Lecture (3): Simple Linear Regression

Concept of Regression Analysis

Regression analysis is defined as a statistical method that studies the relationship between one dependent variable and one or more independent variables, so that this relationship explains the reasons for the changes that occur in the dependent variable and how the independent variables affect these changes, and how these changes are explained by those independent variables, and these changes are done through regression analysis, as regression analysis is based on describing the relationship between the variables in the form of an equation.

Uses of Regression Analysis (Its Objectives)

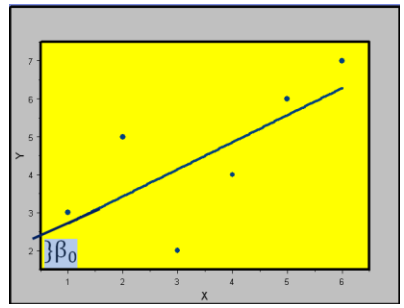
1. **Data Description:** This is done by finding the regression equation that describes that data
2. **Parameters Estimation:** The parameter for the population gives the importance of each independent variable in its effect on y and gives the direction and strength of the effect
3. **Predection:** It is not only the unstudied (future) prediction but it is knowing something unstudied taking into account the data space
4. **Control:** Where y can be controlled by controlling the independent variables

Simple Linear Regression

If we wanted to study the change in a dependent variable and the change in one independent variable, then for every value of X we would have a value of Y. If we took these points, one of which resulted from X and the other from Y, and plotted them in a scatter diagram, that is, we would take X and its corresponding value in Y.

As we notice, there is a line that can pass through these points and has features. It is called the regression line and its equation is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$



Explanation of the regression line model

شرح نموذج خط الأنحدار

y_i: Dependent Variable and also the response value (to the value of X)

β_0 β_1: Regression Parameters, These parameters are fixed values.

β_0 : It is the distance between the intersection of the regression line with the Y axis from the origin. If it is positive, this means that the regression line intersects Y above the origin. If it is negative, the line will be below the origin point. If it is equal to zero, the line passes through the origin point.

β_0 تمثل كذلك مقدار الأستجابة للمتغير المعتمد عندما قيمة X تساوي صفر.

β_1 : : تمثل مقدار التغير (الزيادة أوالنقصان) الحاصل في Y نتيجة زيادة وحدة واحدة في X.

وكذلك فإن β_1 هي عبارة عن الميل $Slop = \frac{\Delta y}{\Delta x} = \frac{\Delta y}{1} = \Delta y = \beta_1$

كذلك β_1 هو عبارة عن ظل الزاوية المحصورة بين خط الأنحدار والمحور X

β_0: also represents the amount of response to the dependent variable when the value of X equals zero.

β_1: represents the amount of change (increase or decrease) in Y resulting from a one-unit increase in X.

Also, β_1 is the slope $Slop = \frac{\Delta y}{\Delta x} = \frac{\Delta y}{1} = \Delta y = \beta_1$

Also, β_1 is the tangent of the angle between the regression line and the X-axis

$$tan\theta = \frac{\beta_1}{1} = \beta_1$$

β_1 is called the regression coefficient of Y on X.

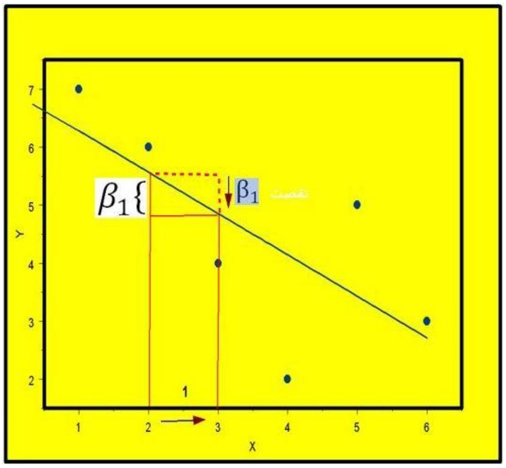


Figure Shows the Negative Relation

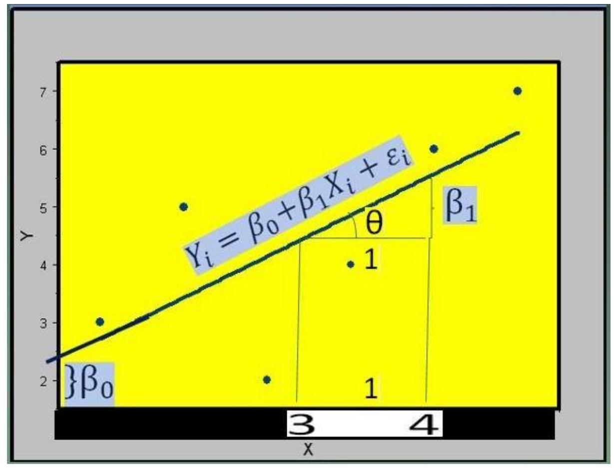


Figure Shows the Positive Relation

The sign β_1 indicates the type of relationship. If it is positive, the relationship is positive, and if it is negative, the relationship is negative. The change in Y can be taken as due to X, but it cannot be said that if Y changes, X changes because X is independent and Y is dependent.

ε_i It is the random error, which is the deviation of the expected values from the actual observed values, as for every X we have two actual observed values and an estimated value that lies on the regression line, and the difference between them represents the error.

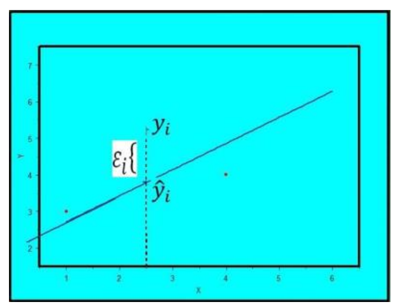
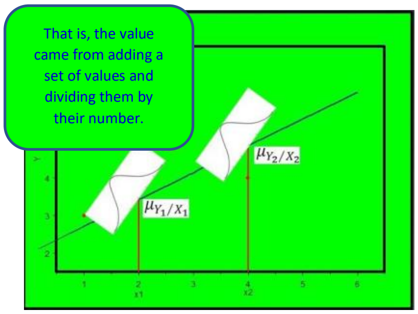


Figure showing the error term

Assumptions of Analysis

1) Y is a random variable (i.e. it has a distribution and is subject to probability) distributed normally with a mean μ and a variance of $\sigma^2_{Y/X}$, i.e. $Y_i \sim N(\mu, \sigma^2_{Y/X})$ and its values are independent of each other (for example, we say y_1 from a specific x, but y_2 is independent of y_1 but has a relationship with x).

The average of the Y values is μ, for example $Y_1 = \mu_{Y_1/X_1}$ and $Y_2 = \mu_{Y_2/X_2}$, and each average is a straight-line function because it results from a straight-line equation and it lies on that straight line, which means that x_i is the arithmetic mean of all the values divided by their number.



2) **Homoscedasticity:** This assumption means that the variances of all observations are homogeneous. It is worth noting that the arithmetic mean is not homogeneous because if it were homogeneous, the regression line would be parallel to the X-axis.

$$V(Y_1) = V(Y_2) = \cdots = V(Y_n) = \sigma^2_{Y/X}$$

3) The random error is normally distributed with a mean equal to zero and a variance equal to σ^2_ε, i.e. $\varepsilon \sim N(0, \sigma^2_\varepsilon)$, and when we write the regression line equation and extract the estimated equation, then $E(\varepsilon_i) = 0$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

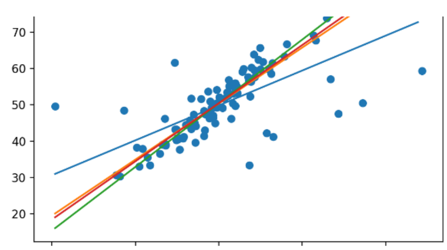
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

4) $\varepsilon_i, \varepsilon_j$ are Uncorrelated to specific time periods, i.e. $cov(\varepsilon_i, \varepsilon_j) = 0 \qquad i \neq j$

Estimation of Regression Parameters by Least Squares Method

This method considers the variable X as a fixed variable and the variable Y as random. The least squares method is used to deal with this case, on the basis of which the values of the two parameters β_0 and β_1 are estimated. The best regression line that represents the relationship between the two variables is the line that passes through these points and gives the sum of squares of the distance from these points as small as possible.

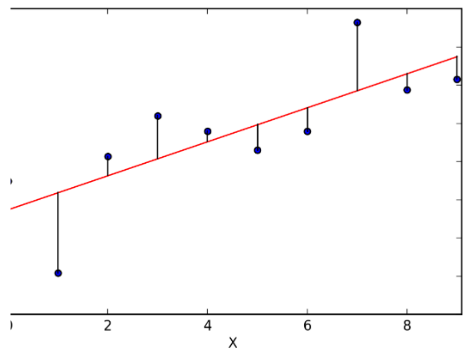
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$
$$\sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = K$$



To make $\sum_{i=1}^{n} {\varepsilon_i}^2$ as small as possible, we must take the partial differential of the quantity K once with respect to β_0 and once again with respect to β_1 and set the result equal to zero, so we get:

$$\frac{\partial K}{\partial \beta_0} = -2\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial K}{\partial \beta_1} = -2\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$$



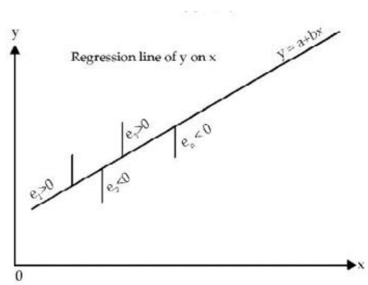
Now $\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

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

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

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

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



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

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

$$\left.\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i &= \sum y_i \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 &= \sum x_i y_i \end{aligned}\right\} \quad Normal\ Equa$$

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

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

We have $\hat{\beta}_0 = \frac{\sum y_i}{n} - \hat{\beta}_1 \frac{\sum x_i}{n} = \frac{\sum y_i \hat{\beta}_1 \sum x}{n}$

We substitute this formula into the second natural equation, as follows:

$$\left(\frac{\sum y_i \hat{\beta}_1 \sum x_i}{n}\right)\sum X_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i$$

$$\frac{(\sum y_i)(\sum x_i) - \hat{\beta}_1 (\sum x_i)^2}{n} + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i$$

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

$$-\hat{\beta}_1 [(\sum x_i)^2 - n \sum x_i^2] = n \sum x_i y_i - (\sum y_i)(\sum x_i)$$

$$\hat{\beta}_1 [n \sum x_i^2 - (\sum x_i)^2] = n \sum x_i y_i - (\sum y_i)(\sum x_i)$$

$$\hat{\beta}_1 = \frac{n \sum X_i y_i - (\sum y_i)(\sum x_i)}{[n \sum x_i^2 - (\sum x_i)^2]} \qquad \div \text{n}$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{(\sum y_i)(\sum x_i)}{n}}{\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]} = \frac{SCP}{SS_x} = \frac{S_{Xy}}{S_{xx}}$$

Where $S_{xy} = \sum(x_i - \bar{x})\, S(y_i - \bar{y})$

$$= \sum x_i y_i - \frac{(\sum y_i)(\sum x_i)}{n}$$
$$= \sum x_i y_i - n\bar{x}\bar{y}$$
$$= \sum (x_i - \bar{x})\, y_i$$
$$= \sum (y_i - \bar{y}) x_i$$

$$S_{xx} = \sum (x_i - \bar{x})^2$$
$$= \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$
$$= \sum x_i^2 - n(\bar{x})^2$$
$$= \sum (x_i - \bar{x})\, x_i$$

Example:

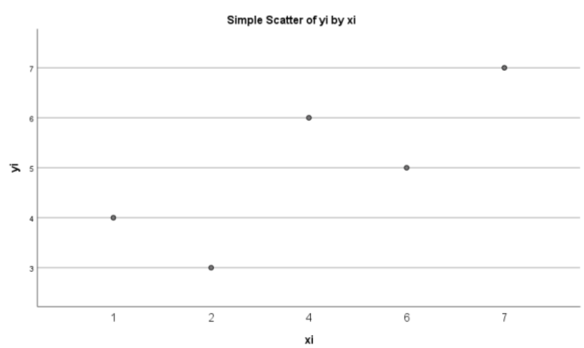
x_i	y_i	x_iy_i	x_i^2	$\hat{y}_i$
2	3	6	4	4
7	7	49	49	6.5
1	4	1	1	3.5
4	6	16	16	5
6	5	36	36	6
20	25	113	106	

If you have the following data with two dependent and two independent variables, you are required to:

1- Plot the scatter plot of this data
2- Find the equation of the regression line
3- Does the equation of the regression line satisfy that the sum of random errors = zero
4- Plot the equation of the regression line
5- Find the expected value of y when $x_0 = 0{,}10$
6- What is the relationship between x_i and y_i
7- What are x_i, y_i, $\hat{\beta}_0$, and $\hat{\beta}_1$

Solution:

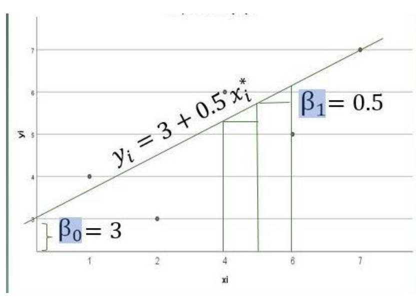
1) Scatter Plot



3)

y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
3	4	-1
7	6.5	0.5
4	3.5	0.5
6	5	1
5	6	-1
		0

4)



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

$$\hat{\beta}_1 = \frac{S_{Xy}}{S_{xx}}$$

$$S_{xy} = \sum x_i y_i - \frac{(\sum y_i)(\sum x_i)}{n}$$

$$= 113 - \frac{(20)(25)}{5} = 13$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$= 106 - \frac{(20)^2}{5} = 26$$

5) $\hat{y}_i$=3+0.5x_i

$$\hat{y}_i = 3 + (0.5)(0) = 3$$

$$\hat{y}_i = 3 + (0.5)(10) = 8$$

$$\hat{\beta}_1 = \frac{13}{26} = 0.5$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$= 5 - (0.5)(4) = 3$$

6) From the sign $\hat{\beta}_1$ we notice that the relationship between x_i and y_i is a positive (direct) relationship, meaning that a one-unit increase in x_i leads to an increase in y_i by 0.5.

7) x_i: independent variable

y_i: dependent variable

$\hat{\beta}_0$: point of intersection of the regression line with the axis

$\hat{\beta}_1$: represents the regression coefficient y/x and represents the amount of change in y when the value of x increases by one unit.