^2 \sigma^2(y_i) = \sigma^2 \sum c_i^2$$

Define $c_i = k_i + d_i$ (k_i are the least square constant and d_i are arbitrary constants)

$$\sigma^2(\hat{\beta}_1) = \sigma^2 \sum c_i^2 = \sigma^2 \sum (k_i + d_i)^2$$

$$= \sigma^2(\sum k_i^2 + d_i^2 + 2 \sum k_i d_i)$$

But

$$\sigma^2 \sum k_i^2 = \sigma^2(\hat{\beta}_1)$$

$$\sum k_i d_i = 0$$

$$\sum k_i d_i = \sum k_i(c_i - k_i) = \sum c_i k_i - \sum k_i^2$$

$$= \sum c_i [\tfrac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2} - \tfrac{1}{\sum (x_i - \bar{x})^2}]$$

$$= [\tfrac{\sum c_i x_i - \bar{x} \sum c_i}{\sum (x_i - \bar{x})^2} - \tfrac{1}{\sum (x_i - \bar{x})^2}] = 0$$

But

$$\sigma^2(\hat{\beta}_1) = \sigma^2(\hat{\beta}_1) + \sigma^2 \sum d_i^2$$

The smallest value of the variance is met when $\sum d_i^2 = 0$ that means that all the values of $d_i = 0$ so $c_i = k_i$

So the least estimator of $\hat{\beta}_1$ has the minimum variance.

1.3.5 Mean Square Error

According to Wackerly (2008), definition: The mean square error of the point estimator $\hat{\theta}$ is the expectation of the square difference between the estimator and its mean

$$\text{MSE}=E\left[\left(\hat{\theta}-\theta\right)^2\right]$$

Where

$$E\left[\left(\hat{\theta}-\theta\right)^2\right]=E\left(\hat{\theta}^2\right)-2\theta E\left(\hat{\theta}\right)+\theta^2$$

If we have an unbiased estimator then the mean square error equals the variance

proof:

$E\left(\hat{\theta}\right)=\theta$, (because $\hat{\theta}$ is an unbiased estimator of θ)

But

$$\text{the variance} = E\left(\hat{\theta}^2\right)-[E(\hat{\theta})]^2$$

Also

$$\text{MSE}=E\left[\left(\hat{\theta}-\theta\right)^2\right]$$

$$= E\left(\hat{\theta}^2\right) - 2\theta E\left(\hat{\theta}\right) + \theta^2$$

$$= E\left(\hat{\theta}^2\right) - 2\theta\theta + \theta^2$$

$$= E\left(\hat{\theta}^2\right) - \theta^2$$

$$= E\left(\hat{\theta}^2\right) - [E\left(\hat{\theta}\right)]^2$$

$$= \text{variance}$$

Finally, it is better to have an estimator with mean square error tends to zero lim $\mathrm{b}_n\mathrm{n} \rightarrow \infty = 0$

Minimum Variance Unbiased Estimator

If we have two un biased estimators $\hat{\theta_1}$ and $\hat{\theta_2}$ for the same parameter θ, then we would prefer the estimator with the smaller variance.

1.3.6 The Method of Maximum Likelihood

According to Wackerly (2008), we use the maximum likelihood method to find the minimal sufficient statistic that guarantees having an un-biased estimator for the target parameter, this method often leads to the minimum variance unbiased estimator unlike other methods

like the method of moment (not necessarily to have an unbiased estimator).

Theorem: Suppose that the likelihood function depends on K parameters $\theta_1, \theta_2, ..., \theta_k$ choose as estimates those values of parameters that maximize the likelihood L $(y_1, y_2, ..., y_k|\theta_1, \theta_2, ..., \theta_k)$. We sometimes write the function as $L(\theta_1, \theta_2, ..., \theta_k)$ because it is a function of those parameters.

Like the method of least square, we will use the method of maximum likelihood to derive the regression parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$)

Since

$$L(l) = \frac{1}{\sigma\sqrt{2\pi}} exp((-\sum y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2/2\sigma^2)$$

1-Take the ln for both sides since it is a positive increasing function

$$Ln(l(\hat{\beta}_0, \hat{\beta}_1, \sigma^2)) = ln(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2}\Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

2-Take the derivative with respect to the parameter that is need to be estimated ($\hat{\beta}_0$)

$$\frac{\partial Ln(L)}{\partial \hat{\beta}_0} = \frac{-1}{2\sigma^2} 2 \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1)$$

3-Set the derivative equal to zero

$$\frac{\partial Ln(L)}{\partial \hat{\beta}_0} = \frac{-1}{2\sigma^2} 2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0$$

$$\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0$$

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \Sigma x_i}{n}$$

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

Again derive with respect to $\hat{\beta}_1$

$$\frac{\partial Ln(L)}{\partial \hat{\beta}_1} = \frac{-1}{2\sigma^2} 2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i) = 0$$

$$\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i) = 0$$

$$\sum (x_i y_i) - \hat{\beta}_0 \sum (x_i) - \hat{\beta}_1 \sum x_i^2 = 0$$

$$\hat{\beta}_1 \sum x_i^2 = \sum x_i y_i - (\bar{y} - \hat{\beta}_1 - \bar{x}) \sum x_i$$

$$\hat{\beta}_1 \sum x_i^2 = \sum x_i y_i - \bar{y} \sum x_i + \hat{\beta}_1 \bar{x} \sum x_i$$

Then

$$\hat{\beta}_1 (\sum x_i^2 - \bar{x} \sum x_i) = \sum x_i y_i - \bar{y} \sum x_i$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2 - \bar{-x} \sum x_i}$$

Or

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

1.3.7 Confidence Interval for β_0 and β_1

According to Wackerly (2008), a confidence interval is an interval estimator that specify a method to find two numbers that represent the endpoints (called the upper and the lower confidence limits) of that interval where that interval contains our target parameter and it will be narrow. The probability that the confidence interval will contain our target parameter is called confidence coefficient, we can say that with a high "confidence coefficient" we will be highly confident that our target parameter is in the interval.

For a large sample, a 95% confidence is an estimate ± 2 standard error (based on the normal distribution), an estimate of ± 1 is 68% interval.

According to Gelman (2007), the 50% confidence limit is easy to express since we have the same probability to have the true values

inside and outside the interval, with an estimate of 2/3 of standard error. Also, a 95% confidence limit is as three times as 50% confidence limit.

For β_0 the (1- α) confidence limit is

$$\hat{\beta}_0 \pm t(1-\tfrac{\alpha}{2}, n-2)s(\hat{\beta}_0)$$

Where s($\hat{\beta}_0$) is the standared error which can expressed as:

$$s(\hat{\beta}_0) = \sqrt{[\tfrac{\sum(y_i-\hat{y}_i)^2}{n-2}][\tfrac{1}{n}+\tfrac{\bar{x}^2}{S_{xx}}]}$$

For β_1 the (1- α) confidence limit is

$$\hat{\beta}_1 \pm t\left(1-\tfrac{\alpha}{2}, n-2\right)s(\hat{\beta}_1)$$

$$s(\hat{\beta}_1) = \sqrt{\tfrac{\sum(y_i-\hat{y}_i)^2/(n-2)}{S_{xx}}}$$

1.4 Correlation

According to Rawlings (1998), there are two famous techniques for investigating the relationship between two quantitative variables

1- correlation

2- linear regression

Correlation measures the strength of the linear relationship between a pair of variables. However, regression expresses the relationship in the form of an equation.

For example, we could use correlation and regression to investigate whether there is a relationship between smoking and the age of death, and whether the age of death can be predicted for a given data.

According to Bewik (2003), to measure the strength of the relationship, we can calculate the correlation coefficient. In algebraic notation, if we have two variables x and y and the data take the form of n pairs as ($[x_1, y_1]$, $[x_2, y_2]$, $[x_3, y_3]$ ... $[x_n, y_n]$), then the correlation coefficient is given by the following equation

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Here, $\bar{x}$ is the mean of the x values, and $\bar{y}$ is the mean of the y values. This is the product moment correlation coefficient (or Pearson correlation coefficient).

The value of r always lies between –1 and +1. A value of the correlation coefficient close to +1 leads to a strong positive linear relationship that means that one variable increase when the other does. A value close to –1 indicates a strong negative linear relationship (one variable decrease when the other increase or the opposite), a value close to 0 indicates no linear relationship between the variables). However, there could be a nonlinear relationship between the variables.

1.5 Test of Significance

Rawlings (1998) mentioned a famous hypothesis in the simple linear regression when we have a zero value of the linear regression coefficient (the slope significance), that means that the dependent variable changes in other cases. We assume other values for the null hypothesis.

Even if the hypothesis is true, the regression coefficient will never

be equal to the hypothesized value. The importance of the test of significance is to protect against being misled by the random variation in the estimates.

If the difference between the observed value of the parameter $\hat{\beta}_1$ and the hypothesized value of the parameter is greater than zero we can be referred to random variation then the null hypothesis is rejected.

A more general case, when the null hypothesis is written as

$H_0 : \beta_1 = \beta^*$, where β^* is any constant of interest which can be equal to zero. The alternative hypothesis is

1-$H_a : \beta_1 \neq \beta^*$

2-$H_a : \beta_1 > \beta^*$

3-$H_a : \beta_1 < \beta^*$

Depending on the expected behavior of β_1 if the null hypothesis is not true. In the first case, we referred to the two -tailed alternative hypothesis (interests in detecting departures of β_1 from β^* in either directions) and leads to a two-tailed test of significance.

The other two alternative hypothesis are one-tailed alternatives and

lead to one-tailed tests of significance.

If the random errors in the model ϵ_i are normally distributed, then y_i and any linear function of the y_i will be normally distributed. Thus, $\hat{\beta}_1$ is normally distributed with mean β_1 and variance $Var(\hat{\beta}_1)$. If the null hypothesis that $\beta_1 = \beta^*$ is true, then $\beta_1 - \beta^*$ is normally distributed with mean zero.

$$\text{t= } (\beta_1\text{-}\beta^*)/s(\beta_1)$$

is distributed as Student's t distribution with degree of freedom determined by the degree of freedom in the estimate of in the denominator.

1.6 Simple Linear Regression in Matrix Terms

According to Kutner (2005), the simple linear regression model can be stated as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, ..., n$$

For n dependent variables, we get the system:

$$y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_2 + \epsilon_2$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_n + \epsilon_n$$

We can define y and ϵ as vectors

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad x = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Since

$$Y = XB + \epsilon$$

Then

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$= \begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$= \begin{bmatrix} \beta_0 + \beta_1 x_1 + \epsilon_1 \\ \beta_0 + \beta_1 x_2 + \epsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \epsilon_n \end{bmatrix}$$

As we can see

$$E(Y) = XB$$

$$E(\epsilon_i) = 0$$

$$var(\epsilon_i) = \sigma^2$$

Since ϵ_i are independent normal variables then the covariance are zero for all ϵ_i.

$var(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$

$$var(\epsilon) = \begin{bmatrix} \sigma^2 & 0 & 0 & ... & 0 \\ 0 & \sigma^2 & 0 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & ... & \sigma^2 \end{bmatrix}$$

We can express this result as

$var(\epsilon) = \sigma^2 I$, where I is the identity matrix

$$\text{I} = \begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & ... & 1 \end{bmatrix}$$

chapter 2
Multiple Linear Regression

2.1 Multiple Linear Regression Definition

According to Neter (1989), sometimes, a single independent variable (in a simple linear regression model) may not be enough to describe the dependent variable. Also, the number of independent variables affect the dependent variable and provide more precision than a simple regression model which may be imprecise.

An experiment may need more than one independent variable because the response variable usually influenced by several variables.

We can define the linear regression model with x_1 ,. . . ,x_{P-1} independent variables as

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{ip-1} + \epsilon_i \quad (2.1)$$

$\beta_0, \beta_1, ..., \beta_{p-1}$ are parameters

x_{i_1} ,..., x_{ip-1} are unknown constants

ϵ_i are independent $\sim$ N(0, σ^2)

As we showed in the previous chapter, in simple regression we use

algebraic notation. However, according to Rawlings (1998) multiple regression is more complicated so we use the matrices in order to develop the normal equation for the general linear model with any number of independent variables and to derive the measures of precision.

2.1.1 The Multiple Linear Regression Model

Rawlings (1998), assumes that the model contains p independent variables then

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

Where the sample size is (n) with $p+1$ parameters $p' = p+1$ where n>p$'$

To express the model in a matrices notation, first we define the variables as matrices, where

y: n×1 column vector on the dependent variable y_i

x: n×p' matrix

β: $p'\times 1$ vector of parameter estimated

ϵ: n$\times$1 vector of random error

As we know

$$y = x\beta + \epsilon$$

In the form of matrices

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

In the x matrix, each column contains the values for some independent variables. Also, the first column contains only the constant (1) which is the corresponding x variable for β_0.

If we multiply the x matrix with the parameters matrix and add the ϵ matrix we will get the several models with the form

$$y_1 = \beta_0 + \beta_1 x_{11} + \ldots + \beta_p x_{1p} + \epsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \ldots + \beta_p x_{2p} + \epsilon_2$$

.

.

.

$$y_n = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_n$$

Also, the variables in y and ϵ matrices are random variables. However, the x matrix is a matrix of known constants where the vector β is a vector of unknown constants (that we need to estimate).

Again, assume that ϵ_i are independent and identically normally distributed with zero mean and variance σ^2 so the covariance between ϵ_i and ϵ_j where $i \neq j$ is zero since they are independent, so the joint probability density function can be written as

$$\prod_{i=1}^{n} \left[(2\pi)^{-1/2} \sigma^{-1} e^{(-\epsilon_i^2/2\sigma^2)} \right] = \left[(2\pi)^{-n/2} \sigma^{-n} e^{(-\sum \epsilon_i^2/2\sigma^2)} \right]$$

But we assumed that x and β vectors are constants so $x\beta$ is constant. Also, ϵ_i is a random independent variable distributed normally, then y_i is a normal random variable with mean

$$\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}+\epsilon_i$$

The covariance between y_i and y_j is zero since they are independent of each other, then the joint probability density function of y_1,y_2,...,y_n is

$$(2\pi)^{-n/2}\,\sigma^{-n}e^{-\sum(y_i-(\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}))^2/2\sigma^2}$$

When we find the confidence interval and do the hypothesis we assume that the estimates are normal (according to the assumption ϵ_i are normally distributed). On the other hand, for the method of the least square the normality would not be required but it is necessary when we apply the maximum likelihood estimator with respect to each parameter $(\beta_0, \beta_1, ..., \beta_p)$ this method guarantees finding the values that would maximize the probability of obtaining the sample that is called the likelihood function.

2.1.2 The Solution of the Normal Equation

According to Rawlings (1998), in a matrix notation, we write the normal equation as

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

Multiply the left side with x'

$$x'x\hat{\beta} = x'y$$

Multiply with $(x'x)^{-1}$ (assuming that the normal equation is consistent), we get

$$\hat{\beta} = (x'x)^{-1}(x'y)$$

Where $x'x$ is a matrix of the form

$$\begin{bmatrix} n & \sum x_{i1} & \dots & \sum x_{ip} \\ \sum x_{i1} & \sum x_{i1}^2 & \dots & \sum x_{i1} \sum x_{ip} \\ \vdots & \vdots & & \vdots \\ \sum x_{ip} & \sum x_{i1} \sum x_{ip} & \dots & \sum x_{ip}^2 \end{bmatrix}$$

However, $x'y$ can be found as shown:

$$\begin{bmatrix} 1 & 1 & 1 & ... & 1 \\ x_{11} & x_{21} & x_{31} & ... & x_{n1} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & x_{3p} & ... & x_{np} \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_{i1} y_i \\ \vdots \\ \sum x_{ip} y_i \end{pmatrix}$$

It is necessary for the inverse of $x'x$ to be exist in order to solve the equation. In other words, the x matrix has to be full ranked that means no linear dependencies between the independent variables [there are no repetition in the x matrix], we can transform the x matrix into a full column rank by eliminating the reduces in the x matrix.

2.1.3 $\hat{Y}$ and Residuals Vectors

According to Rawlings (1998), we can compute the estimated value $\hat{y}$ of the dependent variable y as

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

To express $\hat{y}$ as a function of y

$$\text{since, } \hat{\beta} = (x'x)^{-1}x'y$$

By substituting

$$\hat{y} = (x\,(x'x)^{-1}x')y$$

$$= py$$

Where p is n×n matrix and symmetric matrix ($p' = p$) also($pp = p$).

We can say that $\hat{y}$ is a linear function of y as in simple regression, multiple regression an error residual e can expressed as

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

$$= y - py = (I - p)y$$

Also, p and $(I - p)$ are symmetric, we can express e as follow :

$$\hat{y} + e = py + (I - p)y = (p + I - p)y = y$$

p-matrix is also known as the hat matrix (denoted with H sometimes) with those properties

1-Symmetric

2-Idempotent ($pp = p$)

3-$tr(p) = tr(x(x'x)^{-1}x') = tr(x'x(x'x)^{-1}) = tr(I_k) = k$

2.1.4 Properties of Linear Function of Random Vectors

According to Rawlings (1998), since $\hat{\beta}$ and $\hat{y}$ and e are functions of the random variable y, then they are random vectors can be expressed as

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

$$\hat{y} = py$$

Since $y = x\beta + \epsilon$ and ϵ is independent random variable with mean zero and variance σ^2 then

$$E(\epsilon) = 0$$

$$var(\epsilon) = I\sigma^2$$

Then

$$E(y) = E(x\beta + \epsilon) = x\beta$$

$$var(y) = var(x\beta + \epsilon) = var(\epsilon) = I\sigma^2$$

This is true if the model is correct, otherwise (if some independent variables are omitted or the model is not correct), then the expectation of y would not be $x\beta$.

Under the assumption that the model is correct, then the joint probability density function is

$$(2\pi)^{\frac{-n}{2}}|I\sigma^2|^{(\frac{-1}{2})}e^{(-1/2)[(y-x\beta)'(I\sigma^2)^{-1}(y-x\beta)]}$$

$$=(2\pi)^{\frac{-n}{2}}\sigma^{-n}e^{(-1/2\sigma^2)(y-x\beta)'(y-x\beta)}$$

To study the properties of $\hat{y}$ and $\hat{\beta}$ and e, we need to study the properties of linear functions of random vectors.

Assume that we have a vector (called z) that contains the random variables $z_1, z_2, ..., z_n$ where

$z=(z_1, z_2, ..., z_n)'$

To find the mean and the variance of z

$$\mathrm{E}(z)=\begin{pmatrix} E(z_1) \\ E(z_2) \\ \vdots \\ E(z_n) \end{pmatrix}=\begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_n \end{pmatrix}=M_z$$

To find the variance, we take a special case when z is (3×1) vector

$$var(z)=\begin{bmatrix} var(z_1) & cov(z_1,z_2) & cov(z_1,z_3) \\ cov(z_2,z_1) & var(z_2) & cov(z_2,z_3) \\ cov(z_3,z_1) & cov(z_3,z_2) & var(z_3) \end{bmatrix}$$

$$=\begin{bmatrix} E[(z_1-M_1)^2] & E[(z_1-M_1)(z_2-M_2)] & E[(z_1-M_1)(z_3-M_3)] \\ E[(z_2-M_2)(z_1-M_1] & E[(z_2-M_2)^2] & E[(z_2-M_2)(z_3-M_3)] \\ E[(z_3-M_3)(z_1-M_1)] & E[(z_3-M_3)(z_2-M_2)] & E[(z_3-M_3)^2] \end{bmatrix}$$

$= E([z - E(z)][z - E(z)]')$

So, z is a vector of order n×1 with mean M_z and variance - covariance matrix v_z.

Let U be a transformation where

$U = AZ$, where A is a k × n matrix and U is K×1 matrix.

$$\text{A}=\begin{pmatrix} a_1' \\ a_2' \\ \vdots \\ a_k' \end{pmatrix}$$

$$\text{U}=\begin{pmatrix} a_1'z \\ a_2'z \\ \vdots \\ a_k'z \end{pmatrix}=\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{pmatrix}$$

$$var(u) = v_u$$
$$= E([u - E(u)][u - E(u)]')$$

So

$$V_u = E([AZ - AM_z][AZ - AM_z]')$$
$$=\text{E (A[Z-M}_z][Z - M_z]'A')$$
$$=\text{A (E[Z-M}_z][Z - M_z]')A'$$
$$\text{A[var(Z)]A}'$$
$$=\text{AV}_z A'$$

To find the expectation of $\hat{\beta}$

$$E(\hat{\beta}) = [(x'x)^{-1}x']E(y) \text{ (since the } x \text{ are constants)}$$
$$E(\hat{\beta}) = [(x'x)^{-1}x']x\beta$$
$$=(x'x)^{-1}(x'x)\beta$$
$$= \beta$$

So, $\hat{\beta}$ is unbiased estimator of β. Again, under the assumption that the model is correct, we can find the variance of $\hat{\beta}$ as

$$var(\hat{\beta}) = [(x'x)^{-1}x'][var(y)][(x'x)^{-1}x']'$$

$(AB)' = B'A'$.Since $x'x$ is symmetric then,

$$\begin{aligned} var(\hat{\beta}) &= (x'x)^{-1}x'x(x'x)^{-1}\sigma^2 \\ &= (x'x)^{-1}\sigma^2 \end{aligned}$$

Also, the variance-covariance of the estimated coefficient is given by $(x'x)^{-1}\sigma^2$. Then, $\hat{\beta}$ is multivariate normally distributed

$$\hat{\beta} \sim N(\beta, (x'x)^{-1}\sigma^2)$$

Again

$$\hat{y} = [x(x'x)^{-1}x']y = py$$

Then

$$px = p$$

Also

$$E(\hat{y}) = pE(y) = px\beta = x\beta, \text{ since } (px = x)$$

If the model is correct, we can prove that $px = x$

$$\begin{aligned} px &= [x(x'x)^{-1}x']x \\ &= x[(x'x)^{-1}(x'x)] \\ &= x \end{aligned}$$

We can derive the variance and covariance of $\hat{y}$

$$\begin{aligned}\hat{y} &= x\hat{\beta} \\ var(\hat{y}) &= var(x\hat{\beta}) \\ &= xvar(\hat{\beta})x' \\ &= x(x'x)^{-1}x'\sigma^2 \\ &= p\sigma^2\end{aligned}$$

or

$$\begin{aligned}\hat{y} &= py \\ var(\hat{y}) &= p[var(y)]p' \\ &= pp'\sigma^2 \\ &= p\sigma^2\end{aligned}$$

As we said, p is symmetric and idempotent, since p is a large matrix, if we need to find a variance of a subset of $\hat{y}_i$, determine the rows of x of the correspond to the data points of interest

$$var(\hat{y}_r) = x_r var[(\hat{\beta})]x_r' = x_r(x'x)^{-1}x_r'\sigma^2$$

Since $x\beta$ is constant (where ϵ and y are normally distributed), then $\hat{y}$ is $N \sim (x\beta, p\sigma^2)$.

2.2 Coefficient of Determination

According to Rawlings (1998), the coefficient of determination R^2 can be found by dividing the total corrected sum of squares of the

dependent variable explained by the independent variables in the model

$$R^2 = \frac{SS(Reg)}{SS(Total)} = \frac{SSR}{SST}$$

The aim is to choose a model that accounts for as much of variation in y as is practical.

As known, R^2 doesn't decrease as we add independent variable to the model. If we add all of the independent variables then we will obtain the maximum value of R^2. Also, R^2 increases as we add more important independent variables, we can use R^2 for a model building. We need a judgment if the increase in R^2 from additional variables justified the measured complexity.

2.3 Analysis of Variation in the Dependent Variable

According to Rawlings (1998), the difference between the observed value and the estimated value [that calculated using the regression

equation] is called the residual. Also, each dependent observation y_i can be written as

$$y_i = \hat{y}_i + e_i$$

The total sum of squares of y_i is

$$\text{SS(Total)}=\sum y_i^2, \text{ for each } y_i$$

$$\sum y_i^2 = \Sigma(\hat{y}_i + e_i)^2$$

$$=\text{SS(Model)}+\text{SS(Residual)}$$

$$\text{Since } \Sigma\hat{y}_i e_i = 0$$

Also

$$\text{SS(Model)}=n\bar{y}_+^2 \hat{\beta}_1^2 (x_i - \bar{x})^2$$

$$\text{SS(Res)}=\text{SS(Total)-SS(Model)}$$

Or

$$\sum y_i^2=\hat{\beta}_1^2 (x_i - \bar{x})^2+\Sigma e_i^2$$

$$=\text{SS(Reg)}+\text{SS(Res)}$$

$\sum y_i^2$ is the corrected total sum of squares. SS(Model) is the sum square for the entire model, SS(Reg) is a partial of SS(Model) that

exceeds the correction factor. The correction factor is the sum squares for the model containing only β_0. SS(Reg) is used to measure the additional information about the independent variables. The degree of freedom for each sum of squares depends on the sample size (n) and the number of parameters (p') where the number of parameters equal the number of the independent variables plus 1.

Table 2.1: Sum Square of Dependent Variables

Source of variation	Degree of freedom	sum of square
SS(Model)	p'	$\sum \hat{y}_i^2$
Total correction factor	$n-1$	$\sum y_i^2$
correction factor	1	$n\bar{y}^2$
Due to Regression	$p'-1$	$\Sigma\hat{y}_i^2 - n\bar{y}^2$
Residual	$n-p'$	Σe_i^2

The mean square error can be found by dividing each sum of squares over its degree of freedom.

Table 2.2: Anova Table for Multiple Regression

Source	SS	df	MS	F
Regression	SSR	P	$MSR = \frac{SSR}{P}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$	
Total	SST	$n-1$		

Another measure can be calculated called the coefficient of determination (R^2).

Where

$$R^2 = \frac{SSR}{SST}, O \leq R^2 \leq 1$$

The coefficient of determination is the square of the product moment between y_i and $\hat{y}_i$ (when there is one independent variable) and it is the square of correlation coefficient between x_i and y_i.

For a correct model, the residual mean square is an unbiased estimator of σ^2, the regression mean square is unbiased estimator of $\sigma^2 + \beta_1^2 \Sigma\,(x_i^2)$

Where

$$\Sigma\,(x_i^2) = \Sigma\Big(X_i - \bar{X}\Big)^2$$

These expectations are related to the mean square denoted by $E[MS(Regr)]$ and $E[MS(Res)]$.

2.4 Regression Through the Origin

According to Rawlings (1998), if the regression line passes through the origin (the values of the independent variables are zero), then the mean of the dependent variables is expected to be zero .

Since the linear regression model is passing through the origin, then we can set $\beta_0 = 0$, so the equation is

$$y_i = \beta_1 x_i + \epsilon_i$$

Then we shall estimate only β_1 (using the least square estimation).

$$\hat{\beta}_1 \Sigma \left(x_i^2\right) = \Sigma x_i y_i$$

The estimated value is

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

The regression equation is

$$\hat{y}_i = \hat{\beta}_1 x_i$$

As we know

$$e_i = y_i - \hat{y}_i$$

Also, the uncorrected sum of squares of y partitioned into

$$\text{SS(Model)}=\sum \hat{y}_i^2$$

$$\text{SS(Res)}=\sum (y_i - \hat{y}_i)^2=\Sigma e_i^2$$

SS(Model) has only 1 degree of freedom since the model has one parameter.

2.5 Error Variance Estimation

According to Alharbi (2013), error variance estimation is one of the most important issues in regression models. The estimation of the error variance is essential to assess the variability of the estimated mean of y_i given x_i

Therefore, the error variance plays an important role in regression analysis. For example, it is important to know σ^2 to draw inferences about the mean of y_i and about the regression coefficients, to assess the goodness of fit for the estimated mean function, to obtain a 95% confidence interval for $m(x_i)$ and to predict a new y for a given

x. To sum up, almost in every inferential aspects the knowledge of the error variance is essential. In parametric regression models, the error variance can be a constant or a function of the independent variables.

2.7 Model Selection Methods

According to Tranmer (2008), in some cases, for example when there are a large number of explanatory variables, we can use a statistical method to include or exclude explanatory variables, to make sure that we will have the best equation to predict the dependent variable.

This is a different fundamental approach to the substantive approach where the variables are included on the basis of the research question and this variables are often chosen given the results previous research on the topic and are also influenced by 'common sense' and data availability .

Two types of selection methods are :

1-Backward elimination

2-Farward stepwise selection

3- Stepwise

The main disadvantage of these methods is that we might lose some important theoretical variables, or interactions.

Here is a short description of the two selection methods :

- Backward elimination

This method begins with a model that includes all the explanatory variables. Then, remove that one that is least significant and reform the model after having the least significant explanatory removed, again we remove the least significant explanatory variable from the remaining set then perform the model, and so on, until some 'stopping' criterion is satisfied which is usually that all the explanatory variables that are included in the model are significant.

-Farward stepwise selection

According to Neter, this method chooses the subset models by adding

one variable at a time to the previously chosen subset. This method starts by choosing as the one-variable subset the dependent variable that accounts for the largest co of variation in the dependent variable, this variable will be that has the simple correlation with y. At each step, the variable in the subset of variables that is not already in the model that causes the largest decrease in the residual sum of squares is added to the subset, the method continuous until all variables are in the model.

- Stepwise

More or less the reverse of backward elimination, in that method we start with no explanatory variables in the model, and then build the model up step-by-step. We begin by including the variable that is most highly correlated to the dependent variable in the model. Then include the next most correlated variable, allowing for the first explanatory variable in the model and keep adding explanatory variables until no further variables are significant. In this approach, we may delete a variable that has been included at an earlier step

but is no longer significant, given the explanatory variables that were added later. If we avoid this possibility and do not allow any variables that have already been added to the model to be deleted, this model building procedure is called "forward selection".

Extra Sums of Squares:

The main idea of the extra sum of squares is to measure the marginal reduction in the error sum of squares when we add variable(s) to the model given that the other variables are already in the model. An extra sum of squares can be useful when we intend to test whether the term $\beta_k x_k$ can be dropped from the model, that means testing

$H_0 : \beta_k = 0$

$H_a : \beta_k \neq 0$

2.7.1 Multilevel Modeling

According to Gelman (2006), a Multilevel Modeling is a generalized linear model where the regression coefficients are given by a probability model.

The model has parameters called "hyper parameters" which are estimated from the data. Multilevel model is also called Hierarchical Model because of the structure of the data and the model it self.

Multilevel model is also called Random Effect or Mixed-Effect Model, the reason of that name is the regression efficient that included in the model are supposed to have random outcomes from the process that predict it. On the other hand, a fixed effect model is in common with parameters that do not vary or the parameters that are vary but are not modeled.

It is good to know that a mixed effect model includes fixed and random effects.

-Why do we use multilevel modeling ?

We use multilevel modeling for many reasons

1-learning about the different treatment effects :

The main goal of regression is to study how does a parameter y changes when a special x parameter changes assuming that the other parameters are constants, In some cases, we are not interested in all

of the independent variables x but how does this effect varies in the population.

We usually use interaction to study this case. On the other hand, Multilevel allows us to study the effect that vary by groups.

2-Using all data to perform inference for groups with small sample size: When we have a small sample size, it is not useful to use the classical regression to estimate a group level quality effect.

3-Prediction: A Multilevel Model guarantees many conditions such as finding the group indicators so we can predict a new group from the group level effect predicted (it is essential to have a large sample size).

Being more specified in prediction and estimation (having the right standard deviation error).

In classical regression, a dependent variable may be modeled by including indicator variables for each treatments but some new data can not be predicted in a different treatments because there would not be an indicator for this new data. This problem can be solved

using a multilevel regression. indicator variables and fixed or random effect

According to Gelman, if we include an input variables with j categories into a classical regression, in standard regression we choose one of the categories as a baseline and include the j-1 categories as indicators. However, in multilevel model it is necessary to do this step of choosing one of the levels as a base line. For example, in classical regression the indicators could not be all included because they could be collinear with the constant term but in multilevel model it is not a problem because the indicators are modeled by group-level distribution.

Fixed and random effects

The coefficients (a'_Js and B'_Js) in a multilevel model are called random effect term because of the randomness of the model for the group-level coefficients.

Fixed effect are defined as varying coefficients that are not modeled, in classical regression including j-1 indicators as predictors is

called fixed effect model, that means that coefficients do not vary by groups.

Chapter 3
Data Analysis and Description

According to the Palestinian traffic police records, there has been an increase in the percent of the car accidents, there were (11541) car accidents in the different north Governorate of Palestine in 2017. On the other hand, there has been (12829) car accidents in 2018 which is 11% percent higher than the previous year.

In Jenin city, there has been 1417 car accidents recorded in 2017, a higher number was recorded in 2018 (1470) which is 3.7% percent increase.

In this chapter we will descripe the data icluded in our case study, classify them into different types, talk about the sample size then we will do a data analysis using different types of regression such as Stadard and Hierarchical and stepwise multiple regression then explaining the results that came out.

3.1 Methodology

1- The sample: The data used in this thesis has been collected from the treffic police archive in the city of jenin after emailing brigade Hazem Attalah and having the permission to have this data. The sample used in the data analysis contains car accidents recorded during 2018 (only that caused body injuries or death), there has been 524 car accidents during that period that satisfied those conditions (of course there has been some missing data concerned with some accidents or some unrecorded car accidents).

2- Main variables

In order to study the occurrence of the car accidents in Jenin city during (2018), some variables have been chosen that may be related to the car accidents (considering that those variables can be noticed and measured) and has been registered so we can have them from their source. On the other hand, we ignore the variables that doesn't seem to be related to the dependent variables.

There are three types of variables included in this research

1- Independent variables

1-Gender: male or female.

2-Age: the age is an integer with positive value (greater than zero).

3-Day: include the seven days of the week.

4-Time of accident: the time has been divided into two periods (day and night).

5-Season: including the four seasons of the year (summer, autumn, winter, spring).

6-Age of the car: measured with years (integer number).

7-The license: there are three choices for that variable

a-The driver has a valid license.

b-The driver has no valid license.

c-The driver has no license.

8-The insurance: there are three choices for this variable.

a-The car is registered in an insurance company and the insurance is valid.

b-The car is registered in an insurance company but the insurance is not valid.

c-The car is not registered in an insurance company.

9-The car type: depends on the producer company (Hyundai , kaya, Bmw,...).

10-The car size (small or big).

11-The location: there are three different regions that may affect the occurrence of an accident depends on the feature of that region (city, camp, village).

2-Dependent variables

The number of injuries and number of deaths are considered as a dependent variables that may be zero or any integer number.

3-Some unobserved variables

Some variables may be related to the dependent variables but the problem here that sometimes we can't measure those variables since they can't be noticed or they are not registered, here are some of those variables

1-How many years has the driver been driving?

2-The skills that the driver has.

3-Is the road that the accident took as a place narrow or wide? broken or paved?

4-Is there any traffic signs over the road (if there how many?).

5-Is the driver drunk or stoned?

6-Is there any broken or damaged parts in the car?

7-The number of passengers in the car (that may make some confusion).

8-The car speed just before the accident happens.

9-Is there any traffic jam or large number of vichles?

Although there are some variables that are unobserved or immeasurable, but that factors have a minor effect that don't influence our final results and conclusions.

3.2 Main Results

According to the data collected from the recorded car accidants in the archives of traffic police in the city of jenin, the data included 525 car accidants which caused body injuries for one or more passenger including the driver (those body injuries could be simple such as a fraction or a cut or even a bruise or dangerous injuries that lead to death), other accidants that occures in 2018 and did not cause any body injuries (for example those accidents that caused damage for the properties) are not included in the sample. The data analysis is done using Spss (version 26) and the collected data, the outcomes were as shown below.

1-Variables that concerned with the driver

1-1 The gender To deal with the missing data (that represents a small percent) and since the data is normaly distributed (as we will show) then we will use replace the missing values with the mean.

Table 3.1: Car Accidents According to Gender

	Frequency	Percent	Valid Percent	Cumulative Percent
Missing Data	1	0.2	0.2	0.2
male	496	94.5	94.5	94.7
female	28	5.3	5.3	100.0
Total	525	100.0	100.0	

Table 3.1 shows the number and percent of the drivers who recorded at least one accident divided into two groups according to gender, the percent of the females where 5.3%. However, the males percent has reached 94.5%.

1-2 Age

Table 3.2: Car Accidents According to the Drivers Age Categories

	Frequency	Percent	Valid Percent	Cumulative percent
Missing Data	4	0.8	0.8	0.8
less than 20	24	4.6	4.6	4.6
20-29	203	38.7	39.0	43.6
30-39	136	25.9	26.1	69.7
40-49	98	18.7	18.8	88.5
50-59	42	8.0	8.1	96.5
more than 60	18	3.4	3.5	100.0
Total	525	100.0		

Table 3.2 represents the number and percent of each age category who cause a car accident, we can see that the highest percent belongs to those who are aged (20-29) where the percent has reached 38.7%

(almost one third), the second order was of the share of the age (30-39) which are the closest category to the first one.

This results can be showed in Figure 1.

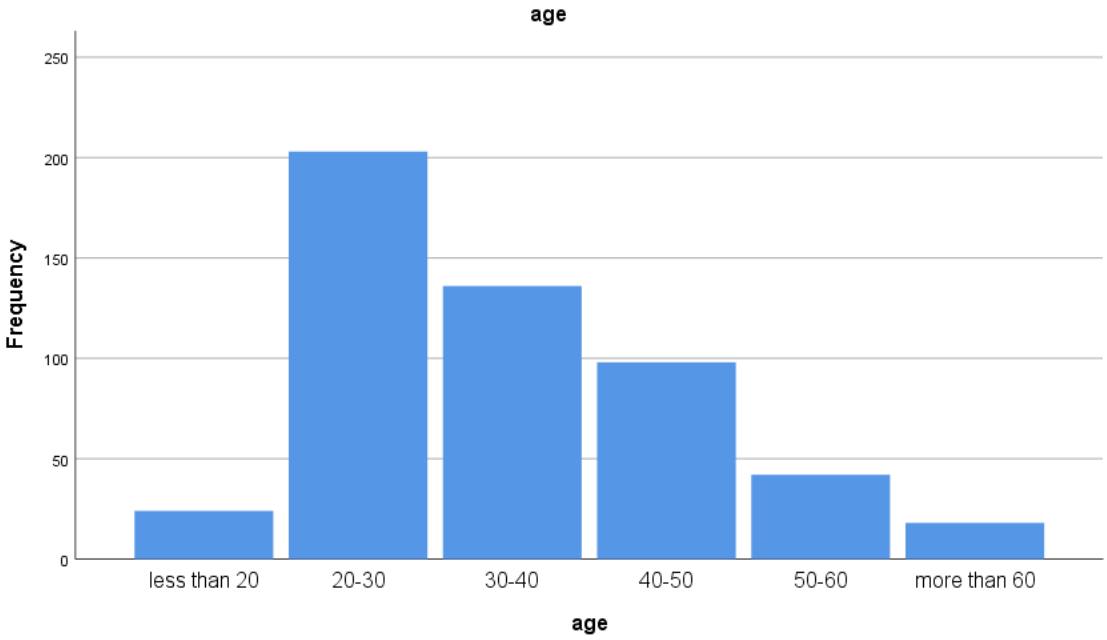

Figure 3.1: Driver Age

1-3 The insurance

Table 3.3: Car Accidents According to the Insurance

	Frequency	Percent	Valid Percent	Cumulative Percent
Missing Data	20	3.8	3.8	3.8
Valid Insurance	481	91.6	95.2	95.2
Un Valid Insurance	15	2.9	3.0	98.2
No Insurance	9	1.7	1.8	100.0
Total	525	100.0		

Table 3.3 shows a real difference between the percent of the

drivers who are registered in an insurance company and has a valid insurance and those who are not registered in an insurance company or their insurance is not valid.

1-4 The license

Table 3.4: Car Accidents According to the License

	Frequency	Percent	Valid Percent	Cumulative Percent
Missing Data	19	3.6	3.6	3.6
Un Valid License	30	5.7	5.7	9.3
Valid License	475	90.5	90.5	99.8
No licence	1	0.2	0.2	100.0
Total	525	100.0	100.0	

Table 3.4 shows a difference between the percent of the drivers who has a valid license 90.5% and those who has no license 0.2% or no valid license 5.7 %.

2-Variables concerned with the car

2-1 Car size

Table 3.5: Car Accidents According to the Car Size

	Frequency	Percent	Valid Percent	Cumulative Percent
Missing Data	22	4.2	4.2	4.2
Big Car	65	12.4	12.4	16.6
Small Car	438	83.4	83.4	100.0
Total	525	100.0	100.0	

Table 3.5 shows that the percent of the cars which are of a small

size were 83.4% which is much greater than the percent of cars of big size which is 12.4% .

2-2 Car type

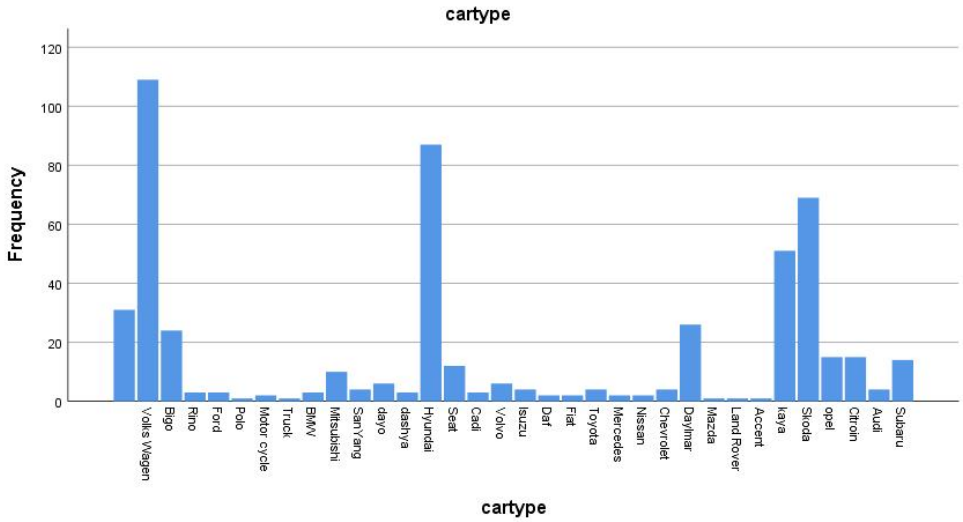

Figure 3.2: Car type distribution

Figure 2 shows the distribution of different types of cars that has been registered, we can see that the highest percent is Volks Wagen followed by Hyundai, that may be because the community of drivers prefer to catch those types of cars fo their properties or availability or cheapness.

2-3 Car age

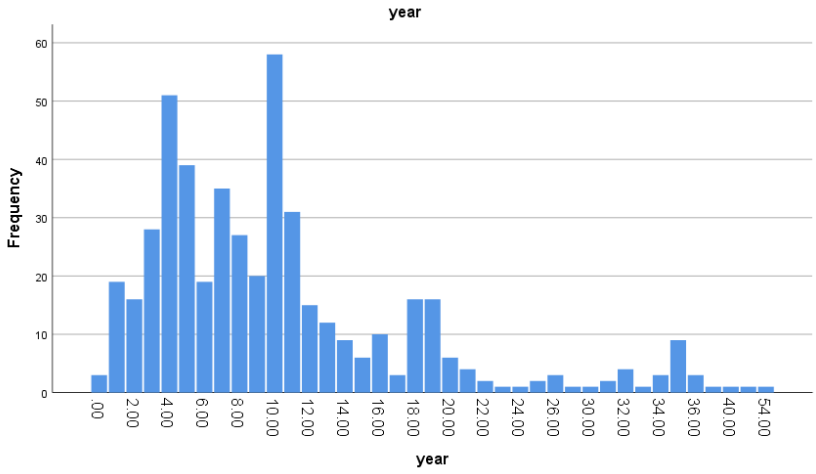

Figure 3.3: Car age distribution

Figure 3 shows the distribution of the different car ages, we can see that the majority of the cars are less than 12 years old, where few cars are old (more than 12 years old), this may be because drivers prefer to buy cars with new models and change those cars when they are old. Also, those cars are not designed for long time use, the majority of the car drivers with short car age are young who are mostly the main category that cause accidents.

3- External variables

3-1 Time of accident

Table 3.6: Car Accidents According to the Time of Accident

	Frequency	Percent	Valid Percent	Cumulative Percent
Missing Data	5	1	1	1
Day	318	60.6	60.6	61.5
Night	202	38.5	38.5	100.0
Total	525	100.0	100.0	

Table 3.6 shows that percent of the accidents during the day is as twice as the accidents during the night.

3-2 The season

Table 3.7: Car Accidents According to the Season

	Frequency	Percent	Valid Percent	Cumulative Percent
Missing Data	6	1.1	1.1	1.1
Summer	156	29.7	29.7	30.9
Autumn	132	25.1	25.1	56
Winter	97	18.5	18.5	74.5
Spring	134	25.5	25.5	100
Total	525	100	100	

According to Table 3.7, in summer the percent is 29.7% , in autumn 25.1%, in winter 18.5%, in spring the percent is 25.5%, that may be refered to the crowded roads in summer since the weather is stable and summer is the vacation season.

3-3 The day

Table 3.8: Car Accidents According to the Day

	Frequency	Percent	Valid Percent	Cumulative Percent
Missing Data	10	1.9	1.9	1.9
Saturday	92	17.5	17.5	19.4
Sunday	83	15.8	15.8	35.2
Monday	61	11.6	11.6	46.9
Tuseday	70	13.3	13.3	60.2
Wednesday	70	13.3	13.3	73.5
Thursday	81	15.4	15.4	89
Friday	58	11	11	100
Total	525	100	100	

Table 3.8 shows that the highest percent of car accidents were on Saturday which is 11.7% different than Sunday which comes at the second order.

3-4 The location

Table 3.9: Car Accidents According to the Location

	Frequency	Percent	Valid Percent	Cumulative Percent
Missing Data	5	1	1	1
city	171	32.6	32.6	33.5
camp	5	1	1	34.5
village	344	65.5	65.5	100
Total	525	100	100	

Table 3.9 shows that the highest percent according to the location returns to the area of the village with 65.5% percent.

3.2.1 Measurements of Precision of the Quantitative Variables

Table 3.10: Measurements of Precision of the Quantitative Variables

	n	minimum value	maximum value	mean	st.deviation
driver age	472	17	77	33.8157	11.67782
car age	472	1	54	10.3665	8.18024
number of injuries	472	1	10	2.563	2.029

Table 3.10 shows the measures of precision for the quantitative variables (driver age, car age, number of injuries), we can see the minimum and maximum value of each variable in addition to the mean and standard deviation. As shown, the minimum age for the a driver who caused a car accident was 17 where the maximum age recorded was 77, the mean value were 33.815, the maximum value of the age of cars recorded was 54. However, the minimum value were 1 year, the mean value for the age of cars was 8.18, the maximum value for the number of injuries which caused by an accident was 10 where the minimum value were 1, the mean value for the number of injuries in one accident was 2.563.

2- **Testing the assumptions of linear regression**

1- Normality and homoscedasticity

Normality means that the relationship between two variables is a straight line also that the variability of the source of the variables is roughly the same for all variables and the residuals of the variables are independent.

In the normal probability plot of regression standarized residuals, the normality of the variables is assessed when no points are extremely deviate from the straight line (diagonal line), then it is assumed that there are no deviation from normality. The properties of normality and homoscedasticity are checked by assessing the residuals scatter plot, we check the overall shape of the scatter plot. If the scatter has almost a rectangular shape and most scores are concentrated around the center, if the variables are normally distributed then homoscedasticity could be assumed. 2- Multicollinearity

If the multicollinearity property is present then the regression coefficient might become insignificant, that refers to the large size of standard errors. The squared multiple correlation of variable (SMC) serves as a dependent variable with the other variables as

independent variables. A high measure of (SMC) means a high correlation between independent variables and therefor results in a multicollinearity.

The (SMC) is computed to the tolerance for multicollinearity. The variance inflation factor (VIF) is the inverse of the tolerance and it is also a measure of multicollinearity. If the tolerence falls below 0.1 or the (VIF) exceeds 10 then is a concern of multicollinearity.

1- Normality

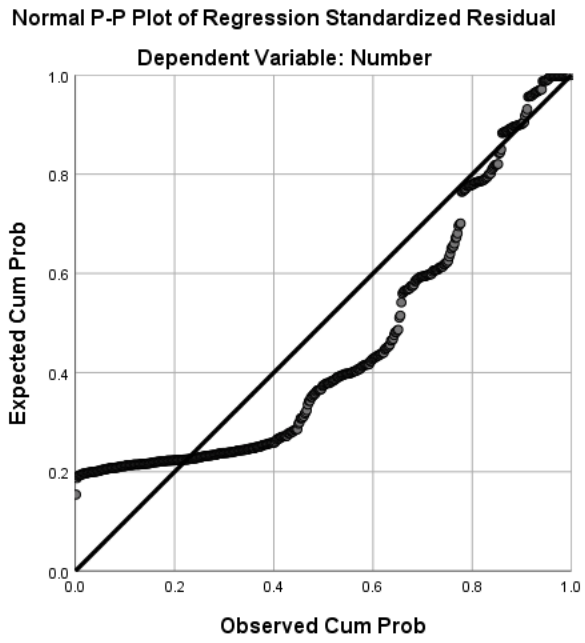

Figure 3.4: Normal P-P plot of Regression Standarized Residual

The P-P plot figure (where the number of injuries had been set as a dependent variable and the driver age and car age are the independent variables) shows that the normality assumption is met with a little deviation.

2- Homoscedasticity

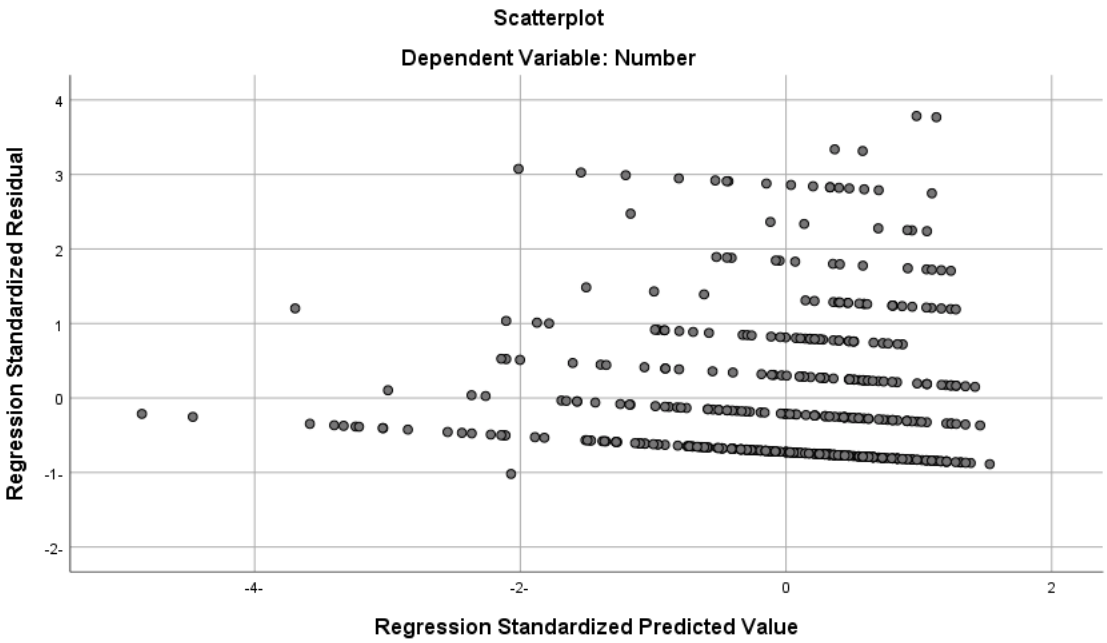

Figure 3.5: Scatter Plot

Since the points are almost equally distributed above and below zero on the X axis and to the left and right to the Y axis then the property of homoscedasticity is satisfied.

3- Multicollinearity

Table 3.11: Coefficients

	β	Std. Error	Beta	t	p-value	Tolerance	VIF
(Constant)	2.907	0.296	9.835	0.000			
driver age	-0.008	0.008	-0.045	-0.992	0.322	1.000	1.000
car age	-0.023	0.011	-0.094	-2.068	0.039	1.000	1.000

We can check the multicollinearity using the VIF value, since the values are below 10 for both independent variables then the property is met.

3.2.2 Standard Multiple Regression

The variables included in the model are the car age and the driver age as independent variables and the number of injuries as a dependent variable.

According table 12 we can built a model to predict the number of injuries where

Number of injuries= 2.907 - 0.008 Driver age - 0.023 Car age

Table 3.12 : Anova Table

	Sum of Squares	df	Mean Square	F	Sig
Regression	20.229	2	10.115	2.658	0.071 [b]
Residual	1803.582	474	3.805		
Total	1823.811	476			

The probability of the F statistic for the overall regression relationship is (0.071) which is more than 0.05 (the level of significance). We fail to reject the null hypothesis that there in no relationship between the dependent variable and the set of independent variables.

Table 3.13: Model Summary

R	R Square	Adjusted R Square	R Square Change	F Change	df1	df2	Sig. F Change
0.105	0.011	0.007	0.011	2.658	2	474	0.071

The multiple R for the relationship between the set of independent variables and the dependent variable is 0.105 which would be characterized as a weak relationship.

3.2.3 Test of Significance

The alpha value (significance level) the probability of rejecting the null hypothesis when it is true. In the social sciences, the alpha

value is typically set at 0.05, this rpresents the amount of the acceptable error or the probability of rejecting the null hypothesis.

The null hypothesis assumes that there is no effect for the independent variable on the dependent variable. On the other hand, the alternative hypothesis suppose that there is an effect for the independent on the dependent variable. Also, the smaller the α-value the stronger we should reject the null hypothesis.

In our case, the α-value is 0.05 that means if the calculated value of the α-value is less than 0.05 we reject the null hypothesis. If the α-value is greater than 0.05, then we fail to reject the null hypothesis.

Table 3.14: Test of Significance

Source	Wald Chi-square	df	Sig.
(Intercept)	13.848	1	0.000
gender	7.343	1	0.007
time	4.719	2	0.094
location	4.695	3	0.196
day	17.722	7	0.013
insurance	1.502	3	0.472
car size	1.428	2	0.490
month	74.975	12	0.000
driver age	5.538	1	0.019
car age	5.124	1	0.024

According to table 3.14, using the standard multiple linear regression analysis, for the (gender, day, month, car age, driver age) we reject the null hypothesis (that there is no effect of those independent variables over the number of injuries).

For the variables (location, insurance, car size), we fail to reject the null hypothesis (we fail to reject that there is no effect of those independent variables on the number of injuries).

3.2.4 Hierarchical Multiple Regression

Table 3.15: Model Summary

R	R Square	Adjusted R Square	R Square Change	F Change	df1	df2	Sig. F Change
0.095	0.009	0.007	0.009	4.332	1	475	0.038
0.105	0.011	0.007	0.002	0.984	1	474	0.322

The R square change statistic for the increase in R^2 associated with the added variables (car age and the number of injuries) is 0.011 using a proportional reduction in error interpretation of R^2 information provided by the added variables reduces our error in predicting number of injuries by 1.1%.

The probability of the F statistic (0.984) for the change in R^2 associated with the addition of the predictor variables to the regression analysis containing the control variable is 0.322 which is more than 0.05, so we cant reject the null hypothesis that there is no improvement in the relationship between the dependent variable and the set of independent variables when the predictors are added.

Table 3.16: Model Summary

R	R Square	Adjusted R Square	R Square Change	F Change	df1	df2	Sig. F Change
0.047	0.002	0.000	0.002	1.034	1	475	0.310
0.105	0.011	0.007	0.009	4.275	1	474	0.039

The R square change statistic for the increase in R^2 associated with the added variables (driver age and the number of injuries) is 0.011

using a proportional reduction in error interpretation of R^2 information provided by the added variables reduces our error in predicting number of injuries by 1.1%.

The probability of the F statistic (4.275) for the change in R^2 associated with the addition of the predictor variables to the regression analysis containing the control variable is 0.039 which is less than 0.05, so we reject the null hypothesis that there is no improvement in the relationship between the dependent variable and the set of independent variables when the predictors are added.

We support the research hypothesis that there is a statistically significant improvement in the relationship between the dependent variable and the set of independent variables.

3.2.5 Stepwise Regression

The best subset of predictors for the number of injuries included the independent variables the (month ,time, car age, gender).

This table illustrates the stepwise method: SPSS starts with zero

predictors and then adds the strongest predictor, month, to the model if it is b-coefficient in statistically significant ($p \leq 0.05$) It then adds the second strongest predictor (time). Because doing so may render previously entered predictors not significant, SPSS may remove some of them -which doesn't happen in this our case. This process continues until none of the excluded predictors contributes significantly to the included predictors. In our example, 4 out of 9 predictors are entered and none of those are removed.

Table 3.17: Anova Table

model		Sum of Squares	df	Mean Square	F	p-value
1	Regression	69.795	1	69.795	18.869	0.000^b
	Residual	1727.420	467	3.699		
	Total	1797.215	468			
2	Regression	106.551	2	53.275	14.684	0.000^c
	Residual	1690.665	466	3.628		
	Total	1797.215	468			
3	Regression	123.540	3	41.180	11.441	0.000^d
	Residual	1673.675	465	3.599		
	Total	1797.215	468			
4	Regression	142.767	4	35.692	10.010	0.000^e
	Residual	1654.448	464	3.566		
	Total	1797.215	468			

The probability of the F statistic for the regression relationship which includes the predictors is less than the level of significance

(less than 0.05). We reject the null hypothesis that there is no relationship between the best subset of independent variables and the dependent variable. We support the research hypothesis that there is a statistically relationship between the best subset of independent variables and the dependent variable.

Table 3.18: Model Summary

R	R Square	Adjusted R Square	R Square Change	F Change	df1	df2	Sig. F Change
0.197	0.039	0.037	0.039	18.869	1	467	0.000
0.243	0.059	0.055	0.020	10.131	1	466	0.002
0.262	0.069	0.063	0.009	4.720	1	465	0.030
0.282	0.079	0.072	0.011	5.392	1	464	0.021

We add the predictors as the following order (month, time, car age, gender)

The multiple R for the relationship between the subset of independent variables that best predict the dependent variable is 0.282 which would be characterized as a weak relationship.

SPSS built a model in 4 steps, each of which adds a predictor to the equation. While more predictors are added, adjusted r-square levels of adding a second predictor to the first raises it with 0.018, but

adding a fourth predictor to the previous 5 only results is a 0.002 point increase. There's no point in adding more than 6 predictors. Our final adjusted r-square is 0.011 which means that our 4 predictors account for 1.1% of the variance in overall satisfaction. This is somewhat disappointing.

3.3 Analysis of the Dummy variables

The dummy variables included in the analysis where the number of injuries is considered as a dependent variable. saturday, sunday, tuesday, monday, wednesday, thurseday friday, car size, city, camp, village, no insurance, valid insurance, un valid insurance,male, female, winter, spring, autumn, valid license, un valid license.

Table 3.19: Model Summary

R	R^2	Adjusted R^2	R^2 Change	F Change	df1	df2	Sig. F Change
0.321	0.103	0.065	0.103	2.751	20	480	0.000

The multiple R for the relationship between the subset of independent variables(the dummy variables included in the model) and the

number of injuries is 0.321 which would be characterized as a weak relationship.

Table 3.20: Anova Table

	Sum of Squares	df	Mean Square	F	p-value
Regression	197.858	20	9.893	2.751	0.000 [b]
Residual	11725.883	480	3.596		
Total	1923.741	500			

The probability of the F statistic for the overall regression relationship is (0.000) which is less than .05 (the level of significance). We reject the null hypothesis that there in no relationship between the dependent variable and the set of independent variables, so the model is statistically significant

Table 3.21: Coefficients (Dummy variables)

	β	Std. Error	Beta	t	p-value
(Constant)	4.825	1.395	3.458	0.001	
male	2.459	1.354	.283	1.815	0.070
female	1.332	1.400	0.148	0.951	0.342
saturday	-0.318	0.410	-0.060	-0.776	0.438
sunday	-0.613	0.408	-0.115	-1.504	0.133
monday	-0.337	0.446	-0.053	-0.756	0.450
tuesday	-0.529	0.423	-0.093	-1.250	0.212
wednesday	-0.975	0.420	-0.168	-2.321	0.021
thursday	-0.462	0.414	-0.083	-1.115	0.266
friday	-0.030	0.440	-0.005	-0.068	0.946
un valid insurance	-0.234	0.535	-0.020	-0.436	0.663
valid insurance	-0.275	0.414	-0.032	-0.665	0.506
no insurance	-0.903	0.700	-0.061	-1.290	0.198
autumn	0.915	0.234	0.202	3.913	0.000
winter	-0.583	0.251	-0.116	-2.317	0.021
spring	-0.023	0.227	-0.005	-0.102	0.919
valid license	-0.225	0.352	-0.029	-0.640	0.523
un valid license	-1.659	0.937	-.198	-1.770	0.077
city	-0.028	0.649	-0.007	-0.043	0.966
camp	0.015	1.292	0.001	0.011	0.991
village	-0.194	0.640	-0.047	-0.303	0.762
small car	-1.898	0.790	-0.336	-2.401	0.017
big car	-2.060	0.817	-0.351	-2.522	0.012

We can see that the (wednesday, autumn, winter, small car, big car) variables are statistically singificant .The other variables are insignificant.

Table 3.22[a]: Coefficients

	B	Std. Error	Beta	t	p-value
age	-0.011	0.007	-0.064	-1.439	0.151
city	0.156	0.184	0.037	0.851	0.395
camp	0.178	0.863	0.009	0.206	0.837

Table 3.22[a] shows that the dummy variables included are insignificant. Also, the number of injuries in the city is .165 higher compared to the village (where the village is not included in the model) controlling for the driver age.

the number of injuries in the camp is 0.178 higher compared to the village (where the village is not included in the model) controlling for the driver age.

Table 3.22[b]: Coefficients

	B	Std. Error	Beta	t	p-value
age	-0.010	0.007	-0.061	-1.376	0.169
friday	0.489	0.324	0.077	1.508	0.132
thursday	0.045	0.291	0.008	0.156	0.876
wednesday	-0.576	0.304	-0.100	-1.895	0.059
tuesday	-0.148	0.297	-0.026	-0.499	0.618
monday	-0.015	0.330	-0.002	-0.044	0.965
saturday	0.078	0.283	0.015	0.277	0.782

Table 3.22 [b] shows that the dummy variables included are insignificant. Also, the number of injuries in friday, thursday, saturday are (0.489, 0.045, 0.078) higher compared to sunday (where sunday is

not included in the model) controlling for the driver age. However, the number of injuries is (0.576, 0.148, 0.015) less in wednesday, tuesday ,monday respectively compared to sunday controlling for the driver age.

Table 3.22[c]: Coefficients

	β	Std. Error	Beta	t	p-value
age	-0.010	0.007	-0.061	-1.376	0.169
spring	0.171	0.209	0.039	0.820	0.413
summer	1.166	0.213	0.258	5.473	0.000
autumn	1.36	0.2	0.254	5.24	0.000

Table 3.22[c] shows that the dummy variables (summer and autumn) are statistically significant. Also, the number of injuries in spring, autumn, summer are (0.171, 1.166, 1.36) higher compared to winter (where winter is not included in the model) controlling for the driver age.

Table 3.22[d]: Coefficients

	β	Std. Error	Beta	t	Sig.
age	-0.010	0.007	-0.061	-1.376	0.169
un valid insurance	0.160	0.842	0.013	0.190	0.850
valid insurance	0.644	0.708	0.069	0.909	0.364

Table 3.22[d] shows that the dummy variables (un valid insurance

and valid insurance) are statistically insignificant. Also, the number of injuries in the un valid and valid insurance are (0.16, 0.644) higher compared to drivers with no insurance (where drivers with no insurance are not included in the model) controlling for the driver age.

Table 3.22[e]: Coefficients

	β	Std. Error	Beta	t	Sig.
age	-0.010	0.007	-.061	-1.376	0.169
valid license	-0.489	0.352	-0.063	-1.388	0.166

Table 3.22[e] shows that the dummy variable (valid license) is statistically insignificant. Also, the number of injuries with the drivers of a valid license is 0.489 less compared to drivers with no valid license (where drivers with no valid license are not included in the model) controlling for the driver age.

3.4 Correlation

Correlation coefficient is a scale (denoted by "r") that explains a relation between two variables which may be positive (that means

when the first variable increase then the other variable does). If we get a negative value, that means that an increase in one variable leads to a decrease with the other variable. Also, the correlation coefficient may take a value($-1 \leq r \leq 1$).

Table 3.23: Correlation Between the Quantitative Variables

	driver age	car age	death
driver age	1	0.014	-0.001
car age	0.014	1	0.079

Table 3.24: Correlation Between the Quantitative Variables

	age	car age	injuries
driver age	1	0.014	-0.077
car age	0.014	1	-0.107

1- Driver age and the car age has a weak positive relation.

2-The driver age and the number of injuries has a weak negative relation.

3-The age of the driver and the number of death has a weak negative relation.

4-The car age has a weak negative relation with the number of injuries.

5-The car age has a positive relation with the number of death.

Conclusions and Future Work

Regression analysis is a statistical method which can be used to study a real life problems using real data. The main two types of regression had been discussed in this work, simple linear regression and multiple linear regression (which has mainly been used).

In this thesis, we deal with a real data concerned with car accidents (the type that causes body injuries or death) in Jenin city during 2018. Analyzing the collected data, we could extract the effect of each independent variable over the dependent variable and decide if there is an effect through a statistical test. Also, we built a regression model which can be used to predict the number of injuries when the different independent variables are known. Several methods were used for analyzing the data such as standard multiple regression, stepwise and hierarchical multiple regression.

Many researches can follow the work presented in this thesis, here are some of them:

1-Applying multiple regression on another case concerned with (health, education, ..)

2-Expand the collected sample size to include another city or other types of car accidents.

3-Using regression to compare between cars accidents in our country and other countries.

Bibliography

M.A Golberg and H.A Cho, Introduction To Regression Analysis ,Wit Press, 2014.

D. Wackerly and W. Menden and R. Scheaffer, Mathematical Statistics With Applications, University of Florida, Cegage Learning, 2008.

J. Rawlings and S. Pantula and D. Dickey, Applied Regression Analysis: A Research Tool, North Carolina State University, 2001.

V. Bewick and L. Cheek and J. Ball, Statistical Review 7: Correlation and regression, (master thesis), University of Brighton, 2003.

G. k. Kanji, 100 Statistical Tests, Sage Publications Ltd, New york, 2006.

Y. Alharbi, Error Variance Estimation in Nonparametric Regression Models,(master thesis), University of Birmingham, 2014.

A. Gelman and J. Hill, Data Analysis Using Regression and Multilevel/Hierarchical Models, Columbia University, 2007.

M. Rouaud, Probability Statistics and Estimation Propagation of Uncertainties in Experimental Measurement, (master thesis), 20017.

A, Esin, Flexibility of Using Com-Poisson Regression Model for Count Data, (Research Paper), 2018.

X. Ma, Refined - Scale Crash Data Analysis Using Multi-Level Regression Models, (Doctoral thesis), Colrado University, 2016.

J. D. Munyon, A Statistical Analysis of Motor Vehicle e Fatalities in the United States, Browling Green State University, (master thesis), 2017.

R. Al jabri, Regression Analysis for Estimation of the Influencing Factors on Accident Injuries in Oman, University of Essex, (master thesis), 2015.

A. Alsherhi, Analysis of Factors Affecting Crash Severity of Pedestrian and Bicycle Crashes Involving Vehicles at Intersections, University of Dayton,(master thesis), 2017.

C. Dhakal, Regression Invention As Statistics, Tribhuvan University,(master thesis), 2018.

P. Gorroochurn, On Galtons Change From "Reversion" to "Regression", (master thesis), 2016.

J. Neter, W. Wasserman, and M.H, Kutner, Applied Linear Regression Models, Homewood, 1989, Australia.

M. tranmer and M. Elliot, Multiple Linear Regression, Cathie Marsh Center for Census and Survey Research, 2008.

M. Kutner and C. Nachtsheim, W. Li, Applied Linear Statistical Models, Emory University, 2004.

J. Danielsson and G. Gistvik, Estimation, Model Selection and Evaluation of Regression Functions in A Least-Squares Monte-Carlo frameWork, (master thesis) Linkoping University, 2014.