TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Quantitative Methods III - Practice 2
Multiple Linear Regression

Prof. Lorenzo Cavallo: lorenzo.cavallo.480084@uniroma2.eu
Prof. Marianna Brunetti: marianna.brunetti@uniroma2.it

Exercise The following measurements have been obtained in a study:

y_i	x_{1i}	x_{2i}	x_{3i}
10	5	8	-0.5
30	20	2	0.4
20	5	9	-0.2
25	15	11	1
15	10	10	0.3

Assuming a multiple linear regression model: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$

1. Obtain the matrices needed to estimate the parameters $\beta_0, \beta_1, \beta_2$
2. Estimate the coefficients of the regression line and describe the relationships between variables, knowing that

$$(X'X)^{-1} = \begin{bmatrix} 4.488 & -0.158 & -0.318 \\ -0.158 & 0.008 & 0.008 \\ -0.318 & 0.008 & 0.028 \end{bmatrix}$$

3. Predict the value of y when $x_1 = 5$ and $x_2 = 8$
4. Calculate the right goodness-of-fit indices of the estimated line to the observed data and comment on the results

Solution

1. We known that

$$\hat{\beta}_{OLS} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_k \end{bmatrix} = (X'X)^{-1}(X'Y)$$

Therefore, we start by defining the matrices X, $(X'X)$ and $(X'y)$.
The X matrix contains the data of the independent variables x_{1i}, x_{2i}, x_{3i} (also called "predictors", "regressors", or "explanatory variables") preceded by the constant, i.e.:

$$X = \begin{bmatrix} 1 & 5 & 8 \\ 1 & 20 & 2 \\ 1 & 5 & 9 \\ 1 & 15 & 11 \\ 1 & 10 & 10 \end{bmatrix}$$

The vector Y instead contains the data relating to the dependent variable. Therefore, in this example it will be:

$$Y = \begin{bmatrix} 10 \\ 30 \\ 20 \\ 25 \\ 15 \end{bmatrix}$$

By applying the rules for matrix calculus it is possible to obtain

$$(X'X) = \begin{bmatrix} 5 & 55 & 40 \\ 55 & 775 & 390 \\ 40 & 390 & 370 \end{bmatrix}$$

and

$$(X'Y) = \begin{bmatrix} 100 \\ 1275 \\ 745 \end{bmatrix}$$

Knowing that $det(X'X) = 30000$ and that therefore the matrix is non-singular, we can proceed to obtain the inverse:

$$(X'X)^{-1} = \begin{bmatrix} 4.488 & -0.158 & -0.318 \\ -0.158 & 0.008 & 0.008 \\ -0.318 & 0.008 & 0.028 \end{bmatrix}$$

2. Using the calculations of the previous point, we can obtain the estimators $\hat{\beta}_{OLS}$:

$$\hat{\beta}_{OLS} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = (X'X)^{-1}(X'Y) = \begin{bmatrix} 4.488 & -0.158 & -0.318 \\ -0.158 & 0.008 & 0.008 \\ -0.318 & 0.008 & 0.028 \end{bmatrix} \begin{bmatrix} 100 \\ 1275 \\ 745 \end{bmatrix} = \begin{bmatrix} 9.8 \\ 1 \\ -0.1 \end{bmatrix}$$

   The estimated model is then:

$$\hat{y}_i = 9.8 + 1x_{1i} - 0.1x_{2i}$$

   The estimated coefficients tell us that the variable x_1 has a positive effect on y, while the coefficient of x_2 correlates negatively with the dependent variable. In particular, it is estimated that as the variable x_2 increases by one unit, the dependent variable Y undergoes a reduction equal to 0.1

3. The required predicted value is:

$$\hat{Y} = 9.8 + 1 \times 5 - 0.1 \times 8 = 14$$

4. The goodness-of-fit indices generally used to evaluate multiple linear regression models are the SER and the Adjusted-R^2 ($\bar{R}^2$).

   The formulas of both indices are recalled below, where with $\hat{u}_i$ we indicate the residuals and with k the number of predictors included in the model (including the constant):

$$SER = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k}}$$

$$\bar{R}^2 = 1 - \frac{n-1}{n-k}\frac{\sum_{i=1}^{n} \hat{u}_i^2}{TSS}$$

   Following the same methodology used in the previous point, it is possible to obtain the predicted values $\hat{Y}_i$ and therefore the residuals $\hat{u}_i = Y_i - \hat{Y}_i$ of the model.

Y_i	$(Y_i - \bar{Y})^2$	$\hat{Y}_i$	$u_i = (Y_i - \hat{Y}_i)$	u_i^2
10	100	14	-4	16
30	100	29.6	0.4	0.16
20	0	13.9	6.1	37.21
25	25	23.7	1.3	1.69
15	25	18.8	-3.8	14.44
100	250	100	0	69.5

The first part of the table also shows the calculations necessary to obtain $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

The dataset used has 5 observations per variable, so $n = 5$. Furthermore, the estimated model uses two variables, in addition to the constant, hence $k = 3$.

It follows that:

$$SER = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k}} = 5.89$$

and that the Adjusted-R^2 ($\bar{R}^2$) is equal to:

$$\bar{R}^2 = 1 - \frac{5-1}{5-3} \times \frac{69.5}{250} = 0.444$$

The value assumed by the index denotes a poor adaptation of the regression line to the observed points.
Indeed, only 44.4% of the variability of Y is explained by the two regressors considered.