2. MULTIPLE LINEAR REGRESSION

The theoretical multiple linear regression model with k independent variables is of the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \ ,$$

This model refers to the population. Y is the dependent variable, i.e. the variable that we wish to explain or predict. The variables $X_1, X_2, ..., X_k$ are the independent variables, $\beta_0, \beta_1, \beta_2, ..., \beta_k$ the model parameters and finally ε are residuals, the random factor in the model and therefore the only source of randomness in the behaviour of Y.

As in the case of simple linear regression, our model is based on some theoretical assumptions. These are:

1) for each value of Y, the residuals ε are independent variables that follow the normal distribution with expected value zero and constant variance. This means that $\varepsilon_i, \varepsilon_j$, for each $i \neq j$ are independent random variables,

$$\varepsilon_i \sim N(0, \sigma^2) \quad \text{for all } i = 1, 2, ..., n$$

2) the variables X_j take fixed values, while in correlation analysis, are random variables. In any case, X_j are independent of the error terms ε. Assuming that X_j take fixed values, we assume that these variables X_j are not random and so, the only randomness in Y result of the residuals ε.

For the case with $k = 2$ independent variables, the model becomes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

thus

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

These relationships are similar to those that apply in the case of simple linear regression. Here, instead of a straight regression, we have a plane. The model parameters can be estimated, as in the case of simple linear, using the method of least squares.

As with the simple linear model, now we want to have estimates of the model parameters that correspond to the minimum possible value of the sum of squares:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \quad .$$

The process can of course be extended to k independent variables. Where $k = 2$, there are three equations and their solutions are the least squares estimates b_0, b_1 and b_2. These are the estimates of the intercept term, the slope corresponding to X_1 and the slope corresponding to X_2.

The three equations for the case of two independent variables are:

$$\sum y = nb_0 + b_1 \sum x_1 + b_2 \sum x_2$$

$$\sum x_1 y = b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2$$

$$\sum x_2 y = b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2$$

The general model of multiple linear regression, contains the intercept β_0 and the k parameters are the regression coefficients for each of the k independent variables.

The value of each parameter $\beta_j,\ j=1,...,k$, Corresponding to the average change (increase if the sign is positive, or decrease if it is negative) of Y when the corresponding variable X_j is increased by one unit of measurement, while all the other independent variables remain unchanged.

A further assumption is that the variables X_j are uncorrelated with each other, so that the value of the estimator of each parameter $\beta_j,\ j=1,...,k$ to reflect the change in $E(Y)$ corresponding to an increase of the corresponding X_j by one unit (the unit of measurement), with all other independent variables being constant.

Conversely if some of the X_j **are lineary associated**, then the parameter estimates of these variables lose their meaning. This is a problem that will occupy us in detail in the next section.

The estimated regression relationship is of the form,

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k$$

where $\hat{Y}$ is the predicted value of Y the observed one. The parameter estimates $b_j,\ \ j=0,...,k$ are the values of the estimators $\hat{\beta}_j,\ j=0,1,...,k$ of the corresponding parameters $\beta_j,\ j=0,1,...,k$, as obtained by applying the least squares method.

So for the sample we have

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki} + e_i, \quad i = 1, 2, ..., n\ ,$$

EXAMPLE

Let us now return to the example of the previous chapter.

For the twenty households in our sample, Table 2.1 shows the annual expenditures (Y) in 1000 euros, the annual income (X_1), in 1000 euros, the number of members of the household (X_2), home ownership (if Yes = 1 if No = 0)(X_3) and the number of children living in the household (X_4). With the last two variables we will deal later.

We saw in Chapter 1 that the relationship between expenditures (Y) and income (X_1), of the 20 households in the sample, is strongly linear (correlation coefficient $r = 0,886$). Thus the simple linear model would be appropriate to describe their relationship.

Let us consider now, as another independent variable, the number of household members, and let's assess how suitable it would be a model that would show the annual expenditures (Y) of the household, as a linear function of two independent variables, the annual income (X_1) and the number of members (X_2) of the households.

Table 2.1: Expenditures (Y) , income (X_1), number of household members (X_2), residential property (X_3) and number of children (X_4), in a random sample of 20 households.

i	Y **Expenditures (in 1000 euros)**	X_1 **Income (in 1000 euros)**	X_2 **Menbers**	X_3 **Property residential**	X_4 **Children**
1	5	5	2	1	1
2	6.5	5	4	0	1
3	6	6	2	0	1
4	5	7	1	1	0
5	6	5	2	0	1
6	10	8	3	0	2
7	9	9	2	0	0
8	8.5	9	1	1	0
9	6.5	10	1	1	0
10	8.5	10	2	1	1
11	9	11	3	1	1
12	12	12	3	0	2
13	11	13	4	1	2
14	11	14	2	1	0
15	14.5	15	4	1	2
16	14	16	4	1	2
17	12	15	2	1	1
18	16	14	2	0	1
19	3	6	1	1	0
20	10	10	3	0	1

For the simple linear regression model, we have:

$$\hat{Y} = 0{,}73 + 0{,}844 \cdot X$$ *where X: income*

For the multiple linear regression model using X_1 and X_2 we have,

$$Y\wedge = -0,299 + 0,763X_1 + 0,770X_2$$

According to the values of the parameter estimates in the model, an income increase by 1,000 euros, for households of the same size, the expected (mean) increase of the annual expenditures is estimated to be 763 euros, while

An increase in household size by one member without changing the annual income is expected to increase the annual expenditure of 770 euros.

2.1. CHECKING THE SUITABILITY OF THE MODEL

A general way to access the suitability of the model is to investigate whether there is a linear regression relationship between the dependent variable and at least some of the explanatory independent variables X_i included in the regression model.

The statistical hypothesis test for a linear relationship between the Y
and at least some of the independent variables X_i *They are:*

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_k = 0 \qquad (2.9)$$

H_1: *at least one of the* $\beta_j (i = 1,...,k) \neq 0$

If the null hypothesis is not rejected, it is an indication that there is no linear relationship between the Y and anyone of the independent variables in the proposed regression relationship. In such a case, the model is clearly totally unsuitable. If on the other hand, the null hypothesis is rejected, there is statistical evidence linear relationship between Y and at least some of the independent variables included in the model.

To conduct this test is useful to calculate the ***Analysis of Variance table*** (ANOVA), which are of similar form to that of a simple linear regression as presented in the previous section, except that here we k independent variables instead of only one.

Let us now consider the deviation of each observed value from the corresponding assessment under the model, the error is, $y-\hat{y}$, The deviation of the estimate of the average of the observations Y, $\hat{y}-\overline{y}$ and the total deviation (the deviation of the observed price Y the average) $y-\overline{y}$. These three deviations satisfy the relation:

$$(y-\overline{y}) \;=\; (\hat{y}-\overline{y}) \;+\; (y-\hat{y})$$

As in the case of simple linear regression, when squaring the deviations and we sum all the data, we take the ratio of the sums of squares.

$$\sum (Y-\overline{Y})^2 = \sum (\hat{Y}-\overline{Y})^2 - \sum (Y-\hat{Y})^2$$
$$\Downarrow$$
$$SS_T = SS_R + SS_E$$

Where SS_T the total sum of squares, SS_R the sum of squares of regression and SS_E the sum of error squares of the errors (residuals).

In simple linear regression, the degrees of freedom of the errors were $n-2$ Because two parameters were estimated using our data (n). In multiple regression estimated we have $k+1$ parameters to be estimated. Therefore, the degrees of freedom for error are $n-(k+1)$. The degrees of freedom for the regression is k , while the total degrees of freedom is $n-1$. Table 2.2 Analysis of Variance Table (ANOVA) is presented for multiple regression model k independent variables.

Table 2.2 Table Analysis of Variance table for multiple regression

Source of Variation	***Sums of squares***	***Degrees of freedom***	***Mean Sums of squares***	***F value***
Regression	$SS_R = \sum (\hat{Y}-\overline{Y})^2$	k	$MS_R = \frac{SS_R}{k}$	$F_{(1,n-(k+1))} = \frac{MS_R}{MS_E}$
Residuals	$SS_E = \sum (Y-\hat{Y})^2$	*n-(k+1)*	$MS_E = \frac{SS_E}{n-(k+1)}$	
Total	$SS_T = \sum (Y-\overline{Y})^2$	$n-1$		

For our example the Analysis of Variance table is given in Table 2.3.

Table 2.3 Table Analysis of Variance of Example

Source of variation	***Sums of squares***	***Degrees of freedom***	***Mean sums of squares***	***F***	***P-value***
Regression	191.762	2	95.881	41,92	0,000
Residuals	38	17	2287		
Total	230.638	19			

Taking p-value = 0,000 , we reject the null hypothesis. So, we conclude that there is a linear relationship between the indicator Y and at least one of the two independent variables.

Note that this test is a primary control, indicative of a relationship between the dependent and at least one of the independent variables included in our model. If H_0 is rejected, then we need to conduct separate checks for the importance of each individual parameter in the final model.

In multiple regression, additional checks are needed to determine the statistically significant of the parameters. These tests are indicative of the contribution of each independent variable to explain the dispersion of the values of our dependent variable. Based on these tests, many of the independent variables show no statistically significant explanatory effect and thus these should be excluded from the regression model. Before we proceed with these individual checks, let us present some goodness of fit criteria of the multiple regression model.

2.2. EVALUATION OF THE MODEL

The mean square error (MSE) is an unbiased estimator of the population variance of errors ε, symbolized by σ_ε^2. The mean square error is defined in equation (2.11):

The mean square error is:

$$MS_E = \frac{SS_E}{n-(k+1)} = \frac{\sum_{j=1}^{n}(y_i - \hat{y}_j)^2}{n-(k+1)} \qquad (2.11)$$

It is already known that the smaller the error, the better the fit of the regression model. The mean square error (MSE) is a measure of goodness of fit of our model. Still, the square root of MSE is an estimator of the standard deviation of error σ_ε. The square root of the MSE is usually denoted by s_ε or simply s and referred to as the standard error of the estimate.

The standard error of the estimate is:

$$s_\varepsilon = \sqrt{MS_E} \qquad (2.12)$$

The s_ε is a goodness of fit criterion (the smaller its value, the better the fit of the model to our data). However it has the major disadvantage that it is measured in units of our dependent variable. So it is a measure that is inappropriate for comparisons between different applications. An alternative goodness of fit criterion of the regression model is the coefficient of determination, denoted as R^2. This is defined as

$$R^2 = SS_R / SS_T = 1 - SS_E / SS_T$$

The coefficient of determination R^2 is a very useful criterion for goodness of fit of the regression model. However, it has some limitations. It can be shown that, for each data set, with increasing the number of independent variables in the model the value of R^2 also increases. This is natural since the greater the number of independent variables in the model, the more the surface approaches the regression data. Since then the adjustment of the multiple regression model becomes better as the number of independent variables increases, R^2 growing and approaching unity or 100% of the variability of Y explaining the model. Thus, the coefficient becomes unsuitable for comparisons between models with different number of independent variables, since every time another independent variable is added to the model, its

value will be increased regardless of whether the contribution of this additional variable is significant or not.

A measure suitable for a comparison between models involving different number of independent variables is the adjusted (or corrected) determination coefficieny factor. The adjusted coefficient of determination denoted by $\bar{R}^2$ is the takes into consideration the degrees of freedom.

The adjusted coefficient of determination is defined as:

$$\bar{R}^2 = 1 - \frac{SSE/[n-(k+1)]}{SST/(n-1)} = 1 - \frac{MS_E}{MS_T} \qquad (2.13)$$

This measure, as seen from the structure, not always grow as additional variables are introduced into the model. If $\bar{R}^2$ grows as we introduce a new independent variable in the model, this is an indication that its presence in the model is appropriate.

The corrected coefficient of determination can still be calculated on the basis of the simple coefficient of determination R^2 by:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-(k+1)} \qquad (2.14)$$

Note that a new variable should be included in the model only if the $\bar{R}^2$ increases.

Among the various multiple regression models with different number of independent variables, the model that minimizes the MSE and maximizes $\bar{R}^2$ is the best. The use of two criteria, and MSE and $\bar{R}^2$, to introduce new independent variables in the regression model would be further discussed in the next section.

In our example, $R^2 = 0,961$, Which means that 96.1% of the total variability of annual household expenditures attributes to their linear relationship with income and household size. Still $\bar{R}^2 = 0,95$, a value that is very close to the non-adjusted rate. So we conclude that the model fits the data very well. Though not yet know if the two independent variables are significant. This will be investigated conducting

individual tests for the significance of each parameters. These tests are presented below.

2.3. TESTING THE SIGNIFICANCE OF THE MODEL PARAMETERS

Until now, we investigated the suitability of the model. using the test statistic F. We also saw how to evaluate the regression model using the coefficient of determination and the adjusted one. It remains to be seen, how can we evaluate whether each one of parameters β_j are significant, or in other words, how important is the existence of each one of the X_j in the model. A test for the significance of the individual parameters of the regression are useful because it investigates whether the variable under examination X_j has explanatory capacity on the dependent variable.

In the last section, we saw that an indication of the benefits we get from the introduction of a particular variable in the regression equation, resulting from the comparison of the adjusted coefficient of determination, which includes the variable of interest, with the corresponding of that model in which that variable is not included. Here we introduce a statistical test for the significance of each parameter β_j.

In Chapter 1 we saw that the hypothesis testing:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

may be carried out using or controlling $t \ \left(= b_1 / s(b_1)\right)$ or control $F \ \left(= MS_R / MS_E\right)$. These two tests proved equivalent. In simple linear regression, there is only one parameter, β_1. And if it is zero there is no linear regression relationship. In multiple regression but where $k > 1$ t these two tests are not equivalent. The F test investigates whether statistically there is significant linear relationship between the Y and at least one independent variable X_j, while the t tests explore the significance of each individual independent variable $X_j, \quad j = 1, 2, ..., k$ including in the model.

In a regression model Y with k independent variables $X_1, X_2, ..., X_k$, there are k significance tests for the slopes $\beta_1, \beta_2, ..., \beta_k$.

If the null hypothesis is true, then the test function following distribution t with $n-(k+1)$ degrees of freedom.

Individual testing of hypothesis of the regression slope is of the form:

(1) $H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$
(2) $H_0 : \beta_2 = 0$
$H_1 : \beta_2 \neq 0$
$\vdots \quad \vdots$
(k) $H_0 : \beta_k = 0$
$H_1 : \beta_k \neq 0$

The control function is of the form $\boxed{\dfrac{\hat{\beta}_j}{s(\hat{\beta}_j)}}$,

By analogy with hypothesis testing confidence intervals for the values of each parameters of our model can be calculated.

The confidence intervals for the parameters β_j, $j = 1, 2, ..., k$ it is easy to calculate using the estimator and the standard error for each of the parameters.

The degree of confidence interval (1-a) 100% for the β_j *They are*

$$\hat{\beta}_j \pm t_{(n-(k+1),1-\alpha/2)} s\left(\hat{\beta}_j\right)$$

If a 95% confidence interval for the parameter β_j contains zero, then the two-sided hypothesis testing $H_0 : \beta_j = 0$ is not rejected and thus there is no indication that the variable X_j having a linear relationship with the Y.

MULTICOLLINEARITY

A problem may arise in drawing conclusions for the regression model parameters are as follows: In multiple regression, we need to have a strong correlation between each one of the independent variables and the dependent

variable Y. But in addition, we do not want the independent variables to be significantly correlated with each other.

When two independent variables are highly correlated with each other, then these variables behave similarly, and consequently they explain the behaviour of Y in a similar way. In such a case, we would say that the one variable “steals" the explanatory power of the other.

This problem is called **multicollinearity**.

Many problems can arise from this confusion. A serious consequence of this problem is that the estimators of the model parameters exhibit high variability, the standard errors are unusually large, making the parameter estimates statistically insignificant. The multicollinearity can also lead to signs of some parameter be opposite to those expected.

In order to avoid multicollinearity, we should avoid to include in the model explanatory variables that are significantly correlated to each other.