University of Belgrade

Faculty of Mathematics

Master thesis

Regression and Correlation

The candidate **Supervisor**

Karima Ibrahim Soufya **Vesna Jevremović**

June 2014

Contains

	References	

Introduction

Regression is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the casual effect of one variable upon another. Regression methods are meant to determine the best functional relationship between a dependent variable Y with one or more independent variables X. The earliest form of regression was the method of least squares, which was published by Legendre in 1805 and by Gauss in 1809. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun. Gauss published a further development of the theory of least squares in 1821. The term “regression” was coined by Sir Francis Galton, while studying the linear relationship between heights of sons and heights of their fathers.

This thesis focuses on tools and techniques for building regression models using real data and assessing their validity. A key theme throughout the thesis is that it makes sense to base inferences or conclusions only on valid models.

We will show that plots are important tool for building regression models and assessing their validity through appropriate diagnostic procedures.

The regression output and plots that appear in thesis have been generated using statistical software R. In addition, real data sets that have appeared in literature are also considered.

The thesis is divided into five chapters. In the first chapter we introduce Pearson`s correlation coefficient, discuss importance of scatter plots and define regression. The second chapter focuses on method of least squares and on statistical properties of regression coefficients. And we show how to test validity of the method and we define prediction intervals. In the third chapter we give examples of nonlinear regression models. The fourth chapter discusses diagnostic procedures for simple linear regression. In the fifth chapter we present our conclusions.

1- Regression Analysis

First we will introduce Pearson`s and Spearman`s correlation coefficients as two different measures of correlation or association between two variables.

1-1 Correlation

We will note that the correlation between the two variables in the population with the Greek letter ρ, and Pearson`s correlation coefficient, as an estimate of ρ, with the letter ***r*** ,Pearson`s correlation coefficient can assume any value in the interval (-1 ,1) .The absolute value of ***r*** (i.e. | ***r*** |) indicate the strength of the relationship between the two variables .As the absolute value of ***r*** approaches 1, the degree of linear relationship between the variables becomes stronger, achieving the maximum when | ***r*** |=1 (i.e. when ***r*** equals +1 or -1). The closer the absolute value of ***r*** is to 0, the weaker the linear relationship is between the two variables. Pearson's correlation coefficient determines the degree to which a linear relationship exists between two variables.

The sign of **r** indicates the nature or direction of the linear relationship which between two variables, the positive sign indicates a direct linear relationship, and the negative sign indicates an indirect linear relationship. A direct linear relationship is one in which a change on one variable is associated with a change on the other variable in the same direction (i.e., an increase on one variable is associated with an increase on the other variable, and a decrease on one variable is associated with a decrease on the other variable).

An indirect relationship is one in which a change on one variable is associated with a change on the other variable in the opposite direction (i.e., an increase on one variable is associated with a decrease on the other variable, and a decrease on one variable is associated with an increase on the other variable).

The use of the Pearson's correlation coefficient assumes that a linear function best describes the relationship between the two variables, and these two variables have normal distribution. If, however, the relationship between the variables is better described by a curvilinear function, Pearson`s correlation coefficient is not appropriate measure of correlation.

Calculation of Pearson's correlation coefficient r -1.1.1

Let (x_1, y_1),(x_2, y_2 ¿ …(x_n, y_n) be *n* paired observations, then Pearson's correlation coefficient is equal to

$$r = \frac{\sum_{i=1}^{n} (x_i - \acute{x})(y_i - \acute{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \acute{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \acute{y})^2}}$$

or simply

$$r=\frac{\frac{\sum_{i=1}^{n} x_i y_i}{n}-\acute{x}\,\acute{y}}{s_x s_y}$$

where $\acute{x}=\frac{\sum_{i=1}^{n} x_i}{n}$, $\acute{y}=\frac{\sum_{i=1}^{n} y_i}{n}$ are sample means and $s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\acute{x})^2}$,

$s_y=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\acute{y})^2}$ are standard sample deviations.

If we use

$$x_i=x_i-\acute{x}$$

$$y_i=y_i-\acute{y}\ ,$$

then we get simplified for Pearson's correlation coefficient

$$r=\frac{\sum_{i=1}^{n} x_i' y_i'}{\sqrt{\sum_{i=1}^{n} x_i'^2}\sqrt{\sum_{i=1}^{n} y_i'^2}}$$

1.1.2- Scatter Diagram

Let us have pairs of values (x_2, y_2 x_1, y_1¿,¿)…(x_n, y_n¿. In scatter diagram the variable X is shown along the *x*-axis and the variable Y is shown along the *y*-axis and all the pairs of values of X and Y are shown by points (or dots) on the graph.

The scatter diagram of these points reveals the nature and strength of correlation between these variable X and Y. Degrees of correlation between two variables are shown on Figure 1. As we can see, when there is no correlation, points on the scatter plot are distributed randomly.
Also, we observe the following:

- If the points lie on a straight line rising from lower left to upper right, then there is a perfect positive correlation between the variables X and Y. If all the points do not lie on a straight line, but their tendency is to rise from lower left to upper right

then there is a positive correlation between the variable X and Y. In these cases the two variables X and Y are in the same direction and the association between the variables is direct.

- If the movements of the variables X and Y are opposite in direction and the scatter diagram is a straight line, the correlation is said to be negative, association between the variables is said to be indirect.

A scatter plot of the data like that given in Figure1 should always be drawn to obtain an idea of the sort of relationship (if any) that exists between two variables (e.g., linear, quadratic, exponential, etc.).

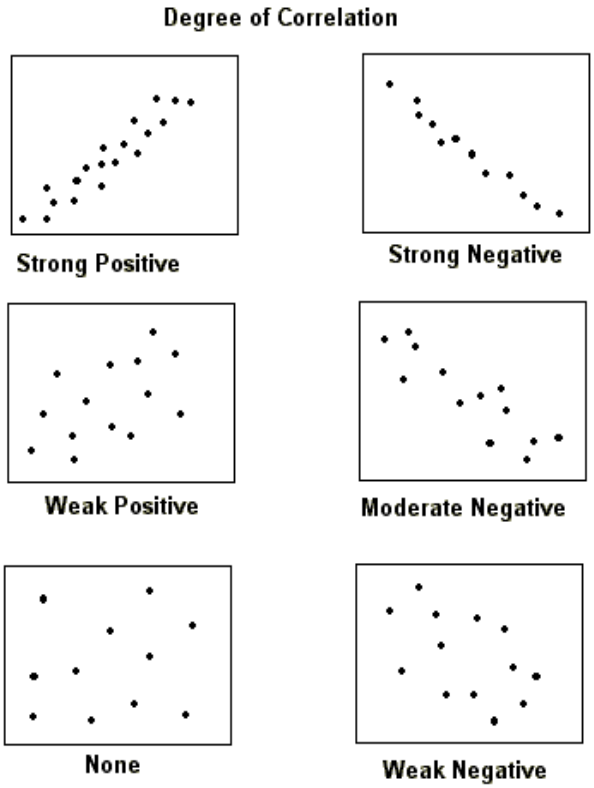

Figure 1. *The degrees of correlation*

Example 1: For the following data draw scatter diagram and calculate Pearson's correlation coefficient.

X	3	5	7	9	11	13	15
Y	5	8	11	13	15	17	19

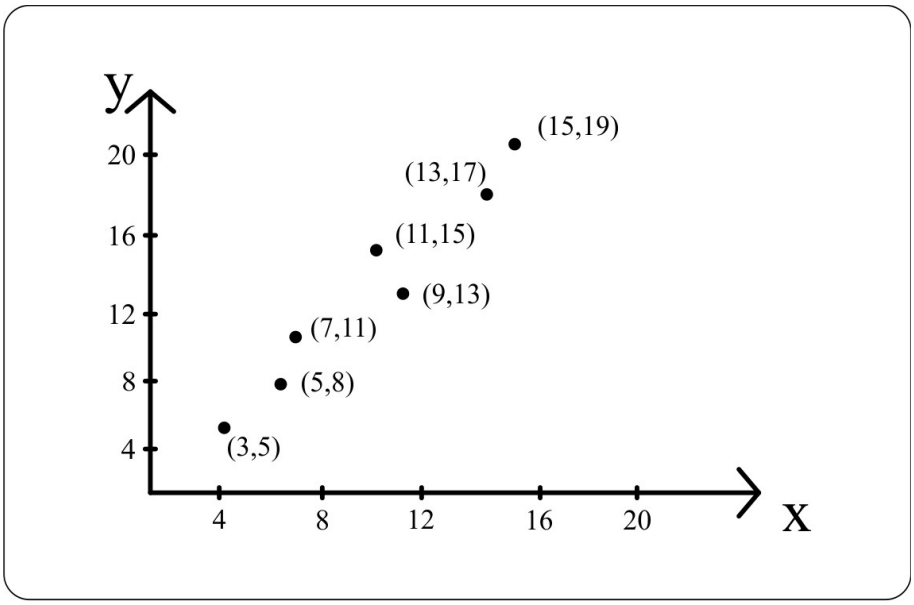

Figure 2. *Scatter plot for given data*

First we calculate sample means, sample standard deviations.

$$\bar{x} = 9, \quad \bar{y} = 12.57, \quad \sum_{i=1}^{7} x_i y_i = 920, \quad s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{7}(x_i - \bar{x})^2} = 4, \quad s_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{7}(y_i - \bar{y})^2} = 4.59$$

And then Pearson's correlation coefficient is equal to

$$r = \frac{\frac{\sum_{i=1}^{7} x_i y_i}{7} - \bar{x}\bar{y}}{s_X \cdot s_Y} = 0.99$$

.

We can see that correlation between variables X and Y is very strong. By looking at the scatter plot it seems that Y is a linear function of X.

It is important to note that strong correlation does not mean that there is cause-effect relationship between two variables. By examining the value of correlation coefficient, we may conclude that two variables are related, but we can`t draw conclusion that one variable causes the other. Variable (or variables) which have not been considered can be responsible for the observed correlation between the two variables.

1.2.3-Spearman`s correlation coefficient

Spearman's correlation coefficient measures degree of monotonic relationship between two random variables. Monotonic relationship can be *increasing* (positive correlation - increase in values of variable X is followed by increase in values of variable Y) or *decreasing* (negative correlation - increase in values of variable X is followed by decrease in values of variable Y). Monotonically increasing, monotonically decreasing and non-monotonic relationship between variables are shown on Figure 3

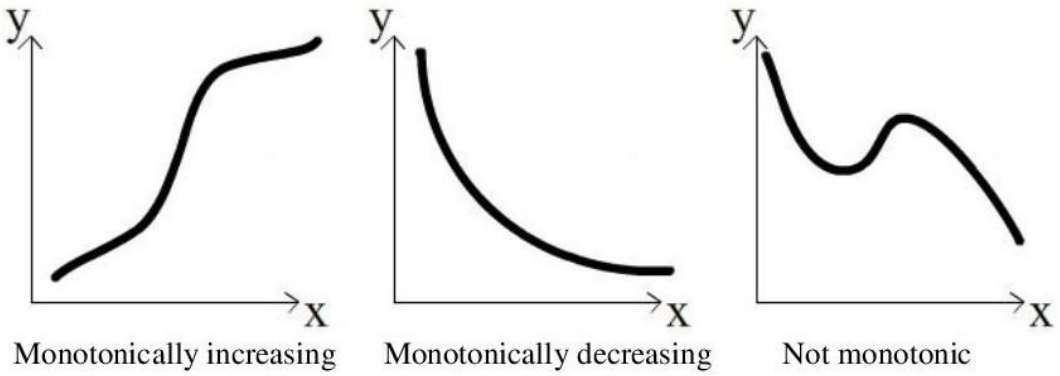

Figure 3. *Monotonic and non-monotonic relationships between variables*

Spearman's population correlation coefficient is denoted with ρ_S and Spearman's sample correlation coefficient with r_S. When this coefficient is equal to -1 or +1, perfect monotonic relationship exists between random variables X and Y.

We will describe how to calculate Spearman's correlation coefficient in three steps. We have pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ of sample values of random variables X and Y.

1. Rank the values of variable X and the values of variable Y, separately. For each element of sample we will have pair of ranks (R_X, R_Y).
2. Form differences of these ranks $d = R_X - R_Y$.

3. Calculate Spearman's correlation coefficient by the formula

$$r_S = 1 - \frac{6\sum d^2}{n \cdot (n^2 - 1)}$$

1.2- Regression

When we are interested what value of random variable Y is expected when X=x, we come to conditional mathematical expectation

$$E(Y \mid X = x),$$

the expected value of Y when X takes the specific value x .

If (X,Y) is bi-dimensional continuous random variable with density function $f(x,y)$ and if $f_x(x)$ is density function of random variable X then

$$Y/¿ = \int_{-\infty}^{\infty} y \frac{f(x,y)}{f_x(x)}$$

$X = x$

¿

E¿

Set of all values $E(Y \mid X = x)$ is set of values of random variable $E(Y \mid X)$, so to calculate $E(Y \mid X)$, we first need to calculate $E(Y \mid X = x)$ for all x.

Random variable $E(Y \mid X) = R(X)$ is called *regression*. For example, if variable X represents day of the week and variable Y sales at a given company, then the regression of Y on X represents the mean (or average) sales on a given day

If (X,Y) is bi-dimensional normally distributed random variable then

$$E(Y \mid X) = \beta_0 + \beta_1 X$$

1.2.1 - Regression Equation

The functional relationship of a dependent variable with one or more independent variables is called a *regression equation*: It is also called prediction equation (or estimating equation).

1.2.2 - Curve of Regression

The graph of the regression equation is called the *curve of regression*: If the curve is a straight line; then it is called the *line of regression*.

1.2.3 - Types of Regression

If there are only two variables under consideration, then the regression is called *simple regression*. If the relationship between X and Y is non-linear, then the regression is *curvilinear* . If there are more than two variables under consideration then the regression is called *multiple regression*. For example, Multiple regression can be used to model relationship between sugar in blood and weight, age and blood pressure of diabetes patients.

1.2.4 - Linear Regression Equation of *Y* on X

Data are collected in pairs (x_1, y_1¿,¿)…(x_2, y_2 x_n, y_n¿, where x_1 denotes the first value of the X -variable and y_1 denotes the first value of the Y -variable. The X variable is called the *predictor variable*, while the Y -variable is called the *dependent variable*.

The regression of Y on X is linear if

$$Y/¿ = \beta_0 + \beta_1 X$$
$$E¿$$

where the unknown parameters β_0 and β_1 determine the intercept and the slope of a specific straight line, respectively. Suppose that $Y_1, Y_2, ..., Y_n$ are independent realizations of the random variable Y that are observed at the values $x_1, x_2, ..., x_n$ of a random variable X. If the regression of Y on X is linear, then for $i = 1, 2, \ldots, n$

$$Y_i = E(Y \mid X = x_i) + e_i = \beta_0 + \beta_1 x_i + e_i$$

where e_i is the random error in Y_i, with zero expectation.

1.2.5 - Assumptions necessary about the regression model

In what follows we shall make the following assumptions:

1. Y is related to x by the simple linear regression model ($y_i = \beta_0 + \beta_1 x_i + e_i$¿.
2. The errors e_1 , $e_2, \ldots, e_n$ are independent of each other.
3. The errors $e_1, e_2, \ldots, e_n$ have a common variance σ^2 .

4. The errors are normally distributed with a mean of 0 and variance σ^2, that is,

$e \vee X \quad N(0,S)$ $e \sim N(0,\sigma^2)$

Because the errors e are normally distributed, we also have that

$Y_1, Y_2, \ldots, Y_i$: $\beta_0 + \beta_1 x, \sigma^2$) then $\dot{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i : N\left(\beta_0 + \beta_1 x, \frac{\sigma^2}{n}\right)$

Methods for checking these four assumptions will be considered in the fourth chapter.

1.2.6- Estimating the population slope and intercept

Suppose for example that variable X represents height and variable Y weight of a person. For a simple regression model the mean weight of individuals of a given height would be a linear function of that height. In practice, we usually have a sample of data instead of the whole population. The slope b_1 β_1 and intercept β_0 are unknown, since these are the values for the whole population. Thus, we wish to use the given data to estimate the slope and the intercept. This can be achieved by finding the equation of the line which "best" fits our data, that is, choose b_0 and b_1 such that $\hat{y}_i = b_0 + b_1 x_i$ $\hat{y} = b_0 + b_1 x_i$ is as "close" as possible to Y_i y_i. We shall refer to $\hat{y}_i$ as $the\, i\, th$ predicted value or the fitted value of y_i Y_i. For estimating these unknown parameters we will use the method of least squares.

2 -The Method of Least Squares

This method of curve fitting was suggested early in the nineteenth century by the French mathematician Adrian Legendre. The method of least squares assumes that the best fitting line is one for which the sum of the squares of the vertical distances of the point (*x*,*y*) from the line is minimal.

For the simple linear regression we can use least squares method to find the estimators b_0 and b_1 such that the sum of the squared distances between value y_i and predicted value $\hat{y}_i$

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

reaches the minimum among all possible choices of regression coefficients β_0 b_o and b_1 β_1 .

Vertical distances of points from regression line are shown on Figure 4

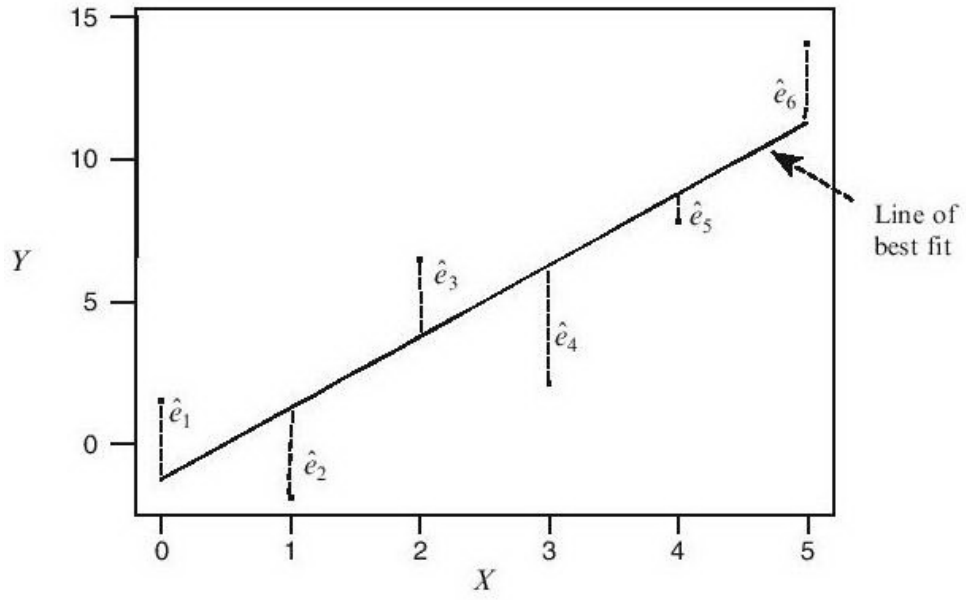

Figure 4. *Vertical distances of points from regression line*

Mathematically, the least squares estimates of the simple linear regression could be obtained by solving the following system:

$$\frac{\partial S}{\partial \beta_0} = 0, \frac{\partial S}{\partial \beta_1} = 0$$

It is more convenient to solve this system using the fitted linear model:

$$\hat{y}_i = \beta_o^{¿} + \beta_1 ¿ \quad ,$$

where

$$\beta_o = \beta_o^{¿} - \beta_1 \acute{x} \quad .$$

Now sum of squared distances equals to

$$S = \sum_{i=1}^{n} \left[y_i - \left(\beta_0^{¿} + \beta_1 \left(x_i - \acute{x} \right) \right) \right]^2$$

and we need to solve the following system

$$\frac{\partial S}{\partial \beta_0^{¿}} = 0, \frac{\partial S}{\partial \beta_1} = 0$$

Taking the partial derivatives with respect to $\beta_0^{¿} \wedge \beta_1$ we get system of equations (called *normal equations*).

$$\sum_{i=1}^{n} \left[y_i - \left(\beta_o^{¿} + \beta_1 \left(x_i - ¿ \right) \right) \right] = 0$$

$$\sum_{i=1}^{n} \left[y_i - \left(\beta_o^{¿} + \beta_1 \left(x_i - \acute{x} \right) \right) \right] \left(x_i - \acute{x} \right) = 0$$

Note that

$$\sum_{i=1}^{n} y_i = n\beta_0^{¿} + \sum_{i=1}^{n} \beta_1 \left(x_i - \acute{x} \right) = n\beta_0^{¿}$$

Therefore, we have

$$b_0^{¿} = \hat{\beta}_0^{¿} = \frac{1}{n} \sum_{i=1}^{n} y_i = \acute{y} .$$

Substituting $b_0^{¿}$ by $\acute{y}$ we obtain

$\acute{x}\big)$

$$\sum_{i=1}^{n}\left[y_i-\left(\acute{y}+\beta_1\left(x_i-¿\right]\left(x_i-\acute{x}\right)=0\right.\right.$$

Now it is easy to see

$$b_1=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\acute{y}\right)\left(x_i-\acute{x}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\acute{x}\right)^2}=\frac{S_{XY}}{S_{XX}}$$

and

$$b_0 \;=\; b_o^{¿} \;-\; b_1\acute{x}=\acute{y}-b_1\acute{x}$$

The fitted value of the simple regression is defined as $\hat{y}_i=b_o+b_1x_i$.

We have $y_1,y_2,\ldots,y_i:N\left(\beta_0+\beta_1x,\sigma^2\right)$ then $\acute{y}_n=\frac{1}{n}\sum y_i:N\left(\beta_0+\beta_1x,\frac{\sigma^2}{n}\right)$

$$b_1=\frac{\frac{1}{n}\sum\left(y_i-\acute{y}_n\right)\left(x_i-\acute{x}\right)}{\frac{1}{n}\sum\left(x_i-\acute{x}\right)^2}=\frac{\sum_{i=1}^{n}y_i\left(x_i-\acute{x}\right)}{\sum_{i=1}^{n}\left(x_i-\acute{x}_n\right)^2}$$

Estimator b_1 is normally distributed, as linear combination of normally distributed random variables Y_i . Then, it also follows that estimator b_0 has normal distribution.

2.1- Residuals

The difference between an observed y_i and the fitted value of $\hat{y}_i$, $e_i=y_i-\hat{y}_i$ is referred to as the ith regression residual. Its magnitude reflects the failure of the least squares line to “model” for that particular point.

We can use these residuals to estimate unknown variance σ^2 of random errors ($\sigma^2 = var(e)$¿ with estimator $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2$.

Example 2: ***A regression model for the timing of production runs***

We shall consider the following data: variable Y represents the time taken (in minutes) for a production run (run time) and variable X the number of items (run size) produced for 20 randomly selected orders. We wish to develop an equation to model the relationship between variables Y and *X*. The data are given in Table 1 and corresponding scatter plot in Figure 5.

Table 1. The production data

Case	Run time	Run size	Case	Run time	Run size
1	195	175	11	220	337
2	215	189	12	168	58
3	243	344	13	207	146
4	162	88	14	225	277
5	185	114	15	169	123
6	231	338	16	215	227
7	234	271	17	147	63
8	166	173	18	230	337
9	253	284	19	208	146
10	196	277	20	172	68

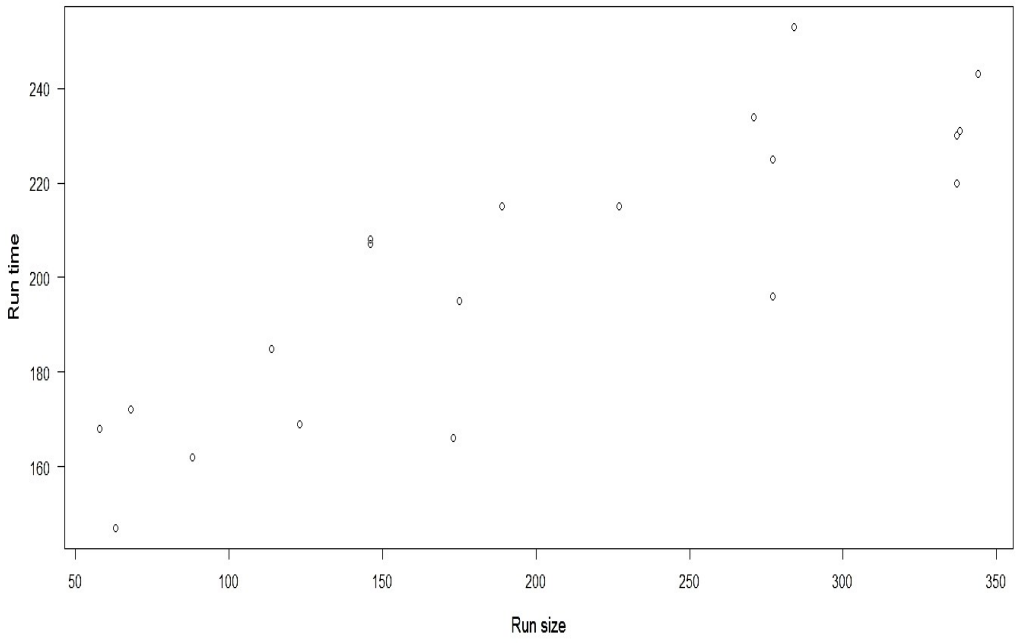

Figure 5. *A scatter plot of the production data*

Scatter plot of the production data shows us that there might be linear dependence between run time and run size. We will consider the linear regression model $Y = \beta_0 + \beta_1 x + e$. Now we will calculate least squares estimators of the unknown parameters β_0 and β_1.

We have that $\bar{x} = 201.75,\ \bar{y} = 202.05\ \sum_{i=1}^{20}(x_i - \bar{x})(y_i - \bar{y}) = 49638.25$, and $\sum_{i=1}^{20}(x_i - \bar{x})^2 = 191473.8$, so the estimators b_1 and b_0 are equal to

$$b_1 = \frac{\sum_{i=1}^{20}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{20}(x_i - \bar{x})^2} = \frac{49638.25}{191473.8} = 0.2592$$

$$b_0 = \bar{y} - b_1\bar{x} = 149.7477$$

The fitted regression line is

$$\hat{y} = 0.26 + 149.75x$$

Now we shall give some properties of estimators in simple linear regression model, always having in mind the assumptions given on page 9.

2.2- Properties of estimator of the slope

Theorem 1: *The least squares estimator b_1 is an unbiased estimator of β_1.*

Proof

Here we take $x_i i = 1,2,\ldots,n$ as constants ,while $y\ Y$ is a random variable .

$$E(b_1) = E\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}} E\frac{1}{n}\sum_{i=1}^{n}(Y_i - \acute{Y})(x_i - \acute{X})$$

Using the fact that

$$\sum_{i=1}^{n}\left(x_i-\acute{X}\right)=0$$

We get:

$$E\left(b_1\right)=\frac{1}{S_{XX}}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\acute{x}\right)Ey_i=\frac{1}{S_{xx}}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\acute{x}\right)\left(\beta_o+\beta_1 x_i\right)=\frac{1}{S_{xx}}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\acute{x}\right)=¿$$

$$¿\frac{1}{S_{xx}}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\acute{x}\right)\beta_1\left(x_i-\acute{x}\right)=\frac{1}{S_{xx}}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\acute{x}\right)^2\beta_1=\frac{S_{xx}}{S_{xx}}\beta_1=\beta_1 \quad .\blacksquare$$

Theorem 2: *Variance of the estimator of the slope is*

$$Var\ (\ b_1\ )=\frac{\sigma^2}{nS_{xx}}$$

Proof

$$¿Y_i-\acute{y}$$

$$(¿)\left(x_i-\acute{x}\right)=\frac{1}{S_{xx}^2}Var\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\left(x_i-\acute{x}\right)\right)$$

$$\frac{1}{n}\sum_{i=1}^{n}¿ \qquad =$$

$$Var\left(b_1\right)=Var\left(\frac{S_{xy}}{S_{xx}}\right)=\frac{1}{S_{xx}^2}Var\,¿$$

$$¿\frac{1}{S_{xx}^2}\cdot\frac{1}{n^2}\sum_{i=1}^{n}\left(x_i-\acute{x}¿^2 Var\left(Y_i\right)=\frac{1}{S_{xx}^2}\cdot\frac{1}{n^2}\sum_{i=1}^{n}\left(x_i-\acute{x}¿^2\sigma^2=\frac{\sigma^2}{nS_{xx}}\right.\quad .\blacksquare$$

Theorem 3: *The least square estimator* b_1 *and* $\acute{y}$ *are uncorrelated. Then under the normality assumption of* y_i *for* i *=1,2,…,n,* b_1 *and* $\acute{y}$ *are normally distributed and independent.*

Proof

As we mentioned before that b_1 and $\acute{y}$ are normally distributed and independent, let us calculate the covariance between b_1 and $\acute{y}$.

$$x$$
$$¿$$
$$x$$
$$¿i-¿$$
$$(¿¿(y_i-\acute{y}),\acute{y}¿)=¿$$
$$¿$$
$$\sum_{i=1}^{n} ¿$$

$$Cov(b_{1,}\acute{y})=Cov\left(\frac{S_{xy}}{S_{xx}},\acute{y}\right)=\frac{1}{nS_{xx}}Cov(S_{xy},\acute{y})=\frac{1}{nS_{xx}}Cov¿$$

$$x_i$$
$$(¿-\acute{x})\,y_{i,}\acute{y}$$
$$\sum_{i=1}^{n} ¿$$
$$¿$$
$$x_i$$
$$¿$$
$$x$$
$$-¿$$
$$(¿y_{i,},\sum_{i=1}^{n} y_i¿)=¿$$
$$\sum_{i=1}^{n} ¿$$
$$¿\frac{1}{nS_{xx}}Cov¿$$

¿

$(¿ x_i - \acute{x}) Cov(y_i, y_j)$

$\sum_{j=1}^{n}$ ¿

¿ $\frac{1}{n^2 S_{xx}} Cov \sum_{i=1}^{n}$ ¿

$Ee_i = 0$ e_i

Note since that, and s are independent, we can write

$Cov(y_i, y_j)$ =E y_i ¿ $-Ey_i$ ¿ $y_j - Ey_i$ ¿] =E ($E i, E j$ ¿ $e_i e_j$) = $\begin{cases} \sigma^2 if\ i=j \\ o\ if\ i \neq j \end{cases}$

Thus, we conclude that

x_i

$(¿ - \acute{x})\sigma^2 = 0$

$Cov(b_1, \acute{y}) = \frac{1}{n^2 S_{xx}} \sum_{i=1}^{n}$ ¿

Recall that zero correlation is equivalent to the independence between two normal variables.

Thus, we conclude that b_1 and $\acute{y}$ are independent. ■

2.3- Properties of estimator of the intercept

Theorem 4. *The least squares estimator* b_0 *is an unbiased estimator of* β_o

Proof

Here also we take $x_i, i = 1,2,\ldots,n$ as constants , while Y is a random variable .

$\beta_o + ¿ \beta_1 \frac{1}{n} \sum_{i=1}^{n} x_i - \beta_1 \acute{x} = \beta_o$

$E b_o = E(Y_i - b_1 \acute{x}) = (\frac{1}{n} \sum_{i=1}^{n} E Y_i) - \acute{x} E b_1 = \frac{1}{n} \sum_{i=1}^{n}$ ¿ . ■

-

Theorem 5*: *Variance of the estimator of the slope is:

$$Var(b_o)=\left(\frac{1}{n}+\frac{\acute{x}^2}{nS_{xx}}\right)\sigma^2$$

Proof

$$Var(b_o)=Var(\acute{y}-b_1\acute{x})=Var(\acute{y})+(\acute{x})^2Var(b_1)=\frac{\sigma^2}{n}+\acute{x}^2\frac{\sigma^2}{nS_{xx}}=\left(\frac{1}{n}+\frac{\acute{x}^2}{nS_{xx}}\right)\sigma^2 \quad .\blacksquare$$

2.4 - Testing the validity of the model

If the slope of the regression line β_1 is equal to zero, then the linear regression model reduces to $Y_i=\beta_0+e_i$ and we can not predict values of variable Y based on the known values of variable X. To test null hypothesis $H_0:\beta_1=0$ against the alternative $H_1:\beta_1\neq 0,$ we can use the test statistics

$$T=\frac{b_1-\beta_1}{\hat{\sigma}}\sqrt{S_X^2}$$

which has under null hypothesis, Student t-distribution with *n-2* degrees of freedom. We used the fact that estimator $b_1:N\left(\beta_1,\frac{\sigma^2}{S_X^2}\right)$ $b_1:N(\beta_1,\frac{\sigma^2}{s^2})$ $S_X^2=\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$, where .

Test is of the form: reject H_0 at level α if $|T|\geq c$ $c=t_{n-2,1-\frac{\alpha}{2}}$, where is obtained from table of Student distribution t_{n-2}. Otherwise, we do not reject null hypothesis and we conclude that linear regression model could be valid model for our data.

2.5- Prediction intervals for the actual value of Y

In this section we consider the problem of finding a prediction interval for the

actual value of Y at x_p, a given value of X. We note $Y_p \vee X = x_p$ as this prediction value of variable Y. We have that $Y_p = \beta_1 x_p + \beta_0 + e_p$ and $\hat{Y}_p = b_1 x_p + b_0$.

First we will derive expectation and variance of $Y_p - \hat{Y}_p$

$$E\left(Y_p - \hat{Y}_p\right) = E\left(\beta_1 . x_P + \beta_0 + e_P - b_1 . x_p + b_0\right) = E\left(b_1 - \beta_1\right) x_p + E\left(b_0 - \beta_0\right) + E\, e_P = 0 \quad \backslash$$

$$Var\left(Y_p - \hat{Y}_p ¿ = Var\left(\beta_1 . x_p + \beta_0 + e_p - b_1 x_p - b_0\right) = Var\left(e_p - b_1 x_p - b_0\right) = ¿\right.$$

$$\begin{matrix} b \\ b \end{matrix}$$

$$(¿¿0) - 2\,Cov\left(e_p, b_0\right) + 2\,Cov\left(b_1 x_p, b_0\right) - 2\,cov\left(e_p, b_1 . x_p\right)$$

$$(¿¿1) + Var ¿$$

$$Var\left(e_p\right) + x_p^2\, Var ¿$$

As we have that $Cov\left(b_1, e_p\right) = Cov\left(b_0, e_p\right) = 0$ and $Cov(b_1, b_0) = -\sigma^2 \frac{\bar{x}_n}{S_X^2}$, where

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

, we get

$$Var\left(\hat{Y}_p - Y_p\right) = x_p^2 \frac{\sigma^2}{S_X^2} + \frac{\sigma^2}{n} + \sigma^2 \frac{\bar{x}_n^2}{S_X^2} + \sigma^2 - 2x_p \sigma^2 \frac{\bar{x}_n}{S_X^2} = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_p - \bar{x}_n)^2}{S_X^2}\right),$$

$$\hat{Y}_p - Y_p : N\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_p - \bar{x}_n)^2}{S_X^2}\right)\right)$$

So, we get that . Standardizing and replacing σ^2 by $\hat{\sigma}^2$ gives

$$T = \frac{\hat{Y}_p - Y_p}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}_n\right)^2}{S_X^2}}} : t_{n-2}$$

A 100(1–α) % prediction interval for Y_p, the value of Y at $X = x_p$, is given by

$$I_{Y_p} = \left(\hat{Y}_p - c\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}_n\right)^2}{S_X^2}}, \hat{Y}_p + c\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}_n\right)^2}{S_X^2}} \right),$$

where $c = t_{n-2,1-\frac{\alpha}{2}}$.

2.6 - Coefficient of determination

The square of the correlation between variables Y and X is called the *coefficient of determination* and is denoted with R^2. It represents the proportion of the total sample variability in the Y's explained by the regression model.

Higher value of R^2 is preferable because it means that regression model is describing well relationship that exists between observed variables.

3- Nonlinear Models

Obviously, not all $x\,y$-relationships can be described by straight lines. Curvilinear relationships of all sorts can be found in every field of application. Most of these nonlinear models can be fitted using least squares method, provided the data have been initially "linearized" by a suitable transformation.

3.1 - Polynomial regression

Polynomial regression is used when dependence between variables Y and X is of the type

$$Y = a_0 + a_1 X + a_2 X^2 + \ldots + a_m X^m + e$$

We will consider the case when $m=2$.

When a set of points exhibits a parabolic trend then we can use quadratic function $Y = a + bx + cx^2 + e$ $y = a + bx + cx^2$ We can find unknown coefficients a, b and c using the method of least squares. We get system of normal equations

$$na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i$$

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i$$

By solving this system of equations, we find estimates of a, b and c.

Example 3 : Find a formula for the line of the form $y = a + bx + cx^2$ $Y = a + bx + cx^2 + e$ to fit the following data

x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
y	3.16	3.20	3.25	3.30	3.27	3.24	3.19	3.15	3.10	2.99

Solution: Substituting the values of

$$\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} x_i^3, \sum_{i=1}^{n} x_i y_i, \sum_{i=1}^{n} x_i^2 y_i, \sum_{i=1}^{n} y_i$$

and n in above normal equations we get

$$10a + 4.5b + 2.85c = 31.85$$

$$4.5a + 2.85b + 2.025c = 14.178$$

$$2.85a + 2.025b + 1.5333c = 8.89$$

Solving these equations, we obtain

$\hat{a} = 3.16,\ \hat{b} = 0.63$ and $\hat{c} = -0.91$

The required equation of quadratic regression is

$$\hat{y} = 3.16 + 0.63x - 0.91x^2$$

Scatter plot with fitted regression function is given below.

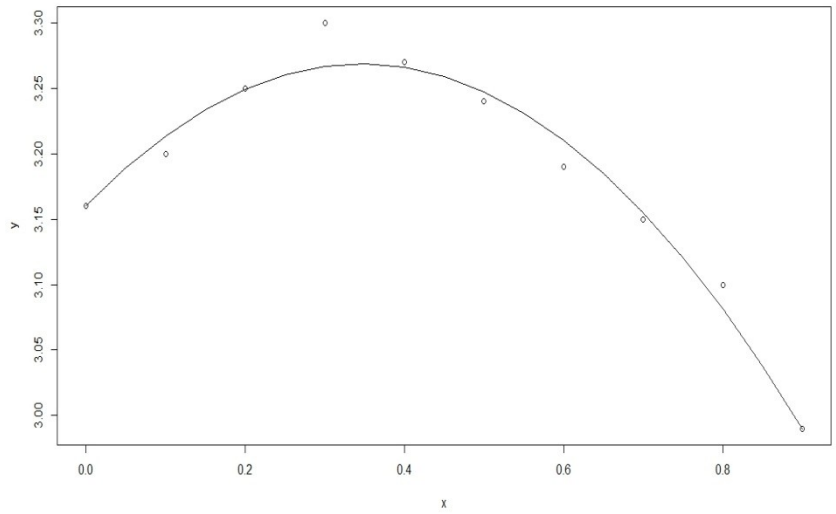

Figure 6. *Scatter plot and fitted quadratic function for given data*

3.2 - Exponential Regression

Suppose the relationship between two variables is best described by an exponential function of the form

$$y = a\,e^{bx}$$

In case it is possible to obtain approximate solutions using numerical methods or the idea of “linearization”.

Transforming the previous equation by taking logarithms on both sides we get

$$\ln y = \ln a + b x$$

which implies that $\ln y$ and x have a linear relationship. When we apply the least squares method to x and $\ln y$ that yield to

$$b = \frac{n\sum_{i=1}^{n} \ln y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} \ln y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and

$$a = \frac{\sum_{i=1}^{n} \ln y_i - b\sum_{i=1}^{n} x_i}{n}$$

Example 4 : Find a least square curve of the form $y = a e^{bx} \quad (a>0)$ to the data given below

x	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0
y	4.46	5.72	6.71	6.31	8.92	11.32	12.78	16.46	20.31	25.54	29.67

Solution :

Consider $y = a e^{bx}$, applying logarithm (with base e on both sides , we get

$$y = \ln a + bx \quad (1)$$

Taking $y = \ln Y$, the equation (1) can be written as $Y = \tilde{a} + bx$ (2) $Y = \grave{a} + bx$, where

$\tilde{a} = \ln a$ $a = \ln a$.

Equation (2) is linear equation of Y on x. The normal equations are

$$n \cdot \tilde{a} + b\sum_{i=1}^{11} x_i = \sum_{i=1}^{11} Y_i$$

$$\tilde{a}\sum_{i=1}^{11} x_i + b\sum_{i=1}^{11} x_i^2 = \sum_{i=1}^{11} x_i Y_i$$

After calculations, we get

$11\tilde{a}$ +22b=26.59

$22\tilde{a} + 48.4b = 57.392$

After solving these equations, we obtain

$\hat{\tilde{a}} = 0.497$ $\hat{b} = 0.96$,

$a = \ln 0.0001 \rightarrow a = 1.0001$ $\hat{a} = e^{\hat{\tilde{a}}} = 1.64$
$\grave{a} = 0{,}0001 \rightarrow \ln$ So we get

The required curve is $\hat{y} = 1.64 \cdot e^{0.96x}$. Scatter plot with fitted regression line is given below.

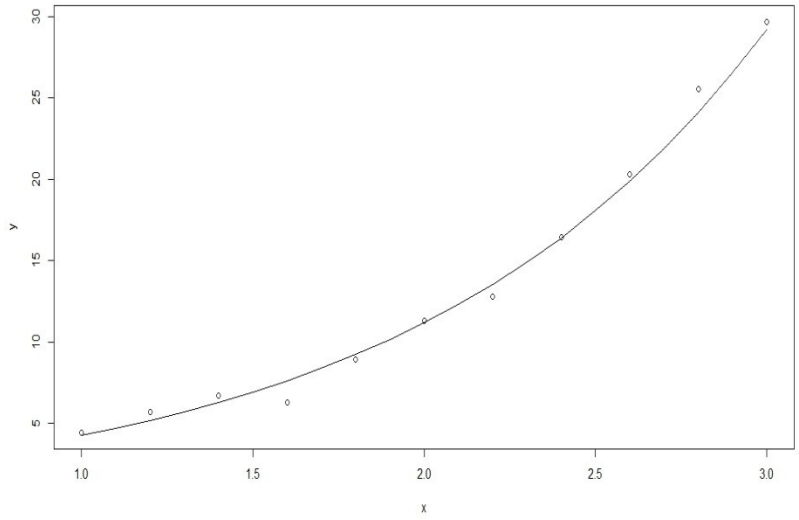

Figure 7. *Scatter plot and fitted exponential function for given data*

3.3 - Other curvilinear models

We will consider two other nonlinear regression models.

1. Let $Y = ax^b$. Taking logarithms of both sides of equation we get $\ln Y = \ln a + b \ln x$.
   Using the notes $\ln Y = U$, $\ln a = \alpha$ and $\ln x = v$ we get linear regression model
   $$U = \alpha + bv$$
   .
   We find estimates of α and b by solving the following system of equations
   $$n\alpha + b\sum_{i=1}^{n} v_i = \sum_{i=1}^{n} U_i$$
   $$\alpha\sum_{i=1}^{n} v_i + b\sum_{i=1}^{n} v_i^2 = \sum_{i=1}^{n} U_i v_i$$
   We find estimate of a as
   $$\hat{a} = e^{\hat{\alpha}}$$
2. Let $Y = \frac{1}{ax + b}$. We use $U = \frac{1}{Y}$ and we get linear regression model
   $$U = ax + b$$
   .
   We find estimates of a and b by solving the following system of equations:

$$a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} U_i$$

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} U_i x_i$$

4 -Diagnostics for simple linear regression

In Chapter 2 we studied the simple linear regression model. Throughout Chapter 2, we assumed that the simple linear regression model was a valid model for the data, that is, the conditional mean of Y given X is a linear function of X and the conditional variance of Y given X is constant. In other words,

$$E(Y/X) = \beta_0 + \beta_1 X \wedge var(Y/X) = \sigma^2$$

In Section 4.1, we start by examining the important issue of deciding whether the model under consideration is indeed valid. In Section 4.2, we will see that when we use a regression model we implicitly make a series of assumptions. We then consider a series of tools known as regression diagnostics to check each assumption.

4.1 -Valid and Invalid Regression Models: Anscombe's Four Data Sets

Throughout this section we shall consider four data sets constructed by Anscombe. This example illustrates the point that looking only at the numerical regression output may lead to very misleading conclusions about the data, and lead to accepting the wrong model. The data are given in the table below (Table 4.1) and are plotted in Figure 8. Values of random variable Y differ in each of four data sets, while x-values of variable X are same in first three sets.

Table 4.1 Anscombe's four data sets

Case	x_1	x_2	x_3	x_4	y_1	y_2	y_3	y_4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.1	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.1	5.39	12.5
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

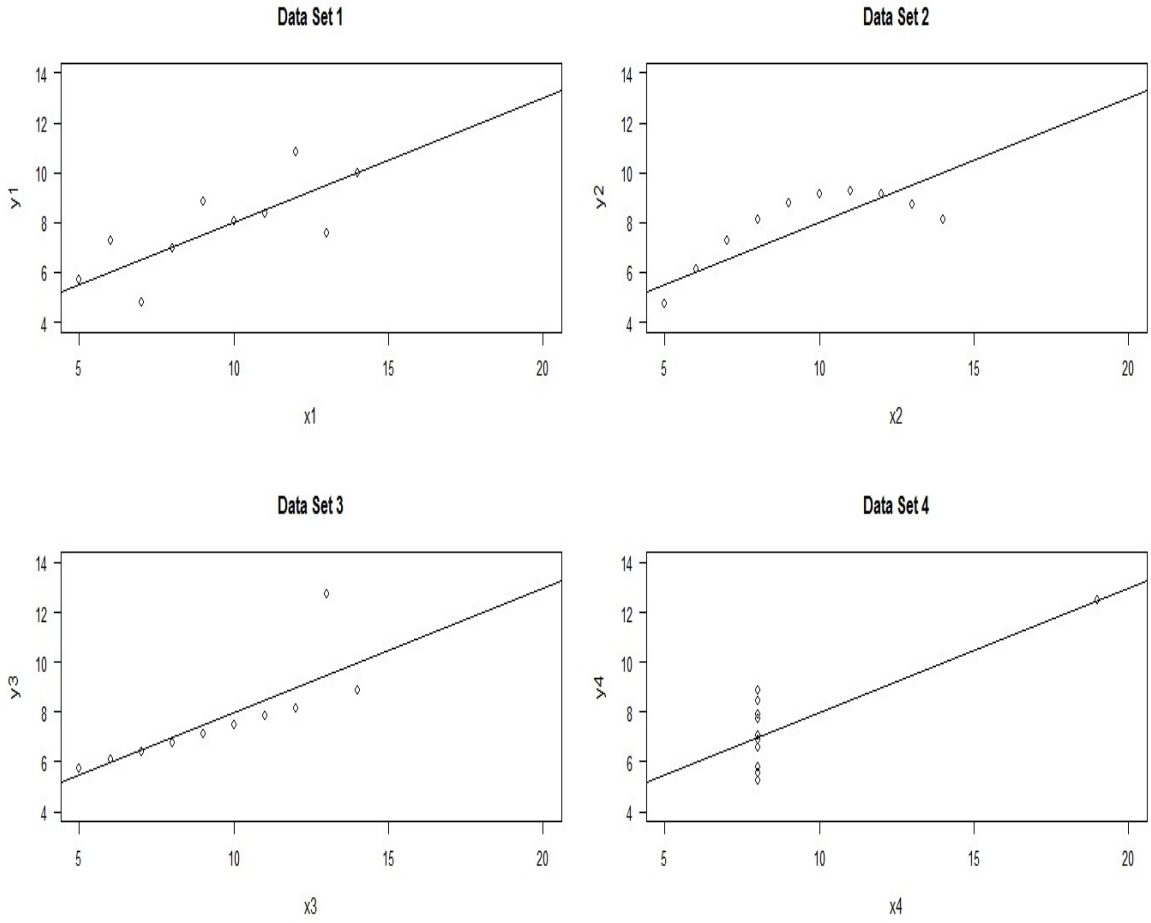

Figure 8. *Plots of Anscombe's four data sets*

When a regression model is fitted to data sets 1, 2, 3 and 4, in each case the fitted regression model is $\hat{y}$ =3.0 +0.5 x,

The regression output for data sets 1 to 4 is given below. The regression output for the four data sets is identical (to two decimal places) in every respect. In all cases, results of *t*-test for validity of a model (slope of regression line is different from zero) are significant (p-value is smaller than 0.05). Looking at Figure 8 it is obvious that a straight-line regression model is appropriate only for Data Set 1. On the other hand, the data in Data Set 2 seem to have non linear rather than a straight-line relationship. The third data set has an extreme outlier that should be investigated. For the fourth data set, the slope of the regression line is solely determined by a single point, namely, the point with the largest x –value.

Regression output from R

First model

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x1            0.5001     0.1179   4.241  0.00217 **
---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6665,     Adjusted R-squared: 0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

Second model

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576
x2             0.500      0.118   4.239  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6662,     Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF, p-value: 0.00217
```

Third model

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0025     1.1245   2.670  0.02562 *
x3            0.4997     0.1179   4.239  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6663,     Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

Forth model

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671  0.02559 *
x4            0.4999     0.1178   4.243  0.00216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6667,     Adjusted R-squared: 0.6297
```

```
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```


This example demonstrates that the numerical regression output should always go together with analysis to ensure that an appropriate model has been fitted to the data. In this case it is enough to look at the scatter plots in Figure 8 to determine whether an appropriate model has been fit. However, in some situations we shall need some additional tools in order to check the validity of the fitted model

4.2- Plots of residuals

One tool we will use to validate a regression model is one or more plots of residuals (or standardized residuals, which will be defined later in this chapter). These plots will enable us to assess visually whether an appropriate model has been fit to the data.

Figure 9 provides plots of the residuals against *X* for each of Anscombe's four data sets. There is no pattern can be recognized in the plot of the residuals for data Set 1 against X_1 .

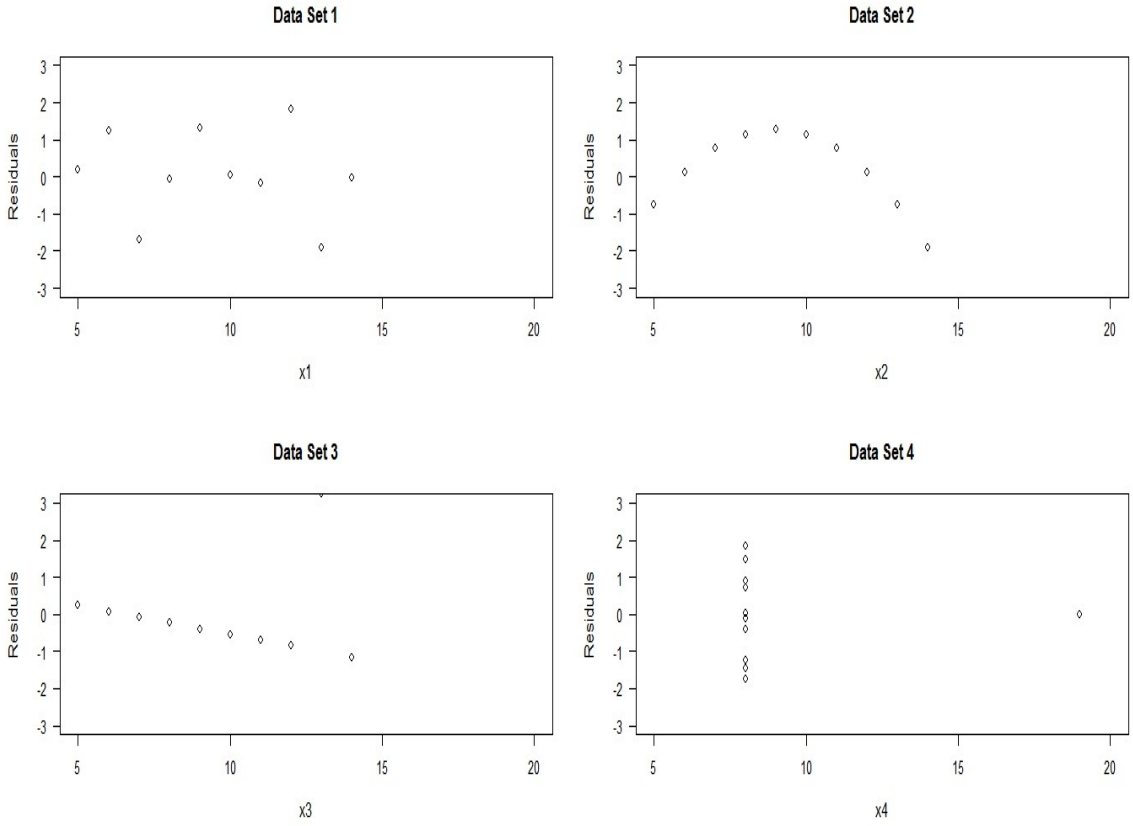

Figure 9. *Residual plots for Anscombe's data sets*

We shall see next that this indicates that an appropriate model has been fit to the data. We shall see that a plot of residuals against *X* that produces a random pattern indicates an appropriate model has been fit to the data. Additionally, we shall see that a plot of residuals against *X* that produces a non random pattern indicates an incorrect model has been fit to the data.

Using Plots of Residuals to Determine Whether the Proposed Regression Model Is a Valid Model

One way of checking whether a valid simple linear regression model has been fit is to plot residuals versus *x* and look for patterns. If no pattern is found then this indicates that the model provides an adequate summary of the data, i.e., is a valid model. If a pattern is found then the shape of the pattern provides information on the function of *x* that is missing from the model.

For example, suppose that the true model is a straight line

$$Y_i=\beta_0+\beta_1 x_i+e_i$$

And we fit a straight line $\hat{y}_i=b_0+b_1 x_i$. Then, assuming that the least squares estimates $b_0 \wedge b_1$ are close to the unknown population parameters $\beta_0 \wedge \beta_1$, we find that

$$\hat{e}_i=Y_i-\hat{y}_i \;=\; (\beta_0-b_0)+(\beta_1-b_1)x_i \;+\; e_i \approx e_i$$

that is, the residuals should resemble random errors. If the residuals vary with *x* then this indicates that an incorrect model has been fit. For example, suppose that the true model is a quadratic

$$Y_i=\beta_0+\beta_1 x_i+\beta_2 x_i^2+e_i$$

and that we fit a straight line

$$\hat{y}_i=b_0+b_1 x_i$$

Then, somewhat simplistically assuming that the least squares estimates $\hat{\beta}_0 \wedge \hat{\beta}_1$ are close to the unknown population parameters β_0 and β_1 , we find that

$$\hat{e}_i=Y_i-\hat{y}_i=(\beta_0-b_0)+(\beta_1-b_1)x_i+\beta_2 x_i^2+e_i \approx \beta_2 x_i^2+e_i$$

Example of a Quadratic Model

Suppose that *Y* is a quadratic function of *X* without any random error. Then, the residuals from the straight-line fit of *Y* and *X* will have a quadratic pattern. Hence, we can conclude that there is need for a quadratic term to be added to the original straight-line regression model. Anscombe's data set 2 is an example of such a situation. Figure 10 contains scatter plots of the data and the residuals from a straight-line model for data set 2. As expected, a clear quadratic pattern is evident in the residuals in Figure 10.

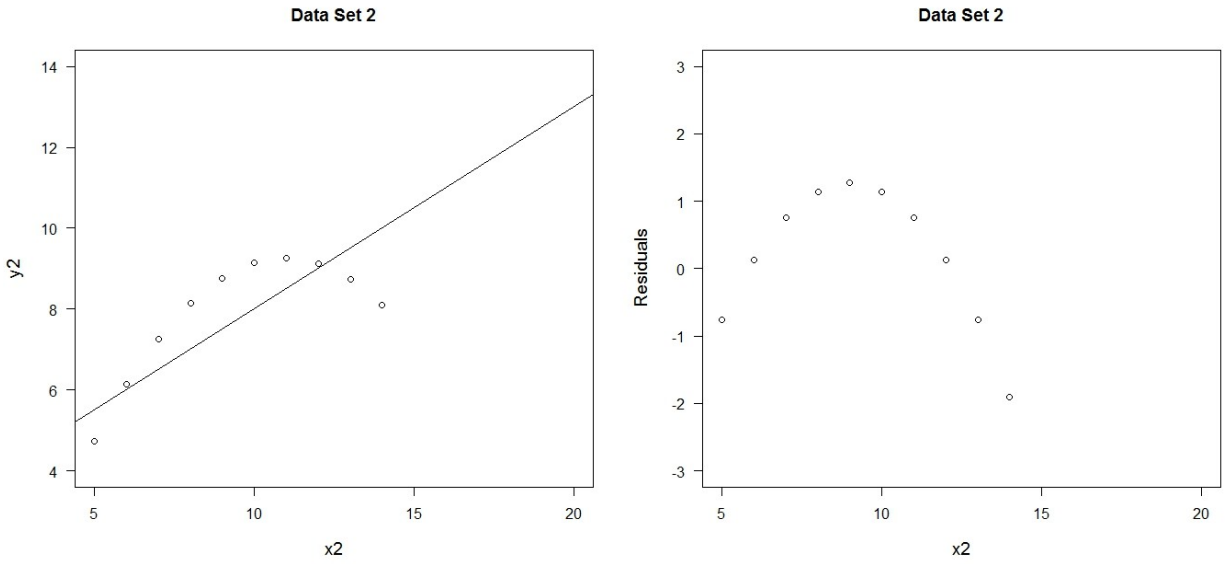

Figure 10 *Scatter plots and the residuals from a straight-line model for data set 2*

4.3 - Regression Diagnostics: Tools for Checking the Validity of a Model

We next look at tools *(called regression diagnostics*) which are used to check the validity of all aspects of regression models. When fitting a regression model we will discover that it is important to:

1. Determine whether the proposed regression model is a valid model (i.e., determine whether it provides an adequate fit to the data). The main tools we will use to validate regression assumptions are plots of standardized residuals. The plots enable us to assess visually whether the assumptions are being violated.

2. Determine which (if any) of the data points have x -values that have an unusually large effect on the estimated regression model (such points are called *leverage points*).

3. Determine which (if any) of the data points are outliers , that is, points which do not follow the pattern set by the rest of the data, when one takes into account the given model.

4. If leverage points exist, determine whether each is a bad leverage point. If a bad leverage point exists we shall assess its influence on the fitted model.

5. Examine whether the assumption of constant variance of the errors is reasonable.

We begin by looking at the second item of the above list, leverage points, as these will be needed in the explanation of standardized residuals.

4.3.1 - Leverage Points

Data points which have considerable influence on the fitted model are called leverage points. To make things as simple as possible, we shall begin by describing leverage points as either “good” or “bad.”

Example of a “good” and a “bad” leverage point

Twenty points are randomly generated from a known straight-line regression model. We get a plot like that shown in Figure 11. One of the 20 points has an x -value which makes it distant from the other points on the x -axis. We shall see that this point, which is marked on the plot, is a good leverage point. True population regression line (namely, $Y_i = \beta_0 + \beta_1 x_i$) and the least squares regression line (namely, $\hat{y}_i = b_0 + b_1 x_1$) are marked on the plot.

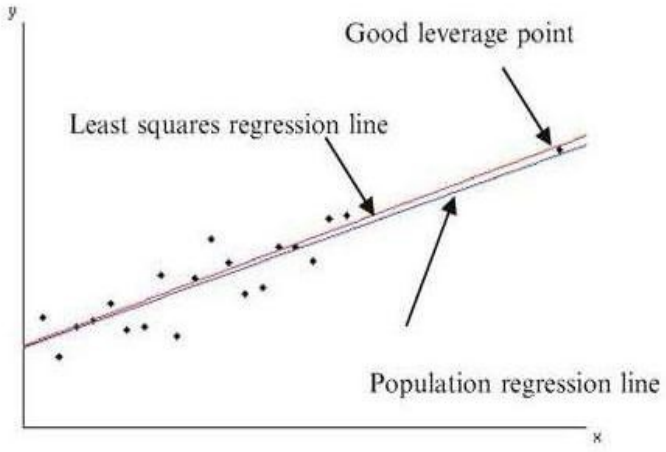

Figure 11. *Good leverage poiint*

Next we move one of the points away from the true population regression line. In particular, we focus on the point with the largest x -value. Moving this point vertically down (so that its x -value stays the same) produces the results shown in Figure 12. Notice how in the least squares regression line has changed dramatically in response to changing the Y -value of just a single point. The least squares regression line has been levered down by single point. Hence we

call this point a *leverage point*. It is a *bad leverage point* since its Y -value does not follow the pattern set by the other 19 points.

In summary, a *leverage point* is a point whose x -value is distant from the other x -values. A point is a *bad leverage point* if its Y -value does not follow the pattern set by the other data points. In other words, *a bad leverage point is a leverage point which is also an outlier*.

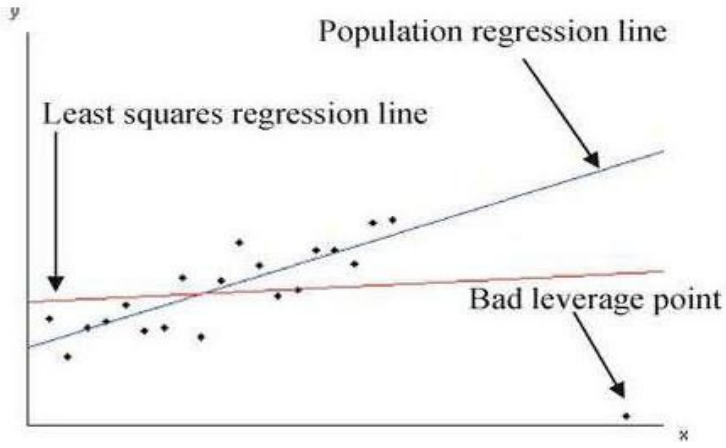

Figure 12. *Bad leverage point*

Returning to Figure 11, the point marked on the plot is said to be a *good leverage point* since its Y -value closely follows the upward trend pattern set by the other 19 points. In other words, *a good leverage point is a leverage point which is NOT also an outlier***.**

Next we investigate what happens when we change the Y -value of a point in Figure 11 which has a central x -value. We move one of these points away from the true population regression line. In particular, we focus on the point with the 11th largest x -value. Moving this point vertically up (so that its x –value stays the same) produces the results shown in Figure 13. Notice how in the least squares regression line has changed relatively little in response to changing the Y -value of centrally located x . This point is said to be an *outlier that is not a leverage point*.

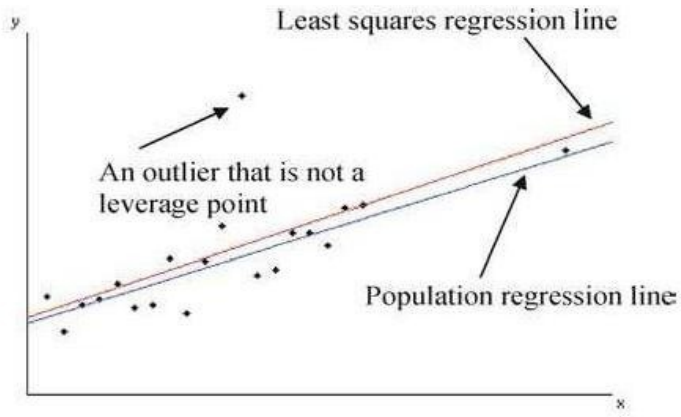

Figure 13 *A plot of* Y *against* x *showing an outlier that is not a leverage point*

Numerical rule that will identify x_i as a leverage point (i.e., a point of high leverage) is based on:

• The distance x_i is away from the bulk of the x's.

• The extent to which the fitted regression line is influenced by the given point.

The second bullet point above deals with the extent to which $\hat{y}_i$ (the predicted value of Y at $¿x_i$) depends on y_i (the actual value of Y at $x=x_i$). We have that

$$\hat{y}_i = b_0 + b_1 x_i$$

where $b_0 = \bar{y} - b_1 \bar{x}$ and $b_1 = \sum_{j=1}^{n} c_j y_j$, $c_j = \frac{x_j - \acute{x}}{S_x^2}$. So that

$$\hat{y}_i = \acute{y} - b_1 \acute{x} + b_1 x_i = \acute{y} + b_1 (x_i - \acute{x})$$

$$x_j - \acute{x} \quad \frac{S_x^2}{} \quad \frac{1}{n}\sum_{j=1}^{n} y_j + \sum_{j=1}^{n}$$

where

$$h_{ij} = \left[\frac{1}{n} + \frac{(x_i - \acute{x})(x_j - \acute{x})}{S_x^2}\right]$$

Notice that

$$x_j - \acute{x} \quad x_i - \acute{x} \quad \frac{1}{n} + (x_i - \acute{x}) = \frac{n}{n} + \quad \sum_{j=1}^{n} h_{ij} = \sum_{j=1}^{n}$$

We can express the predicted value, $\hat{y}_i$ *as*

$$\hat{y}_i = h_{ii}\, y_i + \sum_{j \neq i}^{n} h_{ij}\, y_j \quad (\text{¿})$$

where

$$h_{ii}=\frac{1}{n}+\frac{\left(x_i-\acute{x}\right)^2}{\sum_{j=1}^{n}\left(x_j-\acute{x}\right)^2}$$

The term h_{ii} is commonly called the *leverage* of the ith data point, notice that h_{ii} shows how y_i affects $\hat{y}_i$. For example, if $h_{ii}\cong 1$ then the other h_{ij} terms are close to zero since $\sum_{j=1}^{n} h_{ij}=1$,

$$\hat{y}_i=1\times y_i+other\,terms\cong y_i \ .$$

In this situation, the predicted value $\hat{y}_i$ will be close to the actual value y_i no matter what values of the rest of the data take. Notice also h_{ii} that depends only on the $x's$. Thus a point of high leverage (or a leverage point) can be found by looking at just the values of the $x's$ and not at the values of the $y's.$ We have

$$\frac{1}{n}\sum_{i=1}^{n} h_{ii}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}+\frac{\left(x_i-\acute{x}\right)^2}{\sum_{j=1}^{n}\left(x_j-\acute{x}\right)^2}\right)$$

$$¿\ \frac{1}{n}.\frac{n}{n}+\frac{1}{n}\frac{\left(x_i-\acute{x}\right)^2}{\sum_{j=1}^{n}\left(x_j-\acute{x}\right)^2}=\frac{1}{n}+\frac{1}{n}=\frac{2}{n}$$

Rule for identifying leverage points

A popular rule, which we shall accept, is to classify x_i as a point of high leverage (a leverage point) in a simple linear regression model if

$$h_{ii}>2\times average\left(h_{ii}\right)=2\times\frac{2}{n}=\frac{4}{n}$$

Strategies for dealing with “bad” leverage points

1. *Remove invalid data points*

We ask a question about the validity of the data points corresponding to bad leverage points, Are *these data points unusual or different in some way from the rest of the data?* If so, we remove these points and refit the model without them.

2. *Fit a different regression model*

We ask a question about the validity of the regression model that has been fitted, has *an incorrect model been fitted to the data?* If so, we should consider a different model**.**

"Good" leverage points

While "good" leverage points do not have an adverse effect on the estimated regression coefficients, they do decrease their estimated standard errors as well as increase the value of R^2. Hence, it is important to check extreme leverage points for validity, even when they are so-called "good".

4.4- Standardized Residuals

In this section we will consider standardized residuals and their importance in linear regression diagnostics. First we will show that the $i\,th$ least squares residual has variance given by

$$var\left(\hat{e}_i\right)=\sigma^2\left[1-h_{ii}\right]$$

Proof

Recall from formula (*) that

$$\hat{y}_i=h_{ii}\,y_i+\sum_{j=i}^{n}h_{ij}\,y_j$$

Thus,

$$\hat{e}_i=y_i-\hat{y}_i=y_i-h_{ii}\,y_i-\sum_{j=i}^{n}h_{ij}\,y_j=\left(1-h_{ii}\right)y_i-\sum_{j=i}^{n}h_{ij}\,y_j$$

$$var\left(\hat{e}_i\right)=var\left[\left(1-h_{ii}\right)y_i-\sum_{j\neq i}^{n}h_{ij}\,y_j\right]=\left(1-h_{ii}\right)^2\sigma^2+\sum_{j\neq i}^{n}h_{ij}^2\sigma^2$$

$$¿\sigma^2\left[1-2h_{ii}+h_{ii}^2+\sum_{i\neq j}^{n}h_{ij}^2\right]$$

Next, notice that

$$\sum_{j=1}^{n}h_{ij}^2=\sum_{j=1}^{n}\left[\frac{1}{n}+\frac{(x_i-\acute{x})(x_j-\acute{x})}{S_x^2}\right]^2=n.\frac{1}{n^2}+\frac{2}{n}\sum_{j=1}^{n}\frac{(x_i-\acute{x})(x_j-\acute{x})}{S_x^2}+\sum_{j=1}^{n}\frac{(x_i-\acute{x})^2(x_j-\acute{x})^2}{(S_x^2)^2}=\frac{1}{n}+\frac{(x_i-\acute{x})^2}{(S_x^2)^2}\sum(x_j-\acute{x})^2=\frac{1}{n}+($$

So that,

$$var(\hat{y}_i)=var\left(\sum_{j=1}^{n}h_{ij}y_j\right)=\sum_{j\neq i}^{n}h_{ij}^2\,var(y_j)=\sigma^2\sum_j h_{ji}^2=\sigma^2 h_{ii}$$

Thus, if $h_{ii}\cong$ 1, so that the ith point is a leverage point, then the corresponding residual, $\hat{e}_i$, has small variance (since $1-h_{ii}\cong 0$). This seems reasonable when one considers that if $h_{ii}\cong 1$ then $\hat{y}_i\cong y_i$ so that $\hat{e}_i$ will always be small (and so it does not vary much). We have also shown that $Var\,\hat{y}_i=\sigma^2 h_{ii}$. This again seems reasonable when we consider the fact that when $h_{ii}\cong 1$ then $\hat{y}_i\cong y_i$. In this case, $var(\hat{y}_i)=\sigma^2 h_{ii}\cong\sigma^2=var(y_i)$ The problem of the residuals having different variances can be overcome by standardizing each residual by dividing it by an estimate of its standard deviation. Thus, $the\,i\,th\,standardized\,residual$, r_i is given by

$$r_i=\frac{\hat{e}_i}{\hat{\sigma}_{r_i}}=\frac{\hat{e}_i}{s\sqrt{1-h_{ii}}}$$

where $s=\sqrt{\frac{1}{n-2}\sum_{j=1}^{n}\hat{e}_j^2}$ is the estimate of σ obtained from the model.

A crucial assumption in any regression analysis is that the errors have constant variance. In literature it is recommended that an effective plot to diagnose non constant error variance is a plot of $|\text{Residuals}|^{0.5}$ against x or a plot of $|\text{Standardized Residuals}|^{0.5}$ against x. (or against the fitted values.

When points of *high leverage* exist, it is informative to look at plots of *standardized residuals***.** The advantage of standardized residuals is that they immediately tell us how many estimated standard deviations any point is away from the fitted regression model. For example, suppose that the 6th point has a standardized residual of 4.3, then this means that the 6th point is an estimated 4.3 standard deviations away from the fitted regression line. If the errors are normally distributed, then observing a point 4.3 standard deviations away from the fitted regression line is highly unusual. Such a point would commonly be referred to as an outlier and as such it should be investigated. We shall follow the common practice of labeling points as *outliers* in small- to moderate-size data sets if the standardized residual for the point falls outside the interval from *–2 to 2*. In very large data sets, we shall change this rule to *–4 to 4*. Identification and examination of any outliers is a key part of regression analysis. Recall that a bad leverage point is a leverage point which is also an outlier. Thus, a bad leverage point is a leverage point whose standardized residual falls outside the interval from –2 to 2 for small to moderate size data sets. On the other hand, a good leverage point is a leverage point whose standardized residual falls inside the interval from –2 to 2.

4.5 - Assessing the Influence of Certain Cases

One or more cases can strongly control or influence the least squares fit of a regression model. In this section we look at summary statistics that measure the influence of a single case on the least squares fit of a regression model.

Cook proposed measure of the influence of individual cases which is called Cook`s distance and it is given by:

$$D_i = \frac{r_i^2}{2} \frac{h_{ii}}{1-h_{ii}}$$

where r_i is the ith standardized residual and h_{ii} is the ith leverage value, in literature it is recommended to use $\frac{4}{n-2}$ as "a rough cut-off for noteworthy values of D_i for simple linear regression. (Meaning if Cook`s distance of a point is bigger than $\frac{4}{n-2}$, it has significant influence on fitted regression model).

Recommendations for Handling Outliers and Leverage Points

We conclude this section with some general advice about how to deal with outliers and leverage points.

• Points should not be routinely deleted from an analysis just because they do not fit the model. Outliers and bad leverage points are signals, flagging potential problems with the model.

• Outliers often point out an important feature of the problem not considered before. They may point to an alternative model in which the points are not an outlier. In this case it is then worth considering fitting an alternative model.

4.6 - Normality of the Errors

The assumption of normal errors is needed in small samples for the validity of *t*-distribution based hypothesis tests and confidence intervals and for all sample sizes for prediction intervals. This assumption is generally checked by looking at the distribution of the residuals or standardized residuals.

A common way to assess normality of the errors is to look at what is commonly referred to as a *normal probability plot* or a *normal Q–Q plot* of the standardized residuals. A normal probability plot of the standardized residuals is obtained by plotting the ordered standardized residuals on the vertical axis against the expected order statistics from a standard normal distribution on the horizontal axes. If the resulting plot produces points "close" to a straight line then the data are said to be consistent with that from a normal distribution. On the other hand, departures from linearity provide evidence of non-normality.

Example 6: Recall the example from Chapter 2 on the timing of production runs for which we fit a straight-line regression model to run time from run size. The data are given in Table (along with values of leverages, residuals and standardized residuals) and are plotted in Figure 14.

Table Regression diagnostics for the model in Figure 14

Case	*Run time*	*Run size*	*Leverage*	*Residuals*	*Std. residuals*
1	195	175	0.053	-6.892	-0.132
2	215	189	0.060	-49.226	-0.948
3	243	344	0.145	26.907	0.543
4	162	88	0.141	-0.942	-0.019
5	185	114	0.066	-39.726	-0.767
6	231	338	0.098	54.707	1.075
7	234	271	0.108	-20.743	-0.410
8	166	173	0.124	72.791	1.452
9	253	284	0.197	-61.260	-1.276
10	196	277	0.052	92.291	1.769
11	220	337	0.068	84.691	1.638
12	168	58	0.116	-47.842	-0.950
13	207	146	0.051	-69.693	-1.336
14	225	277	0.080	10.607	0.206
15	169	123	0.112	14.341	0.284
16	215	227	0.060	-11.226	-0.216
17	147	63	0.222	16.308	0.345
18	230	337	0.094	56.524	1.109
19	208	146	0.052	-72.509	-1.390
20	172	68	0.101	-49.109	-0.967

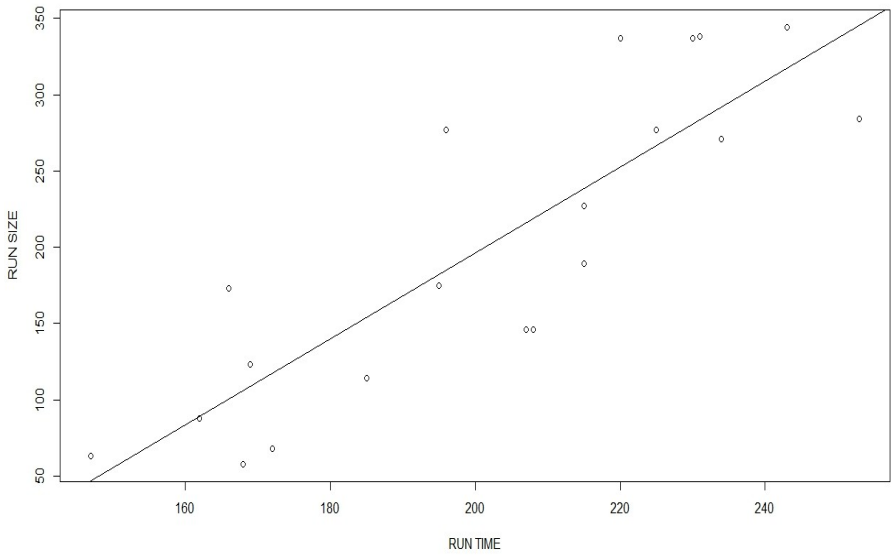

Figure 14. *A plot of the production data with the least squares line included*

We begin by considering the simple regression model where Y = run size and x = run time. Regression output from R is given below.


```
Residuals:
    Min      1Q  Median      3Q     Max
-72.509 -48.159  -3.917  33.857  92.291

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -367.3606    82.4122  -4.458 0.000304 ***
run.time       2.8167     0.4035   6.980 1.61e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53.57 on 18 degrees of freedom
Multiple R-squared: 0.7302,     Adjusted R-squared: 0.7152
F-statistic: 48.72 on 1 and 18 DF,  p-value: 1.615e-06
```

Looking at the given output in R, we could conclude that simple regression output is valid for given data – coefficient of slope of regression line is significantly different from zero (based on results of t test) and 73.02% of variance of variable run size is explained by the regression model (value of coefficient of determination). But, before we make such a conclusion we should look at the diagnostic plots.

Also, we should check leverage values – as we said before, point is classified as high leverage point if its leverage value is bigger that $\frac{4}{n} = 0.2$. Only the 17^{th} point has leverage value bigger than this cut-off point. On scatter plot, we can see that y-value of this point is not distant from other points, meaning that this is good leverage point (also its standardized residual falls inside the interval from *–2 to 2*).

Figure 15 provides diagnostic plots produced by R: on the top left is the plot for randomness of residuals, top right is normal Q-Q plot, down left plot for variance of residuals and on down right is plot for Cook's distance. Variance of the error term appears to be constant. These plots suggest that linear model is appropriate for our production data: residuals seem distributed randomly, having approximately normal distribution with constant variance. None of the points have Cook's distance greater than cut-off point 4/18= 0.22.

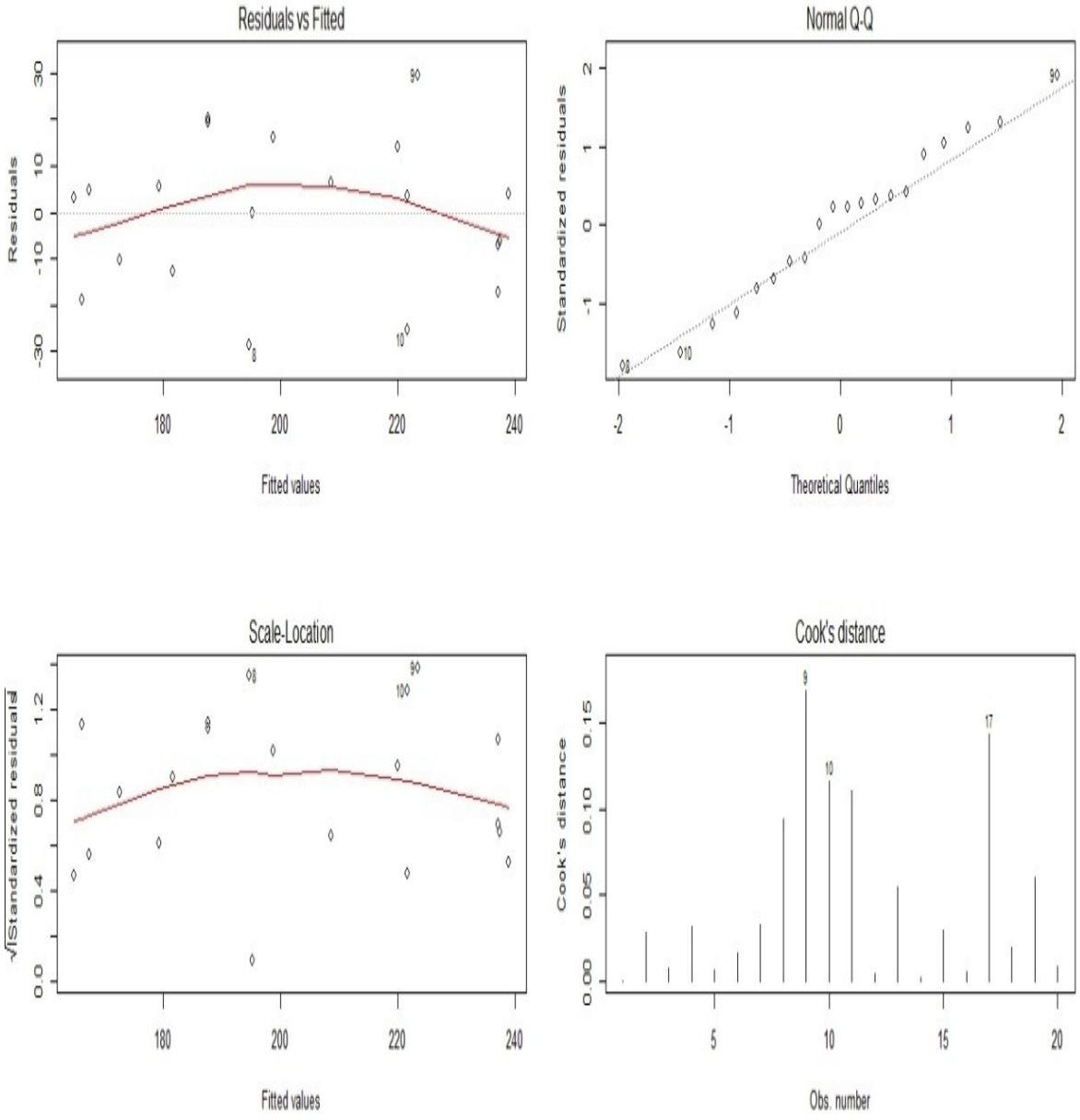

Figure 15. *Diagnostic plots*

Conclusion

This master thesis deals with an important area of statistics - namely correlation and regression. These topics could be also very useful in teaching statistic at the university. We have chosen and explained some aspects of this topic.

First we discussed about correlation as measure of association between variables. We introduced Pearson`s correlation coefficient as measure of linear relationship and Spearman`s correlation coefficient as measure of monotonic relationship. After, we emphasized the importance of constructing the scatter plot of observed variables to reveal the nature and strength of correlation.

We discussed diagnostic plots as methods for checking assumptions of simple linear regression. We presented example based on Anscombe's four data sets which illustrates the point that looking only at the numerical regression output may lead to very misleading conclusions about the data and to accept the wrong model.

We introduced regression models to describe causal relationship between two variables. Further, we dealt especially with simple linear regression. For estimating unknown parameters of linear regression we used method of least squares. We proved properties of unbiasedness and consistency of these parameters, as well as some other properties concerning distribution and independence of statistics we have used.

We talked about testing the validity of the model using Student's t- test and about the significance of coefficient of determination. Regression models are primarily used for prediction, so we considered prediction intervals for values of dependent variables.

Some of the examples given here are only for illustrations of definitions and procedures, some of them, with larger sets of data were solved using statistical package R.

A part of the references given here, I have used teaching materials from my Arabic courses at the University of Libya (Al Mergeb, Zliten).

For future research, it would be interesting to consider other types of regression – for example, when dependent variable is binary (logistic regression) or when we have more than one dependent variable (multivariate regression). Also, we could explore how is regression used to solve problems in biology, medicine and other areas.

References

Cohen,Y. and Cohen, J.Y. (2008), *Statistics and Data with R: An Applied Approach Through Examples.* Wiley, New York.

Jevremović,V and Mališić, J. (2002). *Statističke metode u meteorologiji i inženjerstvu.* Savezni hidrometeorološki zavod, Beograd.

Larson , R. J. and Marx, M.L. (2005). *An introduction to Mathematical statistics and its Applications*. Prentice Hall, New Jersey.

Lindley ,D.V. (1987). *Regression and correlation analysis*. The New Palgrave: A Dictionary of Economics.

Milton, J. S, Corbet, J.J. and McTeer, M.X. (1997). *An introduction to Statistics*. McGraw-Hill Higher Education, Burr Ridge.

Shanker, R. G. (2006) . *Numerical Analysis*, New Age International Ltd., Publishers, New Delhi.

Sheather, S.J. (2009). *A Modern Approach to Regression with R.* Springer, New York.