UNIT 7. CORRELATION AND SIMPLE LINEAR REGRESSION

In this lesson, we will present the models that describe the relationship between two variables. We will first introduce Regression Analysis and Linear Correlation Analysis from a descriptive point of view. We will learn how to obtain the regression line using least squares minimisation, how to measure the goodness of fit, how to interpret the estimations of the parameters and how to predict, distinguishing between interpolation and extrapolation.

We may remember that if two variables show no sign of relationship between them, that is, if they are independent, their joint study is uninteresting. Nevertheless, in many economic phenomena, due to the many interactions, some *kind of association* between the observed values of two or more characteristics is likely to exist. For example, the values of the "*savings of a group of employees"* are expected to be in some way related to the observed values of the *employees' disposable income*. Or, in the case of a company, for instance, the *monthly sales* figures will probably be associated with the amount of the *investment in advertising.* Whenever it is possible to predict accurately the observed values of a variable in terms of the values adopted by other(s) by means of a mathematical function, a special type of dependence called 'functional' occurs.

Whenever, without reaching functional dependence, the data observed show some degree of association between them, we say that a **statistical dependence** between the variables exists, and its analysis is the aim of this unit. When analysing the statistical dependence, two complementary goals will be addressed:

1. The quantitative analysis of the intensity of the dependence, that is, the degree of association. *(To what extent do the observed sales depend on the investment in advertising?).* This is the goal of **Correlation Analysis**.
2. The determination or fit of a function describing the behaviour *(the values and its variations)* of a variable by means of the values of other(s). This objective is carried out through **Regression Analysis**.

SCATTER PLOT

In order to visually analyse the type of relationship existing between the two variables considered X and Y, one of the most usual graphic representations of a joint

frequency distribution is the scatter plot, which is a type of mathematical diagram using cartesian coordinates to display values for a set of bivariate data.

Example:

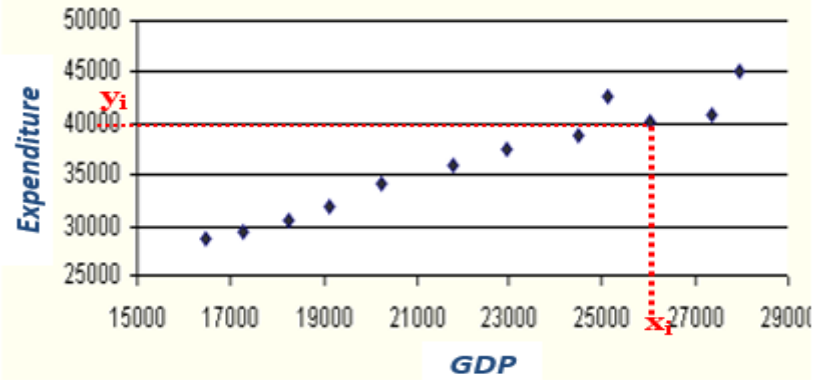


Figure 1. Scatter plot of the Gross domestic product (GDP) versus the expenditure

These diagrams allow us to see whether there is a relationship between the variables, whether this relationship is linear or non-linear, whether it is direct or inverse, and the intensity of the relationship.

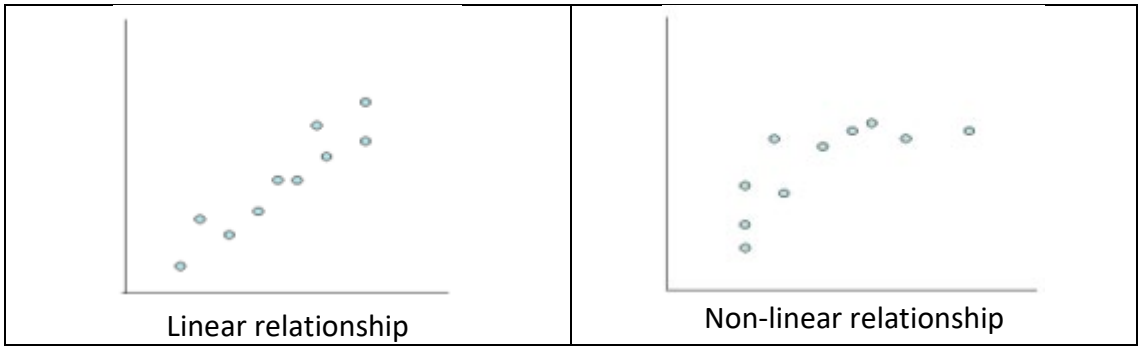


Figure 2. Linear vs. non-linear relationship (both are direct)

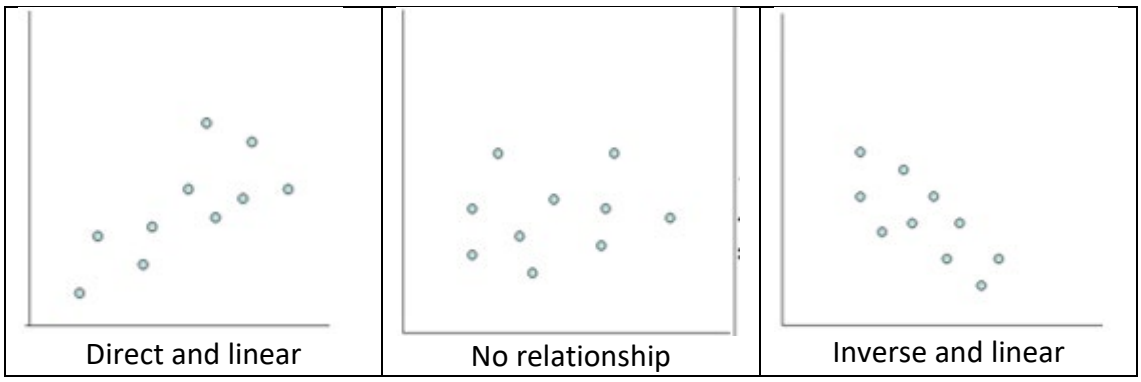


Figure 3. Direct Linear vs. No relationship vs inverse linear relationship

COVARIANCE AND CORRELATION COEFFICIENT

In order to quantify the intensity of the association some coefficients, known as correlation coefficients, are used. They are linked to a key magnitude called Covariance. The Covariance is a measure of how much two variables change together (measures the growth of both at the same time or the growth of one and decrease of the other). The Covariance between two variables is defined as:

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

The covariance can be expressed in a shorthand form as:

$$S_{XY} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

Interpretation of the Covariance

The covariance is a measure of the degree of joint variation of two variables in the sense that its sign informs about the existence or not of the tendency of the cloud of points to be located preferentially in the first and third quadrants (if $S_{XY} > 0$), second and fourth (if $S_{XY} < 0$) or in none of them (if $S_{XY} = 0$) in the diagram that takes as its origin the centroid of the frequency distribution, $(\bar{x}, \bar{y})$. Thus, the points located in the first and third quadrants verify that $(x_i - \bar{x}) \times (y_i - \bar{y}) > 0$ while those located in the second and fourth quadrants $(x_i - \bar{x}) \times (y_i - \bar{y}) < 0$.

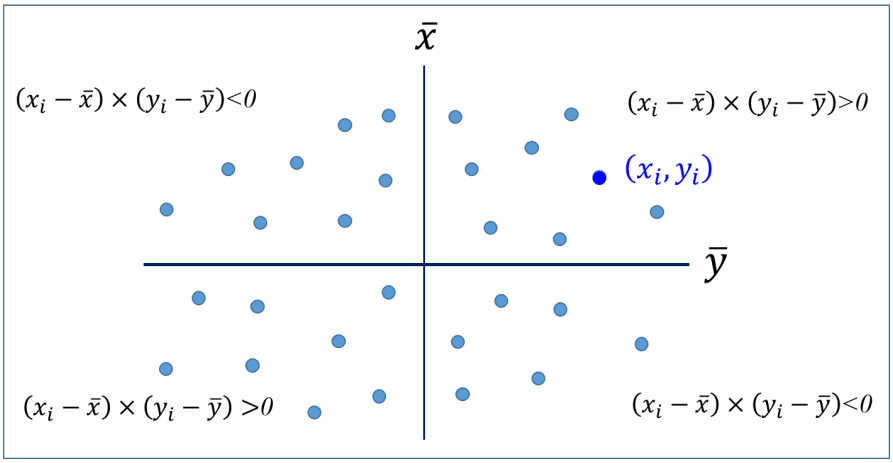


Figure 4. Sign of the covariance terms

The covariance calculates the mean of the above products so that if $S_{XY} > 0$ it is because the number of pairs (x_i, y_i) that verify that $(x_i - \bar{x}) \times (y_i - \bar{y}) > 0$ is greater than those that verify that $(x_i - \bar{x}) \times (y_i - \bar{y}) < 0$ (see Figure 5a), the opposite being true if $S_{XY} < 0$ (see Figure 5b) and both sets being balanced if $S_{XY} \sim 0$ (see Figure 5c and d). Therefore, if $S_{XY} > 0$ this indicates that the relationship between X and Y is direct, i.e. the higher the value of X, the higher the value of Y tends to be (see Figure 5a), while if $S_{XY} < 0$ the relationship is inverse, i.e. the higher the value of X, the lower the value of Y tends to be (see Figure 5b). If $S_{XY} \sim 0$ nothing can be said about the type of relationship: it may not exist (see Figure 5c) or it may not be linear (see Figure 5d).

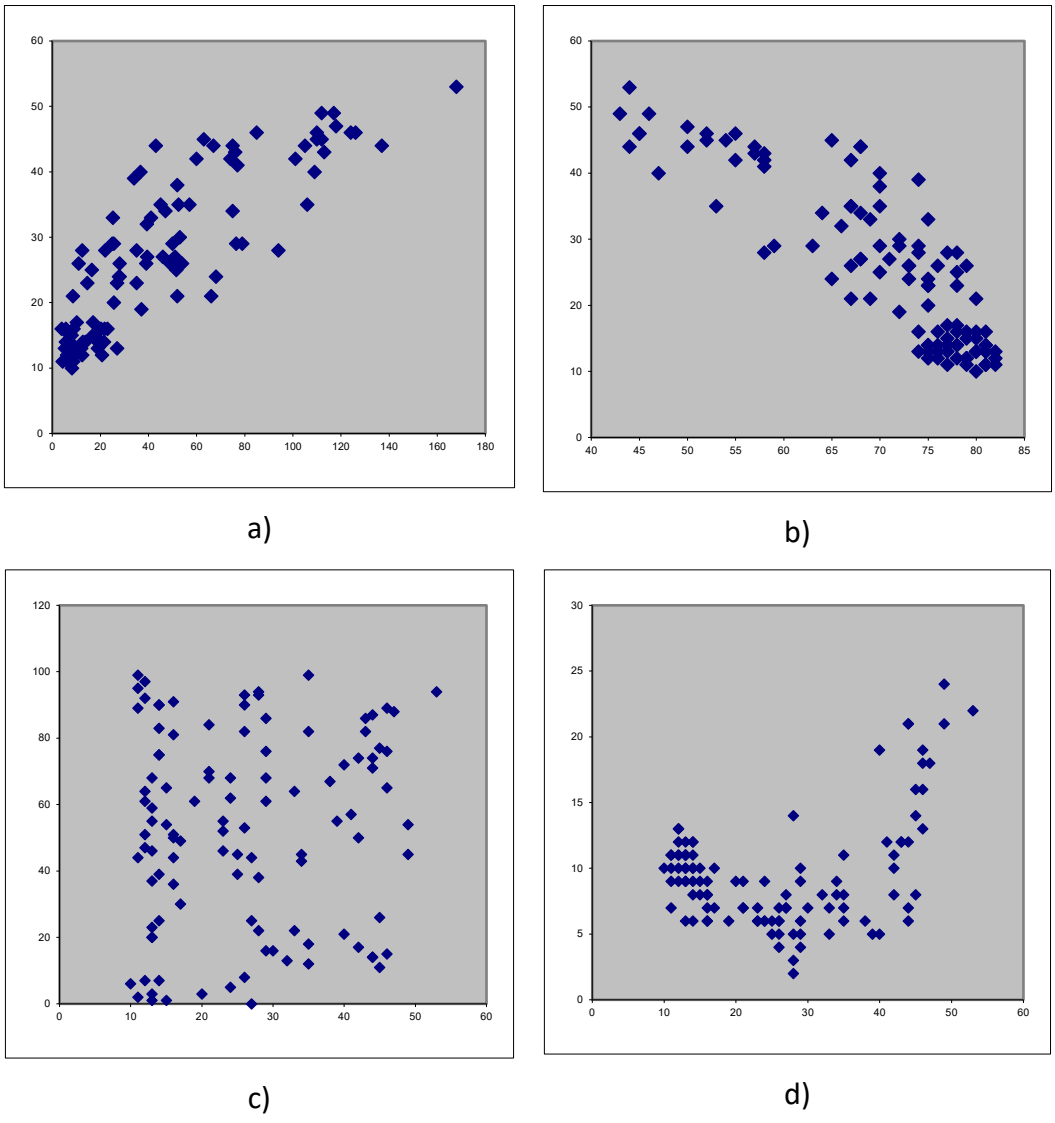


Figure 5. Scatter plots for variables (X,Y) with $S_{XY} > 0$ (a), $S_{XY} < 0$ (b) and $S_{XY} \sim 0$ (c), (d)

Properties of the Covariance

- The Covariance depends on the units of measurement.
- The Covariance is invariant for translation transformations but not for scale changes.

$$\begin{matrix} U = a + bX \\ V = c + dY \end{matrix} \Rightarrow Cov(U,V) = bdCov(X,Y)$$

- The variances of the two variables and the covariance are arranged in a matrix known as the **Covariance matrix:**

$$S = \begin{pmatrix} S_X^2 & S_{XY} \\ S_{XY} & S_Y^2 \end{pmatrix}$$

Correlation Coefficient

It is a dimensionless coefficient associated with the covariance. It is used to measure the intensity of the association and its direction.

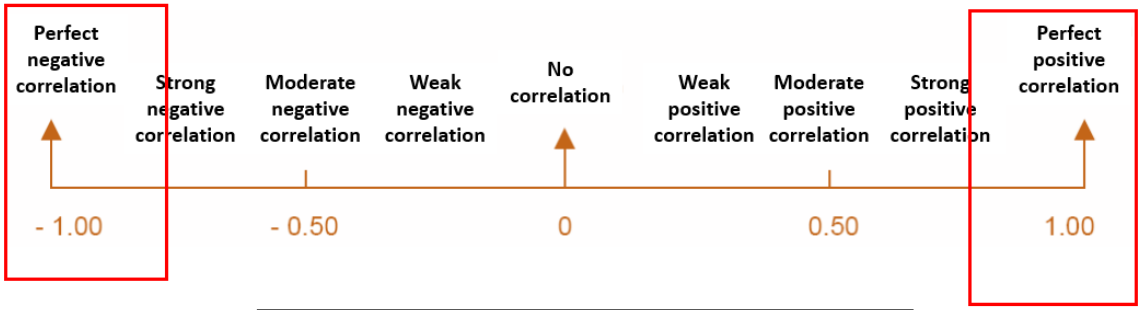
The **Pearson's Linear Correlation Coefficient** measures the degree of linear association between two quantitative variables, in relative terms, with respect to the dispersion of these variables.

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

Properties of the Correlation Coefficient

- Has the same sign as the Covariance
- It is a dimensionless coefficient
- Ranges from -1 to 1
- If a precise (functional) linear relationship between the two variables exists, then its value is 1 or -1
- It is invariant for linear transformations, except for the sign

$$r = \frac{S_{XY}}{S_X S_Y}$$



Values of |r| over 0.75 are commonly required as an indicator of a significant level of linear dependence between the variables X and Y

Example:

The managers of a multinational company want to analyse the possible relationship between the **Annual profits** (Y) and the **Advertising expenditure** (X) for a group of different products. The values are shown in the following table:

Year	X (million €)	Y (million €)
1	2	-6
2	2.8	-3
3	3.9	0
4	4.2	3
5	5.8	6
6	6.2	9
7	7.5	12
8	8.2	15
9	9.3	20
10	10.9	25

Determine the Linear Correlation Coefficient.

Year	X (million €)	Y (million €)	xi^2	yi^2	xi*yi
1	2	-6	4	36	-12
2	2.8	-3	7.84	9	-8.4
3	3.9	0	15.21	0	0
4	4.2	3	17.64	9	12.6
5	5.8	6	33.64	36	34.8
6	6.2	9	38.44	81	55.8
7	7.5	12	56.25	144	90
8	8.2	15	67.24	225	123
9	9.3	20	86.49	400	186
10	10.9	25	118.81	625	272.5
Sum	**60.8**	**81**	**445.56**	**1565**	**754.3**

$\bar{x} = 6.08 \qquad \bar{y} = 8.1$

$S_X^2 = 7.59 \qquad S_Y^2 = 90.89$

$S_{XY} = 26.182$ (million €)^2

$$S_X = \sqrt{7.59} = 2.75 \text{ million €}$$

$$S_Y = \sqrt{90.89} = 9.53 \text{ million €}$$

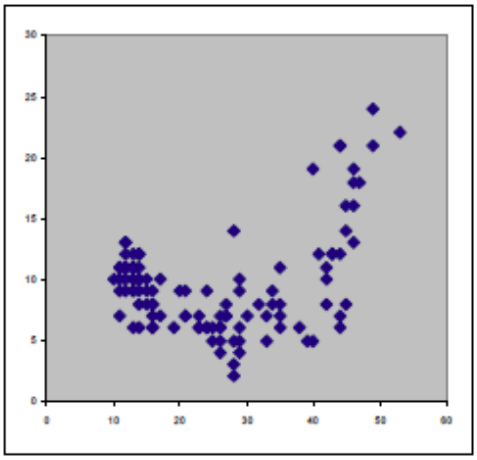
$$r_{XY} = \frac{S_{XY}}{S_X \times S_Y} = \frac{26.182}{2.75 \times 9.53} = \mathbf{0.9969}$$

⇒ **Very strong direct linear** relationship

Let's now define a concept related with that of the independence of two variables.

Definition: X and *Y* are said to be **uncorrelated** if $r_{xy} = 0$ ($S_{xy} = 0$)

If two variables *X* and *Y* are statistically independent ⇒ X and Y are uncorrelated ($r_{xy} = 0$). The opposite is not true: two variables may have a null correlation and still be non-independent. There are numerical counterexamples that corroborate this. For example, in the following figure it can be seen that there may be other associations, of a non-linear type, which can make the covariance equal to zero.



LINEAR REGRESSION MODEL

Regression analysis is aimed at **describing the relationships** between two variables. As the values, or more generally, the behaviour of a variable *Y* is influenced in an

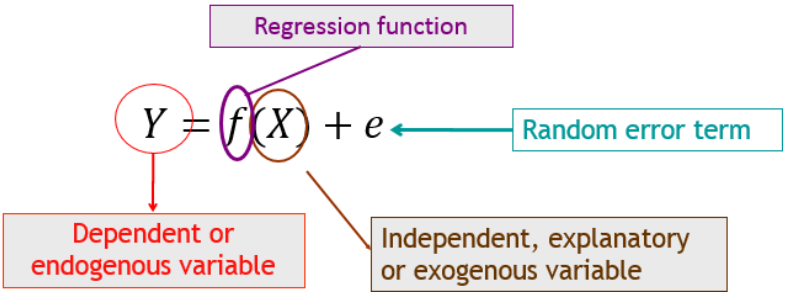
unknown way by the values of a second variable *X*, the regression analysis assesses, through an appropriate function, the dependence **Y=f(X)** and establishes its adequacy.

As the formal relationship is not usually precise, but just an approximation in which other variables of secondary importance are to be omitted, regression models include an additional error term, that will reflect the missing factors whose influence on the variable is secondary and that, individually, are not relevant.

The Regression problem consists of obtaining the equation of a curve that goes "close" to the points represented in the scatter plot, and which adapts as well as possible to the set of these points, fulfilling certain conditions. **Regression Analysis** is applied in two stages:

- **STEP 1: MODEL SELECTION.** Decide the **kind of function (curve)** that **best fits the data set**, that is, that best explains the change in an output or endogenous variable (Y) for each value observed of an input or exogenous variable (X). At this stage, a scatter plot of the data set is often very helpful, since the pattern of dots orientates the choice of the trend line or line of best fit.
- **STEP 2: ESTIMATION OR FITTING.** Once the mathematical family of the function has been set, the **choice of the function** in the family **closest** to the **observed point** has to be made.

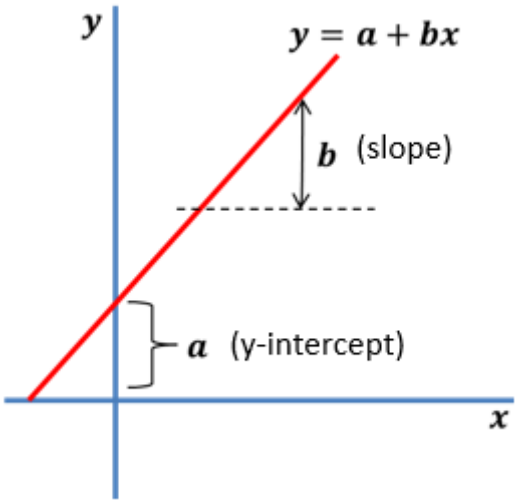
In short, a criterion must be established to determine the best values for the function coefficients or parameters. Formally, a General Regression Model is represented by an equation:



For $f(X) = a + bX$, a and b constants. The linear regression model is expressed as: $Y = a + bX + e$

- Parameter ***b*** is the **slope** of the line, also called Regression Coefficient. Its value represents the change in the dependent variable for every unit change in the explanatory variable. For this reason, in the field of economics it is identified with the so-called **Marginal Propensity**.

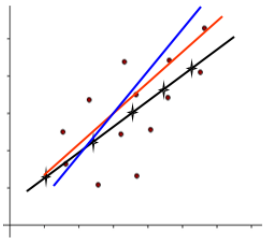
- Parameter ***a*** is the ***y*-intercept** and represents the value of the dependent variable when the explanatory variable is set to zero.



Estimation of the linear regression model

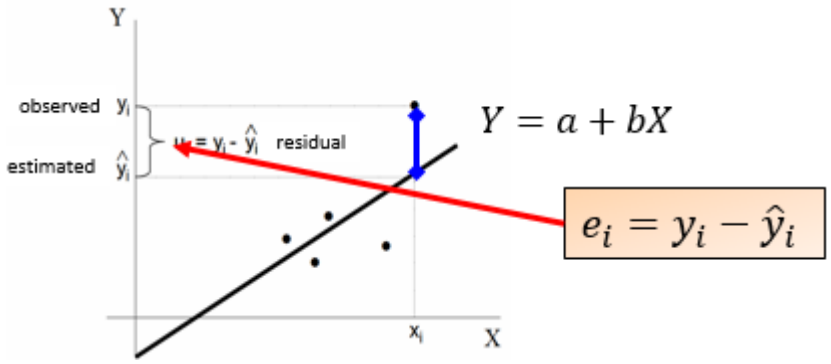
The goal is to estimate, based on the data set, the values of the unknown coefficients *a* and *b* of the model that best fit the points. In the first place, it is useful to use an intuitive approach in order to set some fitting criteria. For that purpose, the scatter plot is used. If the dependence was accurate, the observed data would lie along a straight line. In general, the data are not aligned; they are arranged as a cloud of dots. In such a case, we must look at the straight line as a *"formal"* approach, and the fitting problem as the choice of the "nearest" line to the data set.

$$Y = a + bX$$



In order to measure the closeness of the line to the scatter plot, for each observed value of the variable *X*, x_i, we must take into account two values of *Y*:

- the **observed** value y_i
- The **estimated** value given by the expression $\hat{y}_i = a + bx_i$



The **difference** between the observed and the estimated value is the **residual or error** $e_i = y_i - \hat{y}_i$

Their values take into account the fluctuations of the values of the variable *Y* that are not explained by its relationship with the variable *X.*

It seems reasonable to estimate the unknown constants in a way that they provide the **smaller overall residuals**.

Global measures for the size of the errors or residuals

- The first synthesis strategy would be the sum, or the average: $SRes = \sum_{i=1}^{N} e_i = \sum_{i=1}^{N}(y_i - \hat{y}_i)$. The weak point of this approach is that when making the sum of all positive and negative residuals, the total sum yields a wrong synthesis about the closeness.
- A second alternative synthesis strategy would be to measure the intensity of the residuals, without a sign (absolute value): $SARes = \sum_{i=1}^{N}|e_i| = \sum_{i=1}^{N}|(y_i - \hat{y}_i)|$. Now, the weak point is the difficulty to do algebraic calculations, in particular derivatives.
- The synthesis measure that overcomes the disadvantages of the previous ones is based on the **sum** –or the average- of the **squared residuals**:

$$SSRes = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \Rightarrow MSE = \frac{1}{N}\sum_{i=1}^{N} e_i^2 = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Mean Squared Error

The fit of the regression line on the basis of the MSE consists of the determination of the fit which **minimizes the mean squared error**, giving rise to the most common criterion, called **Method of least squares**.

Getting the best fit using the method of least squares becomes a mathematical problem:

Minimise $\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \min_{a,b} \sum_{i=1}^{N} (y_i - (a + bx_i))^2$

Data (Constants)

Parameters (Variables)

Differentiating with respect to each of the two parameters *a* and *b*, and setting to zero the resulting expressions in order to find the critical points, a system of two equations is obtained:

Normal equations of regression

$$\begin{cases} \sum_i y_i = aN + b\sum_i x_i \\ \sum_i x_i \cdot y_i = a\sum_i x_i + b\sum_i x_i^2 \end{cases}$$

The resolution of this system of equations provides the values of *a* and *b*. Formally, these values are given by:

$$b = \frac{S_{X,Y}}{S_X^2} \qquad a = \bar{y} - b \cdot \bar{x}$$

The regression line of Y on X may also be expressed by the point-slope equation:

$$Y - \bar{y} = \frac{S_{XY}}{S_X^2}(X - \bar{x})$$

From the above expression, it is easy to see that the equation is satisfied for the point $(\bar{x}, \bar{y})$.

Interchanging the roles of the variables and carrying out the same procedure, we obtain the equation of the regression line of X on Y: $\boldsymbol{X} = a_{X|Y} + b_{X|Y}\boldsymbol{Y}$

$$a_{X/Y} = \bar{x} - b_{X/Y}\bar{y}$$

$$b_{X|Y} = \frac{S_{XY}}{S_Y^2}$$

$b_{X/Y}$ – Regression coefficient of *X* on *Y*

Alternatively, the regression line of X on Y may be expressed by the point-slope equation:

$$X - \bar{x} = \frac{S_{XY}}{S_Y^2}(Y - \bar{y})$$

From the above expression, it is easy to see that the equation is also satisfied for the point $(\bar{x}, \bar{y})$, i.e., both regression lines intersect at the centroid.

Properties of the regression coefficients

- Given the regression line Y = *a* + *b* X, If the regression coefficient *b* is positive, the scatter plot is arranged so that the values of Y increase when the values of X increase. In this case, it can be observed that the linear dependency of the variables X and Y is direct. If the regression coefficient *b* is negative, the scatter plot is arranged so that the values of Y decrease when the values of X increase. This is a case of inverse linear dependency. The regression coefficient *b* is invariant for translation transformations but not for scale changes.

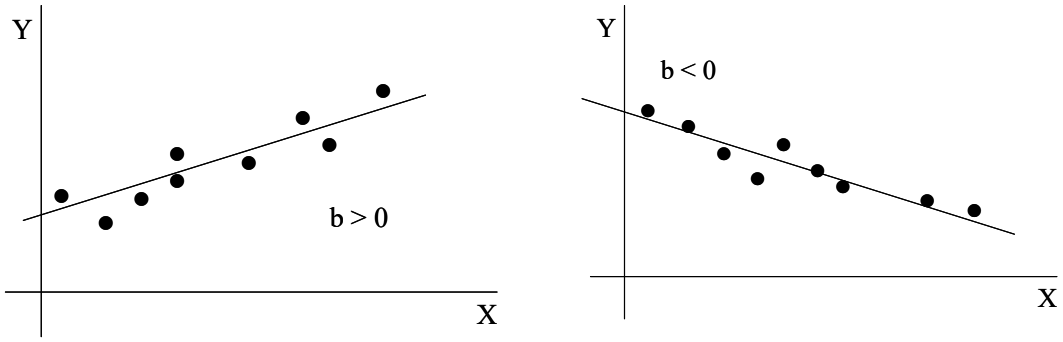


Figure 6. Sign of the linear dependence of X and Y

- The sign of *b* coincides with the sign of the covariance, S_{XY}:

$$b = \frac{S_{XY}}{{S_X}^2} \quad \text{as } {S_X}^2 \geq 0 \Rightarrow sign\left[b = \frac{S_{XY}}{{S_X}^2}\right] = sign[S_{XY}]$$

Properties of the linear regression model

- Both regression lines intersect at the centroid $(\bar{x}, \bar{y})$
- The average of the residuals obtained by the least squares method is zero:

$$\bar{e} = \frac{1}{N}\sum_{i} e_i = 0$$

- The average of the theoretical values given by the linear model equals the average value of the dependent variable. That is to say, on average, the prediction equals the average value of the dependent variable.

$$\bar{\hat{y}} = \frac{1}{N}\sum_{i} \hat{y}_i = \bar{y}$$

GOODNESS OF FIT

We have estimated a linear regression model, and new questions arise: Does it fit the data well? Does it describe the relationship between X and Y? Can it be used to predict? Once a regression model has been estimated, it is necessary to evaluate how well it fits the data, that is, how well it describes the dependency between the data. This is known as **Goodness of fit**. To evaluate it, a numerical measure of the proximity between the model and the data is needed. The measurements that quantify the goodness of fit are based on the values of the residuals or errors. When the residuals are generally small, the fit will be good, and the regression line will be reasonably representative. However, how can we measure the overall size of the residuals? In the following, we will present some of the measures that exist to quantify this intensity.

Residual Variance and Coefficient of Determination R^2

The overall size of the residuals is approximated by averaging their squares -without a sign. This is called **Residual Variance**. It is a key measure to evaluate the degree of goodness of fit.

$$S_{r_Y}^2 = \frac{1}{N}\sum_{i=1}^{N} {e_i}^2 = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

The lower (higher) the residual variance, the better (worse) the fit, and the closer (further) the predictions to the observations. The main drawback is that the magnitude of the residual variance depends on the magnitude of the data and on the units of measurement.

To express the goodness of fit in relative terms, the coefficient of determination (R^2) is defined:

$$R^2 = 1 - \frac{S_{r_Y}^2}{S_Y^2}$$

$$0 \leq R^2 \leq 1 \begin{cases} R^2 = 1 \Rightarrow S_{rY}^2 = 0 \Rightarrow \forall i \;\; e_i = 0 \Rightarrow y_i = \hat{y}_i \\ R^2 = 0 \Rightarrow S_{rY}^2 = S_Y^2 \end{cases}$$

Perfect fit

Poor fit

The amount $1 - R^2 = \frac{S_{rY}^2}{S_Y^2}$ gives the proportion of the residual variance (not explained by the regression model) with respect to the variance of the endogenous variable.

R^2 can also be expressed as: $R^2 = 1 - \frac{S_{rY}^2}{S_Y^2} = \frac{S_Y^2 - S_{rY}^2}{S_Y^2} = \frac{S_{exp_Y}^2}{S_Y^2}$

where the numerator is called **Explained variance**: $S_{exp\,Y}^2 = S_Y^2 - S_{rY}^2$

This explained variance can be calculated as:

$$S_{\exp Y}^2 = \frac{1}{N}\sum(y_j - \bar{y})^2 - \frac{1}{N}\sum(y_j - \hat{y}_j)^2 = \frac{1}{N}\sum(\hat{y}_j - \bar{y})^2$$

From the previous expression, $R^2 = \frac{S_{\exp Y}^2}{S_Y^2}$ which represents the percentage of the variation in Y explained by the regression.

The coefficient of determination is invariant for linear transformations.

In the linear regression model, the explained variance equals the variance of the estimated values: $S_{\exp Y}^2 = S_{\hat{Y}}^2$.

Linear coefficient of determination

In the specific case of linear regression, the coefficient of determination takes a particular expression. By combining the formulas that give the least squares regression line and the definition of the coefficient of determination, the Linear Coefficient of Determination is obtained:

$$R^2 = r^2 = \frac{S_{XY}^2}{S_X^2 S_Y^2}$$

It can be appreciated that the Linear Coefficient of Determination is the square of the Linear Correlation Coefficient, previously defined as:

$$r = \frac{S_{XY}}{S_X S_Y}$$

NOTE: **Correlation does not imply causation**. The presence of a strong association or linear correlation between two variables does not guarantee the existence of a causal connection between them. A **spurious relationship** is a situation in which two or more variables are statistically related but no causal link between them exists (no logical connection). In most cases, there is a third variable that explains both. Some examples are:

- Height and incomes of people.
- Number of single aunts and level of calcium in bone.
- Ice cream sales and number of people fainting in a town.

PREDICTION

In a large number of applications of regression, the ultimate goal is Prediction. Once the regression model has been fitted, the value of the endogenous variable for a specific value of the explanatory variable is approximated through the equation of the regression model. The resulting value is known as Prediction.

Thus, to predict the value of Y given a value X=x_h of the explanatory variable, it is enough to substitute it into the model equation Y=(X).

$$\hat{y}_h = f(x_h)$$

In the specific case of linear regression:

$$\hat{y}_h = a_{y|x} + b_{y|x} x_h$$

If the value x_h for which the model is used falls within the range of observed values of the variable X, it is said that the prediction is an **interpolation**. When the value of interest lies outside the range of observed values the prediction is said to be an **extrapolation**.

Analogously, when we want to predict values of the variable X, we consider the regression line of X on Y. This prediction of X for a specific value y_h of the variable Y would be obtained by simply substituting the value y_h in the equation of the line. That is to say:

$$\hat{x}_h = a_{x|y} + b_{x|y} y_h$$

When extrapolating, it is important to keep in mind that the reliability of the prediction will be lower, as the linearity of the relationship between X and Y may not exist outside the range of the observed values. To sum up:

$$If\ the\ fit\ is\ good\ and\ it\ is\ an\ interpolation \Rightarrow \boldsymbol{Reliable}\ prediction$$

$$If\ the\ fit\ is\ poor\ or\ it\ is\ an\ extrapolation \Rightarrow \boldsymbol{Unreliable}\ prediction$$