REGRESSION AND LEAST SEQURE METHOD AND ITS PROPERTIES

Karima Ibrahim Soufya and Marwa Elhadi Assari

Statistics Department , Alasmaryia Islamic University , Zliten , Libya
Fooqe2007@gmail.com

ABSTRACT

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.

1. INTRODUCTION

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

Plots are shown to be an important tool for both building regression models and assessing their validity. We shall see that deciding what to plot and how each plot should be interpreted will be a major challenge. In order to overcome this challenge we shall need to understand the mathematical properties of the fitted regression models and associated diagnostic procedures. In particular, we shall carefully study the properties of residuals in order to understand when patterns in residual plots provide direct information about model misspecification and when they do not

This paper is divided into three parts . The first part contains an introduction about the work. In the second part we introduce Pearson's correlation coefficient, properties of regression coefficients and types of regression. In the third part we present our conclusions.

2. REGRESSION ANALYSIS

In this part we discuss regression which measures the nature and extent of correlation.

2.1 Correlation

Pearson's correlation coefficient is one of a number of measures of correlation or association. It determines the degree to which a linear relationship exists between two variables.

The statistic computed for the Pearson correlation coefficient is represented by the letter r. r is an estimate of ρ, which is the correlation between the two variables in the underlying population. r can assume any value within the range of -1 to +1. The absolute value of r (i.e., |r|)) indicates 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 either + 1 or - 1). The closer the absolute value of r is to 1, the more accurately a researcher will be able to predict a subject's score on one variable from the subject's score on the other variable. The closer the absolute value of r is to 0, the weaker the linear relationship is between the two variables.

The sign of r indicates the nature or direction of the linear relationship which exists between the two variables. A positive sign indicates a direct

linear relationship, whereas a negative sign indicates an indirect (or inverse) 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). When there is a direct relationship, subjects who have a high score on one variable will have a high score on the other variable, and subjects who have a low score on one variable will have a low score on the other variable.

An indirect/inverse 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). When there is an indirect linear relationship, subjects who have a high score on one variable will have a low score on the other variable, and vice versa.

The use of the Pearson's correlation coefficient assumes that a linear function best describes the relationship between the two variables. If, however, the relationship between the variables is better described by a curvilinear function, the value of r computed for a set of data may not indicate the actual extent of the relationship between the variables .

Calculation of Pearson's correlation coefficient r

Let $(x_1, y_1),(x_2, y_2)\ldots(x_n, y_n)$ be n paired observations, then Pearson's correlation coefficient is equal to

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

or simply

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

Where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ are sample means and $s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ are standard sample deviations.

If we use

$\grave{x_i} = x_i - \bar{x}$,

$$\grave{y_i} = y_i - \bar{y}$$

, then we get simplified for Pearson's correlation coefficient

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

2.2 Scatter Diagram

Let us have pairs of values $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$. 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 paper.

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 that exists between two variables

(e.g., linear, quadratic, exponential, etc) .

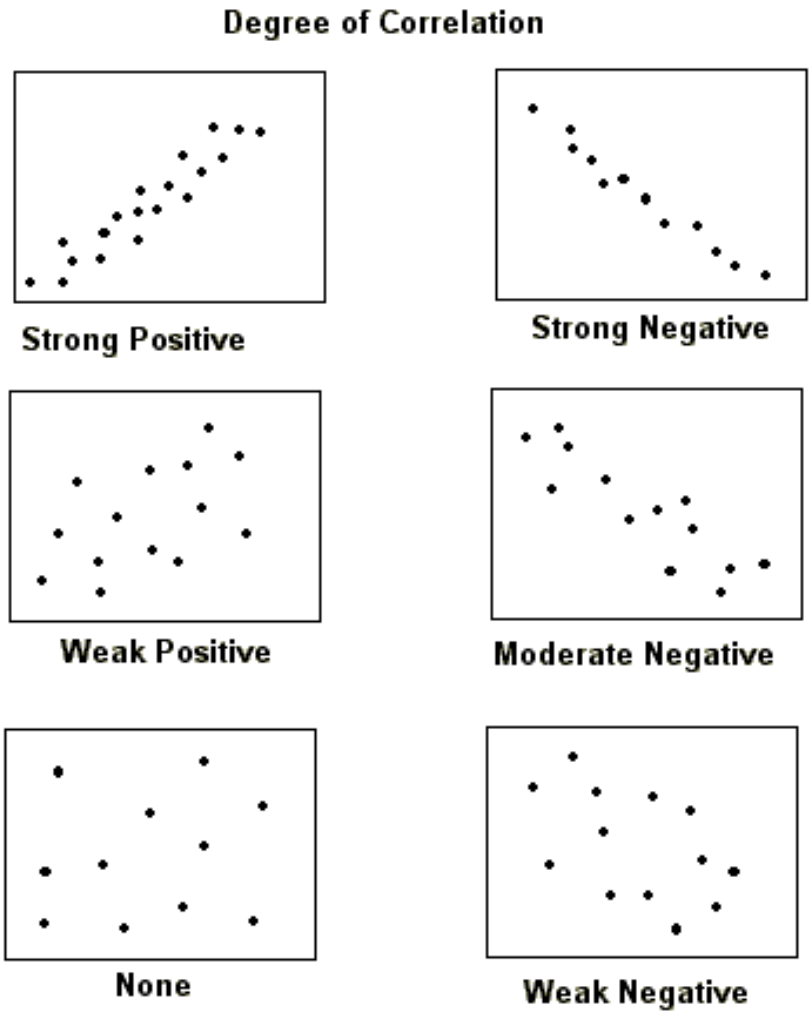

Figure 1. The degrees of correlation

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

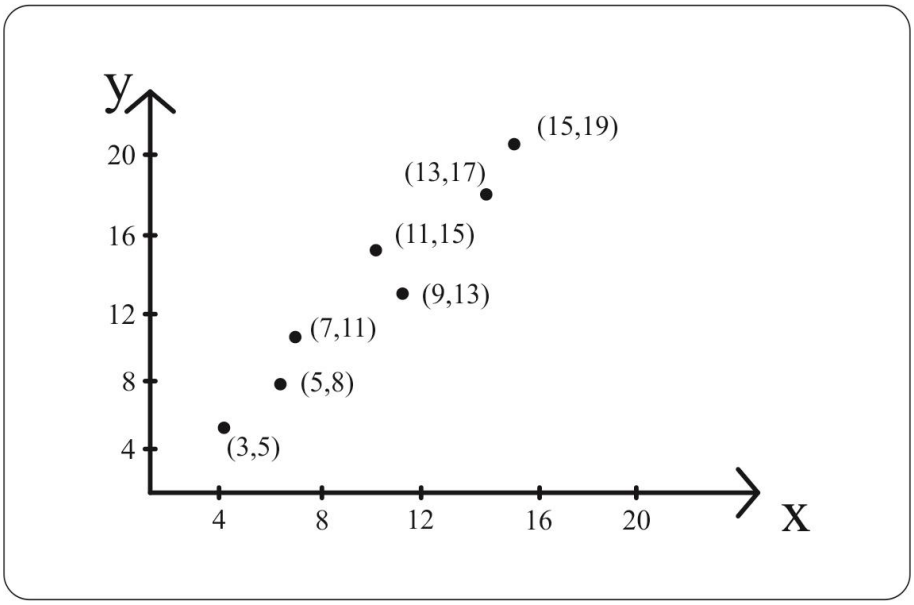

Figure 2. Scatter plot for given data

First we calculate sample means, sample standard deviations and then Pearson's correlation coefficient is equal to

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

$$s_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{7}(y_i - \bar{y})^2} = 4.59$$

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

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 correlation does not imply causation. Consequently, if

there is a strong correlation between two variables , a researcher is not justified in concluding that one variable causes the other variable. Although it is possible that when a strong correlation exists one variable may, in fact, cause the other variable, the information employed in computing the Pearson's correlation coefficient does not allow a researcher to draw such a conclusion. This is the case, since extraneous variables which have not been taken into account by the researcher can be responsible for the observed correlation between the two variables.

2. 3 Regression

Regression is typically used to model the relationship between dependent variable Y and one or more independent variables X, so that given the specific value of X, that is X=x, we can predict the value of Y. Mathematically, the regression of a random variable Y on a random variable X is

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

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

2.3.1 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).

2.3.2 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*.

2.3.3 Types of Regression

If there are only two variables under consideration, then the regression is called *simple regression*. For example, in the case of a study of regression between heights and age for a group of persons the relationship is linear. If there are more than two variables under considerations 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. If the relationship between X and Y is non linear, then the regression is *curvilinear regression.* For example, volume of oil tanker is a cubic function of its length. In some cases polynomials are selected to predict or estimate; which is called *polynomial regression*.

2.3.4 Linear Regression Equation of Y on X

Data are collected in pairs (x_1, y_1) , $(x_2, y_2) \ldots (x_n, y_n)$, where x_1 denotes the first value of the so-called X -variable and y_1 denotes the first value of the so-called Y -variable. The X variable is called the *explanatory* or *predictor variable* , while the Y -variable is called the *response variable* or the *dependent variable***.**

The regression of Y on X is linear if

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

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 and is such that E(e | X) = 0.

The random error term is there since there will almost certainly be some variation in Y due strictly to random phenomenon that cannot be predicted or explained. In other words, all unexplained variation is called *random error* . Thus, the random error term does not depend on, nor does it contain any information about Y (otherwise it would be a systematic error).

2.3.5 Assumptions necessary about the regression model

Throughout this section 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, $e \sim N(0, \sigma^2)$

Methods for checking these four assumptions will be considered in the third chapter. In addition , since the regression model is conditional on x we can

assume that the values of the predictor variable, $(x_1, x_2, \dots, x_n)$ are known fixed constants

2.4 Estimating the population slope and intercept

Suppose for example that variable X represents height and variable Y weight of a person. For a line 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 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 $\hat{y}_i = b_0 + b_1 x_i$ is as "close" as possible to Y_i.

Here the notation $\hat{y}_i$ is used to denote the value of the line of best fit in order to distinguish it from the observed values of Y that is y_i. We shall refer to $\hat{y}_i$ as the i th predicted value or the fitted value of Y_i. For estimating these unknown parameters we will use the method of least squares.

3. 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 in the curve for which the sum of the squares of the vertical distances of the point (*x*, *y*) from the line is minimal.

The Least squares principle for the simple linear regression model is to find the estimators b_0 and b_1 such that the sum of the squared distances from

actual response y_i and predicted response $\hat{y}_i = \beta_o + \beta_1 x_i$ reaches the minimum among all possible choices of regression coefficients b_o and b_1, we are searching for values which minimize the sum:

$$S = \sum_{i=1}^{n} [y_i - (\beta_o + \beta_1 x_i)]^2$$

The motivation behind the least squares method is to find parameter estimates by choosing the regression line that is the most "closest" line to all data points (x_i, y_i), $i = 1, n$ (Figure 3).

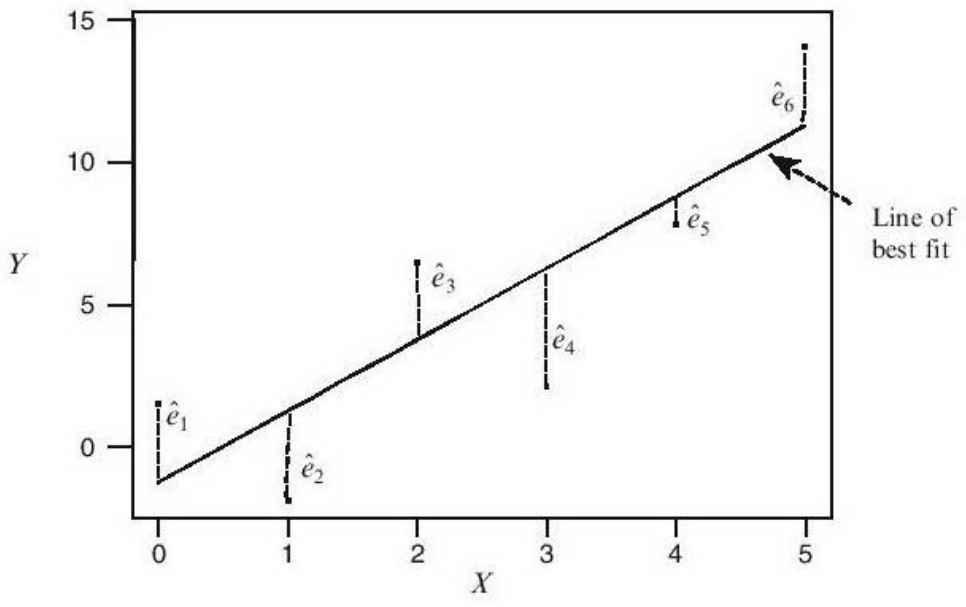

Figure 3. Vertical distances of points from regression line

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

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

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

$$\hat{y}_i = \beta_o^* + \beta_1(x_i - \bar{x}) + \varepsilon_i ,$$

where

$$\beta_o = \beta_o^* - \beta_1 \bar{x} .$$

Now sum of squared distances equals to

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

and we need to solve the following system

Taking the partial derivatives with respect to β_0^* and β_1 we have:

$$\sum_{i=1}^{n} [y_i - (\beta_o^* + \beta_1(x_i - \bar{x}))] = 0$$

$$\sum_{i=1}^{n} [y_i - (\beta_o^* + \beta_1(x_i - \bar{x})](x_i - \bar{x}) = 0$$

Note that

$$\sum_{i=1}^{n} y_i = \; n\,\beta_0^* + \sum_{i=1}^{n} \beta_1\,(x_i - \bar{x}) = \quad n\,\beta_0^*$$

Therefore, we have

$$b_0^* = \hat{\beta}_0^* \quad = \frac{1}{n}\sum_{i=1}^{n} y_i = \; \bar{y}.$$

Substituting b_0^* by $\bar{y}$ we obtain

$$\sum_{i=1}^{n} [y_i - (\bar{y} + \; \beta_1(x_i - \bar{x}))](x_i - \bar{x}) = 0$$

Now it is easy to see

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

and

$b_0 = \; b_o^*$ - $\; b_1\bar{x} \; = \bar{y} - b_1\bar{x}$

The fitted value of the simple regression is defined as $\widehat{y_i} = b_o + b_1 x_i$.

3.1 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.

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 4.

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

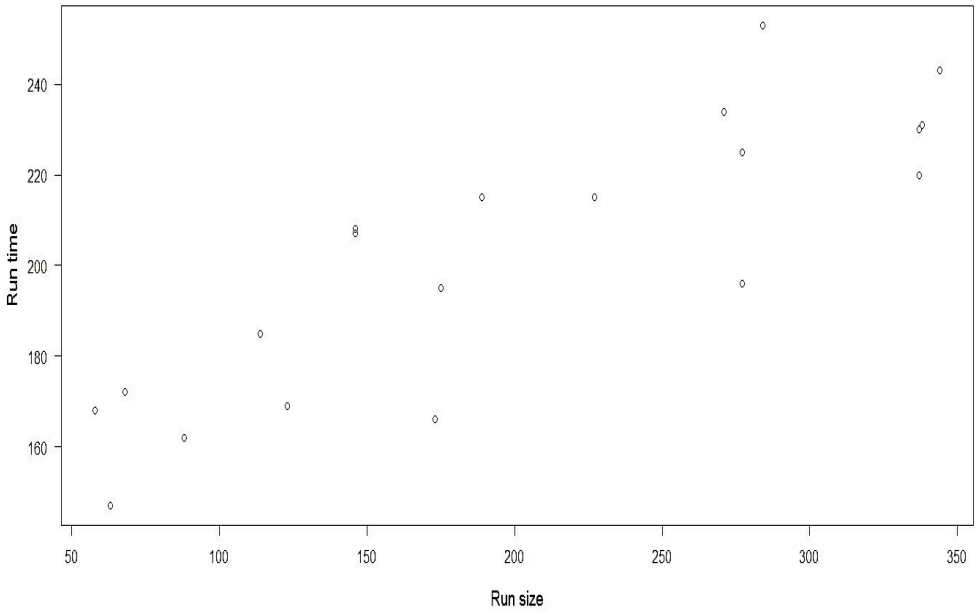

Figure 4. A scatter plot of the production data

Now we shall give some properties of estimators in simple linear regression model

3.2 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 , … , n as constants ,while Y is a random variable .

$$E(b_1) = E\left(\frac{Sxy}{Sxx}\right) = \frac{1}{Sxx} E \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{X})$$

Using the fact that

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

We get:

$$E(b_1) = \frac{1}{Sxx} \cdot \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})Ey_i = \frac{1}{Sxx} \cdot \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(\beta_o + \beta_1 x_i) = \frac{1}{Sxx} \cdot \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{Sxx} \cdot \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\beta_1(x_i - \bar{x}) = \frac{1}{Sxx} \cdot \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\beta_1 = \frac{Sxx}{Sxx}\beta_1 = \beta_1 .$$

Theorem 2: Variance of the estimator of the slope is

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

Proof:

$$Var(b_1) = Var\left(\frac{sxy}{sxx}\right) = \frac{1}{S_{xx}^2}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{y})\right)(x_i - \bar{x}) = \frac{1}{S_{xx}^2}\,Var\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\,(x_i - \bar{x})\right) =$$

$$= \frac{1}{S_{xx}^2} \cdot \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\,Var(Y_i) = \frac{1}{S_{xx}^2} \cdot \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2 = \frac{\sigma^2}{nS_{xx}}.$$

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

Proof:

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

$$= \frac{1}{nS_{xx}} Cov\left(\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \bar{y}), \bar{y}\right) =$$

$$= \frac{1}{nS_{xx}} Cov\left(\sum_{i=1}^{n} (x_i - \bar{x}) y_{i,} \bar{y}\right) = \frac{1}{n^2 S_{xx}} Cov\left(\sum_{i=1}^{n} (x_i - \overline{x}) y_{i,}, \sum_{i=1}^{n} y_i\right) =$$

$$= \frac{1}{n^2 S_{xx}} Cov \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - \bar{x})\; Cov\,(y_i, y_j)$$

Note since that, $Ee_i = 0$ and ei`s are independent, we can write

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

Thus, we conclude that

$$Cov\,(b_1, \bar{y}) = \frac{1}{n^2 S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \sigma^2 = 0$$

Recall that zero correlation is equivalent to the independence between two normal variables. Thus, we conclude that b_1 and $\bar{y}$ are independent.

3.3 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 .

$Eb_o = E(Y_i - b_1\bar{x}) = (\frac{1}{n}\sum_{i=1}^{n} E\,Y_i) - \bar{x}Eb_1 = \frac{1}{n}\sum_{i=1}^{n}\beta_o + \beta_1\frac{1}{n}\sum_{i=1}^{n} x_i - \beta_1\bar{x} = \beta_o.$

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

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

Proof:

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

3.4 Estimator of variance of random error and its property

As we said before, in linear regression model $Y_i = \beta_0 + \beta_1 x_i + e_i$ random errors e_i are normally distributed with mean equal to 0 and variance σ^2 .We wish to estimate variance $\sigma^2 = Var(e)$.

We can estimate these errors by replacing β_0 and β_1 by their respective least squares estimates b_0 and b_1 giving the residuals

$$\hat{e}_i = Y_i - (b_0 + b_1 x_i)$$

These residuals can be used to estimate σ^2. We will estimate σ^2 with $\frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2$ and we will show that this estimator is asymptotically unbiased estimator of σ^2.

$$E\,\widehat{\sigma^2} = E\left(\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)$$

$$= \frac{1}{n-2}\,E\left(\sum_{i=1}^{n}(b_1 x_i + b_0 + e_i - \beta_1 x_i - \beta_0)^2\right) =$$

$$= \frac{1}{n-2}E\left(\sum_{i=1}^{n}\left((b_1-\beta_1)x_i + (b_0-\beta_0) + e_i\right)^2\right) =$$

$$= \frac{1}{n-2}\left(\sum_{i=1}^{n} x_i^2\frac{\sigma^2}{Sx^2} + \frac{\sigma^2}{n} + \sigma^2\,\frac{\overline{x_n^2}}{Sx^2} + \sigma^2 - 2b_0 b_1 x_i + 2E\beta_0\beta_1 - 2x_i E\beta_1 e_i - 2E\beta_0 e_i\right)$$

We have that

$$E\beta_1 e_i = \frac{1}{Sx^2}E\left(\sum_{j=1}^{n}(x_j - \bar{x}_n)(y_j - \bar{y}_n)e_i\right) = 0,$$

$$E\beta_0 e_i = E(\bar{y}_n - \beta_1\bar{x}_n)e_i = E(b_0 + b_1\bar{x}_n + \bar{e}_n - \beta_1\bar{x}_n)e_i = \frac{\sigma^2}{n}$$

$$E\,\beta_0\beta_1 = E\,\beta_1.(\bar{y}_n - \beta_1.\bar{x}_n) = E\beta_1(b_0 + b_1\bar{x}_n + \bar{e}_n - \beta_1.\bar{x}_n)$$

$$= b_0 b_1 - \bar{x}_n . \frac{\sigma^2}{Sx^2} + \frac{1}{Sx^2} \sum_{i=1}^{n} (x_i - x_n) (E e_i \bar{e}_n - E \bar{e}_n^2) = b_0 b_1 - \bar{x}_n . \frac{\sigma^2}{Sx^2} \quad .$$

When we include these results in the expression for expectation of estimator of variance, we get

$$E\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} x_i^2 . \frac{\sigma^2}{Sxx^2} + \frac{\sigma^2}{n-2} + \frac{n}{n-2} \frac{\sigma^2 . \bar{x}_n^2}{Sxx^2} + \frac{n\sigma^2}{n-2} - \frac{2b_0 b_1}{n-2} \sum_{i=1}^{n} x_i + \frac{2}{n-2} b_0 b_1 \sum_{i=1}^{n} x_i$$

$$= \frac{n}{n-2} \sigma^2 \xrightarrow[n \to \infty]{} \sigma^2$$

4. CONCLUSION

The correlation coefficient is a measure of linear association between two variables ,values of the correlation coefficient are always between (-1 and +1). Regression analysis is widely used for prediction and forecasting . Where its use has substantial overlap with the field of machine learning . Regression analysis is also used to understand which among the independent variables are related to the dependent variable ,and to explore the forms of these relationships. There are many examples of use of regression analysis could be used for a predicting yield of a crop , for different doses of a fertilizer ,and regression analysis also could be used to estimate the height of a person at a given age ,by finding the regression of height on age .

REFERENCES

[1] Cohen,Y. and Cohen, J. Y.(2008), Statistics and Data with R: An Applied Approach Through Examples. Wiley.

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

[3] Larson , R. J. and M. L. Marx (2005). An introduction to Mathematical statistics and its Applications. Prentice Hall.

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

[5] Milton, J. S. and J. J. Corbet, P. M. McTeer (1997). An introduction to Statistics. McGraw-Hill Higher Education

[6] Shanker, R. G. (2006) . Numerical Analysis, New Age International (P) Ltd, Publishers.

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