1. SIMPLE LINEAR REGRESSION

In this chapter we will refer to modeling the relationship between two variables when the relationship is linear.

Let us consider two variables, X and Y .

Consider then the variable Y as the dependent variable, i.e. one whose behaviour (variation) depends on the behaviour (variation) of another variable, X , which we call independent or explanatory variable.

Let us even assume that the relationship between the two variables is linear. In such a case, a simple linear regression model would be appropriate to describe the relationship between X and Y .

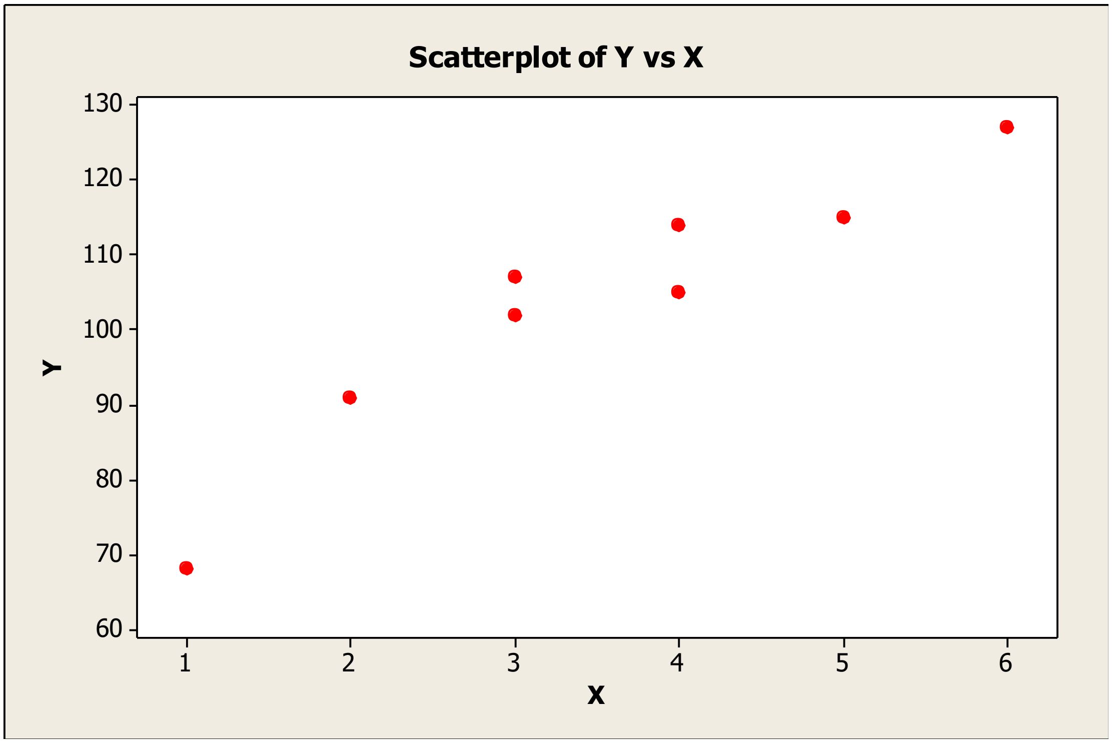
We could in our model introduce several independent (explanatory) variables, i.e. variables that co-decide for the behaviour of the dependent variable (Y). In the latter case and if the relationship of each independent variable with the dependent is linear, we get a multiple linear regression model with which we will deal in the next chapter.

Example

Consider the height in centimetres (Y) and the age in completed years of life (X) of a sample of eight pre-schoolers. The observed pairs of values of the two variables are given below:

Y ***Height***	***X*** ***Age***
68	1
91	2
102	3
107	3
105	4
114	4
115	5
127	6

1. Diagrammatic illustration of the X, Y pairs of observations: (x, y).



2. Calculation of correlation coefficient

We will need:

$$\boxed{\begin{aligned} \sum Y_i &= 829 \\ \sum X_i &= 28 \\ \sum Y_i^2 &= 88\ 133 \\ \sum X_i^2 &= 116 \\ \sum X_i Y_i &= 3090 \end{aligned}}$$

To calculate:

$$SS_X = \sum\left(X_i - \bar{X}\right)^2 = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n} = 116 - \frac{28^2}{8} = 18$$

$$SS_Y = \sum\left(Y_i - \bar{Y}\right)^2 = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n} = 88133 - \frac{829^2}{8} = 2227{,}875$$

$$SS_{XY} = \sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = \sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n} = 3090 - \frac{28 \cdot 829}{8} = 188{,}5$$

Therefore the correlation coefficient between the two variables in this sample is,

$$\boxed{r = \frac{SS_{XY}}{\sqrt{SS_X \cdot SS_Y}} = \frac{188{,}5}{\sqrt{18 \cdot 2227{,}875}} = 0{,}9413}$$

Based on these calculations, to assess the significance of correlation coefficient ρ of the population, we have to test the hypothesis:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

The test statistics (control variable) is

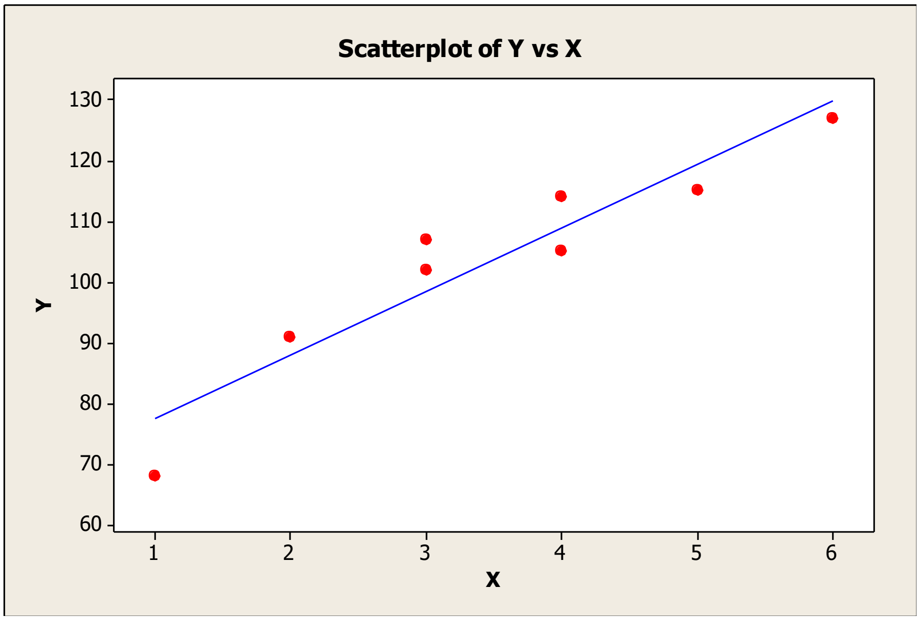
,

$$t_{(n-2)} = \frac{0,9413}{\sqrt{(1-0,9413^2)/(8-2)}} = 6,86 \; > \quad t_{0,975}^{(n-2)} (= 2,447) => H_0 \text{ rejected}$$

Thus, with a risk α= 0.05 (5%) we conclude that the ρ can be considered as statistically significant

So it makes sense to implement a simple linear model to describe the relationship between Y and X

In this step we must decide which is the dependent and which the independent variable.



Theoretical background

The theoretical simple linear regression model is

$$\boxed{Y = \beta_0 + \beta_1 X + \varepsilon}$$

Here Y It is the dependent variable, i.e. the variable that we wish to explain or predict.

and X is the independent variable, which is also called the explanatory variable, and finally

ε are the errors, which is the only random term in the model and therefore the only source of randomness in the behaviour of Y .

The model of simple linear regression, consisting of two parts: the non-random part, which is the straight line itself and the random parts i.e. the residuals or errors.

The non-random part of the model, the straight line, is the relationship which expresse the expected value of Y for any given value x_i as a linear function of X .

$$\boxed{E(Y\big|X) = \beta_0 + \beta_1 X}$$

Thus

$$\boxed{Y = E(Y\big|X) + \varepsilon}$$

Theoretical assumptions of the model:

1) ***The relationship between X and Y is linear.***

2) ***The different values of the independent variable X is fixed (not random). The only source of randomness in Y therefore stems from the presence of residuals ε.***

3) ***The residuals ε (As many as the distinct values X) are random variables that follow the normal distribution with expected value 0 and constant variance σ^2.***

Using statistical notations, we have:

$$\varepsilon \sim N(0,\sigma^2)$$

The theoretical regression model describes the relationship between two variables X and Y in the reference population. To estimate the parameters of the model, we use the observed pairs of values of the two variables. Based on this set of observed values X and Y we cannot calculate the regression model parameters β_0 and β_1 but simply estimate them. The classical way for estimating these two model parameters, is the method of ordinary least squares, as described in the next section.

ORDINARY LEAST SQUARES METHOD (OLS)

The challenge, as already mentioned in the previous section, is to calculate the best possible estimates of the parameters of our model. The desired line is the one that passes as closely as possible the observed pairs (x,y). Still want the estimators of the model parameters β_0 and β_1, let us denote them by $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively, to meet the criteria of unbiasedness and efficiency, so that, based on these, using the observed data, to calculate the best possible point estimates of parameters β_0 and β_1. The observed values of $\hat{\beta}_0$ and $\hat{\beta}_1$, will be denoted as b_0 and b_1 respectively.

One method that will give our estimators these desired features are the OLS method.

This method is not unique to tailor a straight line to the data. There are alternative methods, such as minimization of the sum of the absolute values of errors. However the method of least squares is the most widely used method for estimating the parameters of a regression model.

The estimation of our model in this sample of observations X and Y will be,

.

$$Y_i = b_0 + b_1 X_i + e_i \qquad i = 1, 2, \ldots, n$$

where

b_0 is the estimate of β_0 and b_1 is the estimate β_1, while e_i are the observed values of the residuals., which are the observed values of the actual population errors ε_i .

The linear relationship of X and Y in this sample is given by

$$\hat{Y} = b_0 + b_1 X$$

where $\hat{Y}$ the fitted value of Y

$$\begin{tikzcd} \textit{Parameters} & \textit{estimators} & \textit{estimates} \\ \beta_0 \arrow[r] & \hat{\beta}_0 \arrow[r] & b_0 \\ \beta_1 \arrow[r] & \hat{\beta}_1 \arrow[r] & b_1 \end{tikzcd}$$

Having defined the estimated regression relationship, errors, and fitted values of Y, we will now describe the principles of the least squares method, which give us unbiased and efficient estimators.

Our aim is to minimize all errors, which are sometimes positive and sometimes negative, but if we raise to the square, these become positive. We consider the sum of squares of errors,

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i)]^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i)^2$$

Our challenge now is to identify the estimators $\hat{\beta}_0$ *and* $\hat{\beta}_1$ so that they will give us the minimum possible price SS_E.

Taking the first partial derivatives of SS_E and setting them equal zero
The normal equations are:

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \,.$$

The two equations with two unknowns if solved would give us formulas for calculating the estimator $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize SS_E.

The estimators of the parameters of the simple linear regression model based on the least squares method are:

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_X}$$

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

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables having smallest possible variance, as is apparent from the theorem of Gauss-Markov.

These estimators can be used, along with the assumption of normality, to construct confidence intervals and to conduct hypothesis testing about parameters β_0 and β_1 the theoretical model. This methodology will be given below.

EXAMPLE (cont.)

Returning to our example, using the sample data of Table 1.1 can fit a simple linear model to height (Y) and (X).

We have:

$$SS_X = \sum\left(X_i - \overline{X}\right)^2 = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n} = 116 - \frac{28^2}{8} = 18$$

$$SS_Y = \sum\left(Y_i - \overline{Y}\right)^2 = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n} = 88133 - \frac{829^2}{8} = 2227,875$$

$$SS_{XY} = \sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = \sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n} = 3090 - \frac{28 \cdot 829}{8} = 188,5$$

$$\boxed{\begin{gathered} \hat{\beta}_1 = \frac{SS_{XY}}{SS_X} \Rightarrow b_1 = 10,47 \\ \\ \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \Rightarrow b_0 = 66,972 \end{gathered}}$$

$$\boxed{\hat{Y} = 66,972 + 10,472 \cdot X}$$

THE VARIATION OF RESIDUALS

Let us remember that σ^2 the variance of errors ε in the theoretical regression model (i.e. that concerning population) is supposed to be constant for all variables ε associated with various values of X. The variation of errors, σ^2 is an important parameter because it is a criterion that shows how dispersed are the data around the regression line. Generally, the smaller the variation is, the closer to the population data passes the straight regression. It is important to understand that the variance of the errors is the variation of the dependent variable Y corresponding to any given value of X. This easily we can see since, as previously presented, for any given value of X we have got,

$$Y \mid X = E\left(Y \mid X\right) + \varepsilon,$$

note that $E(Y \mid X)$ as population parameter is stable with zero variation. Thus,

$$Var\left(Y \mid X\right) = Var\left(\varepsilon\right) = \sigma^2$$

As a population parameter, σ^2 is unknown and thus it must be estimated from the observed data. An unbiased estimator for σ^2 is the mean square error (MS_E) defined as,

$$MS_E = s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}$$

An unbiased estimator of σ^2, is

$$MS_E = SS_E / \left(n-2\right),$$

where SS_E is the sum of squares for the errors.

$$SS_E = \sum (Y - \hat{Y})^2 = SS_Y - \frac{(SS_{XY})^2}{SS_X} = SS_Y - b_1 SS_{XY}$$

The positive square root of this, $s = \sqrt{MS_E}$ is the standard error of the residuals. This is criterion of goodness of fit of the model and is often referred as a standard error of the estimate.

Note that this criterion is expressed in units of our dependent variable. Therefore it cannot be used as a criterion of comparison between different applications of the linear model when the dependent variables of these applications have the same metric. But this criterion can be an alternative model comparison tool on the same dependent variable.

In our example we take,

$$SS_E = \sum (Y - \hat{Y})^2 = SS_Y - b_1 SS_{XY} = 2227,875 - 10,472 \cdot 188,5 = 42,31$$

=> $s = \sqrt{MS_E} = \sqrt{42,31} = 6,505$.

The standard errors of the estimators of model parameters.

The standard error of $\hat{\beta}_0$ *is*

$$s\left(\hat{\beta}_0\right) = \frac{s\sqrt{\sum x_i^2}}{\sqrt{n \cdot SS_x}} \quad ,$$

where $s = \sqrt{MS_E}$.

The standard error of $\hat{\beta}_1$ *is*

$$s(\hat{\beta}_1) = \frac{s}{\sqrt{SS_X}}$$

in our example:

$$s\left(\hat{\beta}_0\right) = \frac{s\sqrt{\sum x_i^2}}{\sqrt{n \cdot SS_x}} = \frac{6{,}505\sqrt{116}}{\sqrt{8 \cdot 144}} = 5{,}838$$

$$s(\hat{\beta}_1) = \frac{s}{\sqrt{SS_X}} = \frac{6{,}505}{\sqrt{144}} = 1{,}533$$

CONFIDENCE INTERVALS FOR THE PARAMETERS OF THE MODEL

The confidence intervals for the parameters β_0 and β_1 are easy to calculate using the estimators and the standard error for each one of the two parameters.

The (1-a) 100% confidence interval for the β_0 *is,*

$$\hat{\beta}_0 \pm t_{[(n-2),1-\alpha/2]} \cdot s\left(\hat{\beta}_0\right)$$

The (1-a) 100% confidence interval for the β_1 *is:*

$$\hat{\beta}_1 \pm t_{[(n-2),1-\alpha/2]} \cdot s\left(\hat{\beta}_1\right)$$

in our example

$$b_0 \pm t_{(n-2),1-\alpha/2} s\left(b_0\right) = 66,972 \pm 2,447 \cdot 5,84 => 52,7 < \beta_0 < 81,3$$

$$b_1 \pm t_{(n-2),1-\alpha/2} s\left(b_1\right) = 10,472 \pm 2,447 \cdot 1,53 => 6,8 < \beta_1 < 14,2$$

HYPOTHESIS TESTING FOR THE PARAMETERS OF THE MODEL

$H_0: \beta_1 = \beta_{1(0)}$

$H_1: \beta_1 \neq \beta_{1(0)}$

with control function:

$$t_{(n-2)} = \frac{\hat{\beta}_1 - \beta_{1(0)}}{s(b_1)}$$

where $\beta_{1(0)}$ the price of β_1 under the null hypothesis.

In our example, $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$ the observed value of the test function equal to 6.831. This value should be compared to the critical value of t distribution with $(n-2)$= 6 degrees of freedom. Here the sample size used is equal to 8. Then the critical value for a = 0.05 in two-sided test, from tables of t distribution is ± *2447* and since 6.831> 2.447. Thus we conclude that we reject the null hypothesis at α=0.05. Note that the value of the observed level of significance (p-value) is very small, almost zero.

The corresponding test for the significance of the constant β_0 , given by most statistical packages, is of lesser importance compared to the one for the statistical significance of the parameter β_1. Here, the non-rejection of H_0 connected to the conclusion that when the independent variable X take the value 0, then the mean value of the dependent variable Y is 0. However, even if the test lead to the conclusion that, the constant term remains in the model, since though it's insignificant contribution, is not a problem for our fit.

In our example, $H_0: \beta_0 = 0$ versus $H_1: \beta_0 \neq 0$ the observed value of the control function equal to 11.472. which is much higher that the critical value 2,447. We thus reject the null hypothesis.

GOODNESS OF FIT CRITERIA

We have already presented a measure to evaluate the fit of the regression line to the observed pairs of observations of the two variables, i.e. the mean square error (MS_E).

The MS_E is an estimate of the dispersion of the errors (residuals) and a measure of observed data dispersion around the regression line. Consequently, the smaller the value, the better the fit. However as mentioned above the MS_E measured in the unit of our dependent variable and therefore the value is dependent on both the measure of the dependent variable, and the price levels of Y. It therefore cannot be used as a comparison goodness of fit criterion between different fits where the dependent variables are measured in different units of measurement, or when the dependent variable, although having the same unit, moves in substantially different price levels.

What we need then is a measure of the degree of data dispersion around the regression line, liberated by units. Such a measure will enable us to compare the results of different fits.

The measure looking should be a measure that compares the dispersion of Y around the regression line to the total dispersion Y. It can be shown that this measure is the square of the estimated correlation coefficient r, i.e. the coefficient of determination r^2.

The coefficient of determination r^2 is a descriptive measure of the goodness of fit, an indicator of how well the regression line describes the observed relation between Y and X. Moreover the coefficient of determination r^2 is an estimator of the corresponding population parameter ρ^2, Which is the square of the population correlation coefficient between two variables X and Y.

Let us now see how the coefficient of determination is defined.

$(y-\bar{y}) =$	$(y-\hat{y})$	$+$	$(\hat{y}-\bar{y})$
Total deviation	unexplained deviation		explained deviation

Considering now the squares of these three variations for each point y and taking the sums of those terms for all data points, after simple calculations we conclude:

$$\begin{array}{ccccc} \sum_{i=1}^{n}(y_i-\bar{y})^2 & = & \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 & + & \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\ SS_T & = & SS_E & + & SS_R \\ \text{total sum of squares} & & \text{sum of squares of the residuals} & & \text{sum of squares of regression} \end{array} \qquad (1.29)$$

The term SS_R the part of the variability of Y explained by the model, while SS_E is the sum of the square errors, i.e. the part of the variability Y which are not explained by the model, i.e. the dispersion of Y values that cannot be interpreted. The sum of these two terms gives us the overall dispersion in Y , i.e. SS_T .

We define the coefficient r^2 as the ratio of the regression sum of squares SS_R the total sum of squares SS_T .

$$r^2 = \frac{SS_R}{SS_T}$$

According to (1.29) , the sum of SS_E and SS_R give us SS_T. Thus r^2 can also be considered as

$$r^2 = 1 - \frac{SS_E}{SS_T}$$

The coefficient of determination r^2 is defined as the part of total dispersion of Y interpreted by the regression model.

ANALYSIS OF VARIANCE TABLE (ANOVA)

The ***Analysis of Variance Table*** or simply ANOVA Table is given below.

Table 1.2 Analysis of Variance table – ANOVA table

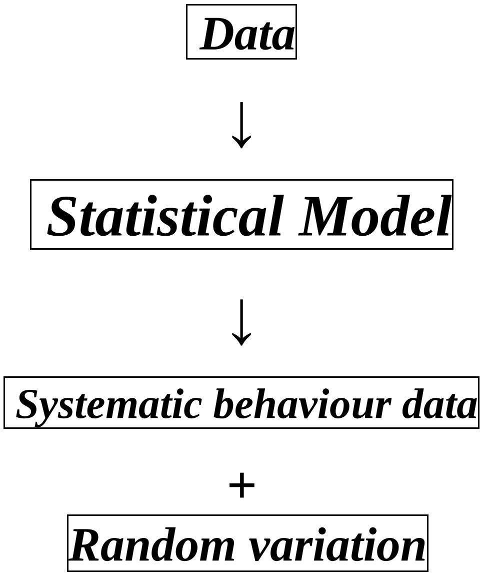
Source of Variation	***Sums of squares***	***Degrees of freedom***	***Mean sums of squares***	***F statistic***
Regression	$SS_R = \sum\left(\hat{Y} - \bar{Y}\right)^2$	1	$MS_R = \dfrac{SS_R}{1}$	$F_{(1,n-2)} = \dfrac{MS_R}{MS_E}$
Errors (Residuals)	$SS_E = \sum\left(Y - \hat{Y}\right)^2$	$n-2$	$MS_E = \dfrac{SS_E}{n-2}$	
Total	$SS_T = \sum\left(Y - \bar{Y}\right)^2$	$n-1$		

In our example we take

Table 1.3 Table Analysis of Variance - ANOVA

Source of Variation	***Sums of squares***	***Degreeso of freedom***	***Mean sums of squares***	***F value***	***p-value***
Regression	*SSR = ,974.014*	1	$MS_R = 1974.014$	$F_{(1,n-2)} = 46,656$	*0,000*
Residuals	*SSE = 253.861*	*6*	$MS_E = 42.310$		
Total	*SST = 227.875*	*7*			

$$\mathbf{F\ (1,\ n\text{-}2) = (t^{(n\text{-}2)})^{2}}$$



Data

↓

Statistical Model

↓

Systematic behaviour data

+

Random variation

Data behaviour has two components (systematic and random). The model describes the systematic behaviour of the data leaving out the contribution of the random agent (errors or residuals).

ASSECING THE ASSUMPTIONS OF THE RESIDUALS

- **INVESTIGATION OF THE INDEPENDENCE OF RESIDUALS**

- **EXPLORING THE CONSTAND VARIANCE OF THE RESIDUALS**

- **INVESTIGATION OF THE NORMALITY ASSUMPTION**

USING THE REGRESSION MODEL FOR PREDICTIONS

As mentioned in the first section of the chapter, there are several uses of the regression model. The first is to describe the relationship between two variables X and Y.

The ultimate goal of any statistical course of the investigation is the prediction. Using the equation $\hat{Y} = b_0 + b_1 X$, We are able to estimate (predict) the value of the dependent variable Y for any value of the independent variable X. This way we provide point estimates (forecasts) on our dependent variable. It should be noted that the projections should be made to the data area used in the estimation process. Using linear regression to draw conclusions out of the range of values X, Our estimate is highly uncertain and the risk of erroneous prediction becomes serious as the estimated relationship may not be appropriate outside the range of the independent variable used in our model.

POINT FORECASTS

The calculation of point forecasts using the estimated regression equation is very simple. We simply model the price X for which we want to predict Y And so calculate the predicted value of Y.

In our example, for children aged 4 years, using our model we get that the expected height is 109 cm,

$$\boxed{\hat{Y} = 66{,}972 + 10{,}472 \cdot 4 = 109}$$

PREDICTION INTERVALS

The interval degree prediction (1-a) 100% for Y when the $X = x$ is given

$$\boxed{\hat{y} \pm t_{[(n-2),1-\alpha/2]} \cdot s\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{SS_X}}}$$

In our example for $x = 4$, The 95%confidence interval for Y , is

$$109 \pm 2,447 \cdot \sqrt{42,3.(1+\frac{1}{8}+\frac{(4-3,5)^2}{18}} \quad = \quad 109 \pm 17,0 \quad = \quad [92\,;126]$$

It is evident that the ranges are dependent on the distance of that price X (for which we want to predict the Y) from the mean $\overline{X}$. The greater this distance, the larger the standard error of $\hat{Y}$, The larger the width of the confidence interval and hence the less effective our prediction.

CONFIDENCE INTERVALS OF THE EXPECTED VALUE OF Y FOR GIVEN X

The (1-α)100% confidence interval for the $E\left(Y \mid X\right)$ is :

$$\boxed{\hat{y} \pm t_{\left[(n-2),1-\alpha/2\right]} \cdot s\sqrt{\frac{1}{n}+\frac{(x-\overline{x})^2}{SS_X}}}$$

In our example the 95% confidence interval for the expected value of Y if X=4, is

$$\boxed{109 \pm 2,447 \cdot \sqrt{42,3.(\frac{1}{8}+\frac{(4-3,5)^2}{18}} \quad = \quad 109 \pm 5,9 \quad = \quad \left[103\,;115\right]}$$

MODEL BUILDING

1. Select the model

↓

2. Estimation of parameters using the data

↓

3. Model: Validation - residual analysis

If the model is not appropriate, then return to Step 1,
if appropriate, then

↓

4. Using the model for estimates and projections

Another example

Table 1.1 shows the annual total expenditure (Y) n thousands of euros, and the annual total net income (X) also in thousands of euros, for a random sample of twenty households.

<u>Table 1.1.</u>

Expenses (Y) And income (X)
a random sample of 20 households (in thousand euros).

Y ***Consumption*** ***(In th. Euros)***	***X*** ***Income*** ***(In th. Euros)***
5	5
6.5	5
6	6
5	7
6	5
10	8
9	9
8.5	9
6.5	10
8.5	10
9	11
12	12
11	13
11	14
14.5	15
14	16
12	15
16	14
3	6
10	10

Figure 1.2 shows graphically the relationship of observed values of two variables in this sample of twenty households.

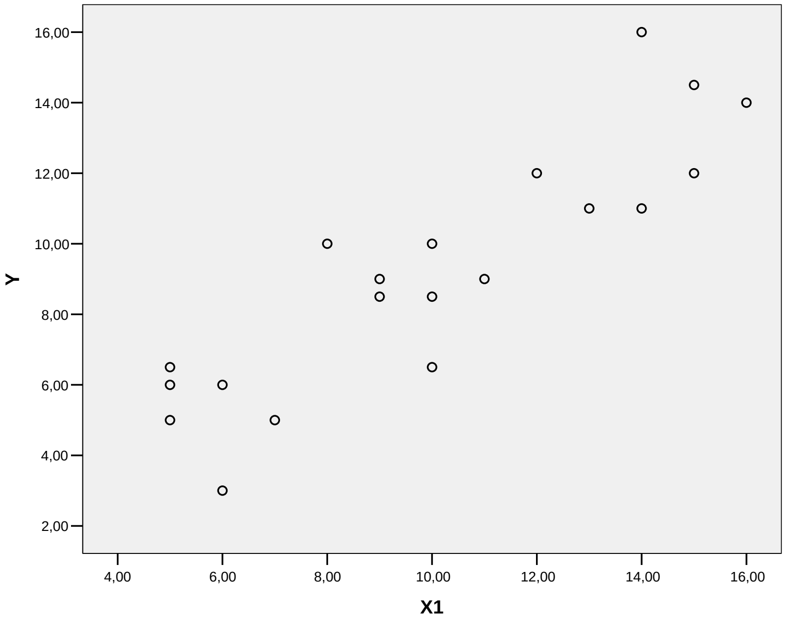


Figure 1.2: Relationship between expenditure (Y) And income (X)

We compute:

$$SS_X = \sum (x_i - \bar{x})^2 = \sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n} = 2254 - \frac{200^2}{20} = 254$$

$$SS_Y = \sum (y_i - \bar{y})^2 = \sum y_i^{\,2} - \frac{\left(\sum y_i\right)^2}{n} = 1914 - \frac{183,5^2}{20} = 230,6$$

$$SS_{XY} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 2049,5 - \frac{183,5 \cdot 200}{20} = 214,5$$

Having calculated the values of the sums of squares, we now from the relationship 1.14 to calculate through estimator $\hat{\beta}_0$ and $\hat{\beta}_1$, The point estimates of the parameters of the simple linear model (b_0 and b_1):

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_X} \Rightarrow b_1 = 0,844$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \Rightarrow b_0 = 0,73$$

So our model is:

$$\hat{Y} = 0{,}73 + 0{,}844 \cdot X$$

Based on the assessment of the fixed term of the model, the average consumption of zero-income households is estimated at 730 euros. Still, according to the estimate of the regression coefficient, an increase in annual household income of 1000 euros, the expected increase of the annual expenditure equal 844 euros.

Figure 1.3 is a plot of pairs of observations X and Y and the straight regression fitted using the least square method.

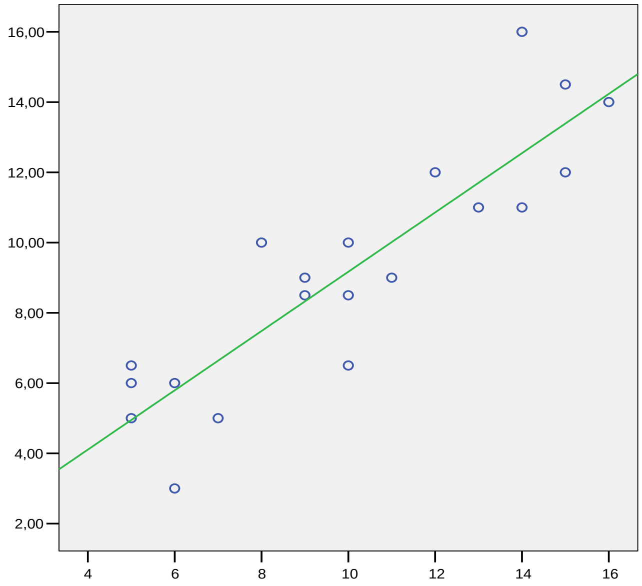


Figure 1.3: Simple linear model to describe the relationship between costs (Y) and income (X)

Let us then build the 95% confidence intervals for β_0 and β_1. Using the respective formulas we take

$$s\left(b_0\right) = \frac{s\sqrt{\sum x^2}}{\sqrt{nSS_X}} = \frac{1{,}858 \cdot \sqrt{2254}}{\sqrt{(20)\cdot(254)}} = 1{,}105$$

$$s\left(b_1\right)=\frac{s}{\sqrt{SS_X}}=\frac{1,658}{\sqrt{254}}=0,104$$

So the confidence interval for the β_0 is:

$$b_0 \pm t_{(n-2),1-\alpha/2}s\left(b_0\right)=0,730 \pm 2,093\cdot(1,105)=[-1,583\ ;\ 3,043]$$

where the value 2,093 is the taken from the table of t distribution, for $1-\alpha/2=0,975$ and $n-2=18$ degrees of freedom. So, with 95% confidence, the constant term of the regression model have a value between 1,583 and 3,043.

If the confidence interval includes the value 0, we can say that the constant term of the model is not statistically significant. This means that at zero X, the dependent variable Y gets a value not significantly different from 0. This is a reasonable in our case.

The 95% confidence interval for the β_1, is

$$b_1 \pm t_{(n-2),1-\alpha/2}s\left(b_1\right)=0,844 \pm 2,093\cdot(0,104)=[0,626\,;1,062]$$

So, based on our results, 95% of the population households, an increase of annual income in thousand euros expected to increase their annual costs by 626 to 1062 euros. If the confidence interval does not include the value 0 we can say that with 95% confidence, the regression coefficient is statistically significant and therefore our vision, that Y linearly related to the X is correct and therefore our model is appropriate.

$$SS_T = SS_Y = 230,64$$

$$SS_R = b_1 SS_{XY} = 0,844\cdot 214,5 = 181,143$$

$$SS_E = SS_T - SS_R = 49,495$$

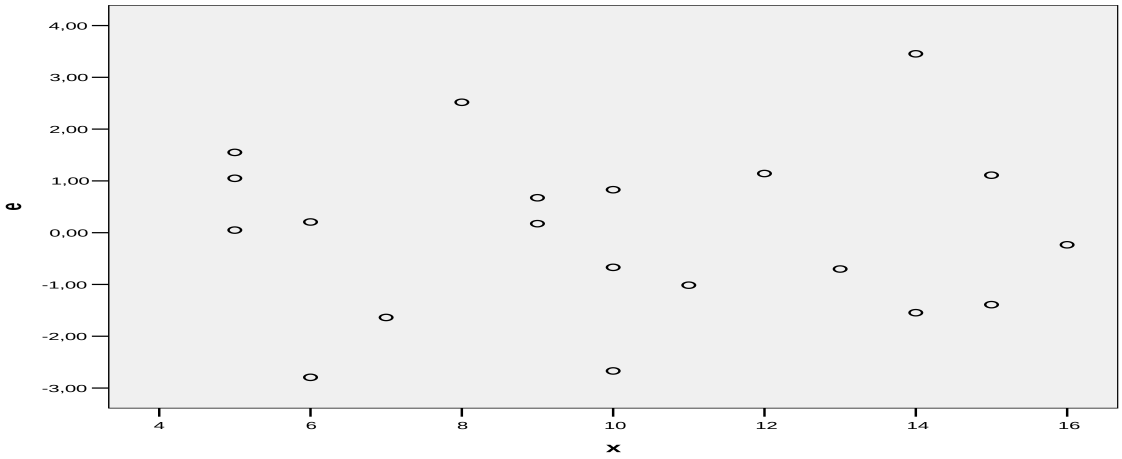
$$r^2 = SS_R / SS_T = 0,785$$

Table Analysis of Variance - ANOVA

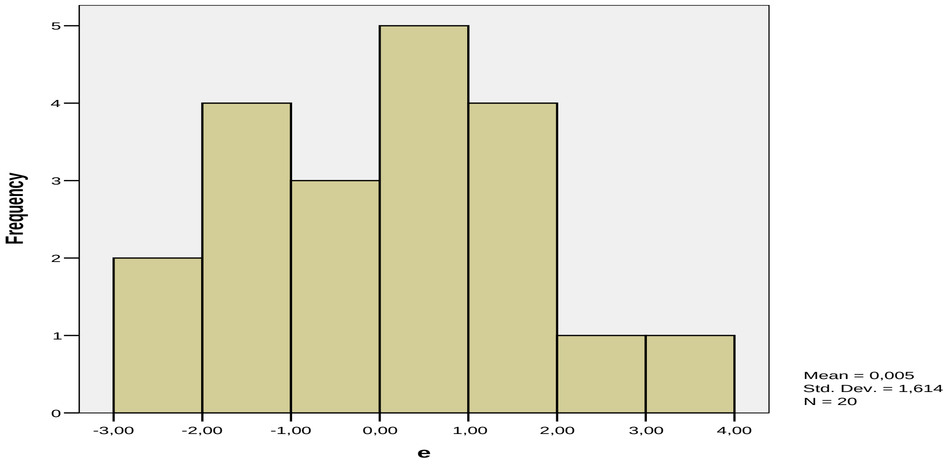
source of Variation	***sums of squares***	***Degrees of freedom***	***Through squares Sums***	***F value***
straight regression	*SSR = 181143*	1	$MS_R = 181.143$	$F_{(1,n-2)} = 65,877$
bugs (Remnants)	*SSE = 49,495*	*18*	$MS_E = 2,750$	
Total	*SST = 230 638*	*19*		

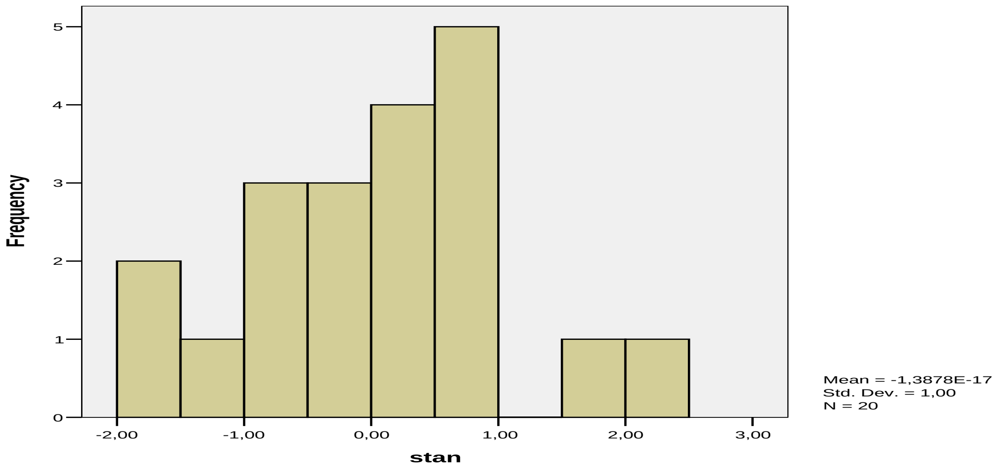
$F_{(1,18)}$ They are equal to 4.41

INVESTIGATION OF THE INDEPENDENCE OF RESIDUES



INVESTIGATION OF THE REGULARITY OF RESIDUES




Frequency
0
1
2
3
4
5
-2,00
-1,00
0,00
1,00
2,00
3,00
stan
Mean = -1,3878E-17
Std. Dev. = 1,00
N = 20

PREDICTIONS

In our example, if the annual income equal to 10 000 euro on the basis of our model and our calculations in this set of observations of (X,Y), the expected level of expenditure, equal to 9170 euros, since $\hat{Y} = 0,73 + 0,844 \cdot 10 = 9,17$.

For $x = 10$, the 95% prediction interval for Y, is

$$9,3 \pm 2,101 \cdot 1,658 \sqrt{1 + \frac{1}{20} + \frac{(10-10)^2}{254}} \quad = \quad 9,3 \pm 3,57 \quad = \quad [5,73\,;12,87]$$

So, based on the model and our calculations in this set of observations, in the population, the 95% of households with an annual income equal to 10 000 euros, have annual expenditures in the range between 5730 to 12 870 euros.

The 95% confidence interval for the expected value of Y for $x = 10$, is

$$9,3 \pm 2,101 \cdot 1,658 \sqrt{\frac{1}{20} + \frac{(10-10)^2}{254}} \quad = \quad 9,3 \pm 0,78 \quad = \quad [8,52\,;10,08]$$

So, based on the model and our calculations in this set of observations, the average annual consumption of households with an annual income equal to 10 000 euro in the population, with 95% confidence is expected to have a value between 8520 and 10080 euros.