TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Multiple Linear Regression

Quantitative Methods III

1 Problem Statement

We want to examine how two independent variables — **Education** (measured in years of schooling) and **Experience** (measured in years of work experience) — affect the dependent variable **Income** (annual income in thousands of dollars).

2 Data Description

We consider a small dataset with 5 observations for simplicity. The variables are defined as follows:

- **Income (Y)**: Annual income (in thousands of dollars)
- **Education ($\mathbf{X_1}$)**: Years of schooling
- **Experience ($\mathbf{X_2}$)**: Years of work experience

The dataset is:

Observation	Income (Y)	Education (X_1)	Experience (X_2)
1	40	10	5
2	50	12	7
3	65	14	10
4	55	12	8
5	80	16	15

3 Model Specification

We assume a linear relationship between the variables, which can be modeled as:

$$\text{Income} = \beta_0 + \beta_1 \times \text{Education} + \beta_2 \times \text{Experience} + u$$

where:

- β_0 is the intercept,
- β_1 is the coefficient for Education,
- β_2 is the coefficient for Experience,
- u is the error term.

The goal is to estimate the coefficients β_0, β_1, and β_2 using the Ordinary Least Squares (OLS) method.

4 Estimation via OLS

Traditional Resolution (Formulas)

In matrix notation, the model is written as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where $\mathbf{X}$ is the design matrix that includes a column of ones (for the intercept) and columns for Education and Experience.

The OLS estimator is given by:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

This formula minimizes the sum of squared residuals:

$$\min_{\beta_0, \beta_1, \beta_2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2$$

In practice, we compute $\mathbf{X}^\top \mathbf{X}$, invert it, and then multiply by $\mathbf{X}^\top \mathbf{Y}$ to obtain the estimated coefficients.

The matrix X (including the intercept) is

$$X = \begin{pmatrix} 1 & 10 & 5 \\ 1 & 12 & 7 \\ 1 & 14 & 10 \\ 1 & 12 & 8 \\ 1 & 16 & 15 \end{pmatrix}$$

Matrix multiplication: the rows of the first matrix are multiplied by the columns of the second one.

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 10 & 12 & 14 & 12 & 16 \\ 5 & 7 & 10 & 8 & 15 \end{pmatrix} \cdot \begin{pmatrix} 1 & 10 & 5 \\ 1 & 12 & 7 \\ 1 & 14 & 10 \\ 1 & 12 & 8 \\ 1 & 16 & 15 \end{pmatrix} =$$

$$\begin{pmatrix} 1\cdot1+1\cdot1+1\cdot1+1\cdot1+1\cdot1 & 1\cdot10+1\cdot12+1\cdot14+1\cdot12+1\cdot16 & 1\cdot5+1\cdot7+1\cdot10+1\cdot8+1\cdot15 \\ 10\cdot1+12\cdot1+14\cdot1+12\cdot1+16\cdot1 & 10\cdot10+12\cdot12+14\cdot14+12\cdot12+16\cdot16 & 10\cdot5+12\cdot7+14\cdot10+12\cdot8+16\cdot15 \\ 5\cdot1+7\cdot1+10\cdot1+8\cdot1+15\cdot1 & 5\cdot10+7\cdot12+10\cdot14+8\cdot12+15\cdot16 & 5\cdot5+7\cdot7+10\cdot10+8\cdot8+15\cdot15 \end{pmatrix} =$$

$$\begin{pmatrix} 5 & 64 & 45 \\ 64 & 840 & 610 \\ 45 & 610 & 463 \end{pmatrix}$$

and its transpose is

$$X^\top = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 10 & 12 & 14 & 12 & 16 \\ 5 & 7 & 10 & 8 & 15 \end{pmatrix}$$

The vector Y is

$$Y = \begin{pmatrix} 40 \\ 50 \\ 65 \\ 55 \\ 80 \end{pmatrix}$$

Multiplying $X^\top$ by X gives

$$X^\top X = \begin{pmatrix} 5 & 64 & 45 \\ 64 & 840 & 610 \\ 45 & 610 & 463 \end{pmatrix}$$

After computing the determinant (which is 252), the inverse is given by

$$(X^\top X)^{-1} = \begin{pmatrix} 66.75 & -8.66 & 4.92 \\ -8.66 & 1.15 & -0.67 \\ 4.92 & -0.67 & 0.41 \end{pmatrix}$$

The estimator is computed as

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

First, we compute

$$X^\top Y = \begin{pmatrix} 40 + 50 + 65 + 55 + 80 \\ 10 \cdot 40 + 12 \cdot 50 + 14 \cdot 65 + 12 \cdot 55 + 16 \cdot 80 \\ 5 \cdot 40 + 7 \cdot 50 + 10 \cdot 65 + 8 \cdot 55 + 15 \cdot 80 \end{pmatrix} = \begin{pmatrix} 290 \\ 3850 \\ 2840 \end{pmatrix}$$

Multiplying, we obtain

$$\hat{\beta} = \begin{pmatrix} 66.75 & -8.66 & 4.92 \\ -8.66 & 1.15 & -0.67 \\ 4.92 & -0.67 & 0.41 \end{pmatrix} \begin{pmatrix} 290 \\ 3850 \\ 2840 \end{pmatrix}$$

Carrying out the multiplications step by step:

$$\hat{\beta}_0 = \left(66.75 \cdot 290 - 8.66 \cdot 3850 + 4.92 \cdot 2840\right) = -5.16,$$

$$\hat{\beta}_1 = \left(-8.66 \cdot 290 + 1.15 \cdot 3850 - 0.67 \cdot 2840\right) = 3.65,$$

$$\hat{\beta}_2 = \left(4.92 \cdot 290 - 0.67 \cdot 3850 + 0.41 \cdot 2840\right) = 1.83$$

Thus, the vector of estimated coefficients is

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} -5.16 \\ 3.65 \\ 1.83 \end{pmatrix}$$

5 Estimation Using R

Below is an R script that creates the dataset and estimates the regression model using the `lm` function.

```r
# Step 1: Create the dataset
data <- data.frame(
  Income = c(40, 50, 65, 55, 80),
  Education = c(10, 12, 14, 12, 16),
  Experience = c(5, 7, 10, 8, 15)
)

# Step 2: Fit the multiple linear regression model
model <- lm(Income ~ Education + Experience, data = data)

# Step 3: Display the summary of the model
summary(model)
```

Running this script in R will provide you with the estimated coefficients, standard errors, t-statistics, p-values, and the overall goodness-of-fit (R-squared).

```
Call:
lm(formula = Income ~ Education + Experience, data = data)

Residuals:
      1       2       3       4       5 
-0.4762 -1.4286  0.7937  1.7460 -0.6349 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5.159     14.557  -0.354    0.757
Education      3.651      1.911   1.910    0.196
Experience     1.825      1.145   1.595    0.252

Residual standard error: 1.782 on 2 degrees of freedom
Multiple R-squared:  0.9932,	Adjusted R-squared:  0.9863 
F-statistic: 145.5 on 2 and 2 DF,  p-value: 0.006827
```

6 Interpreting the Results

Suppose the R output gives the following estimated coefficients:

- **Intercept** ($\hat{\beta}_0$): -5.16
- **Education** ($\hat{\beta}_1$): 3.65
- **Experience** ($\hat{\beta}_2$): 1.83

Interpretation:

- **Intercept** ($\hat{\beta}_0 = -5.16$): This is the predicted income when both Education and Experience are zero.
- **Education** ($\hat{\beta}_1 = 3.65$): Holding Experience constant, an additional year of education is associated with an increase of approximately 3.65 thousand dollars in annual income.
- **Experience** ($\hat{\beta}_2 = 1.83$): Holding Education constant, an additional year of work experience is associated with an increase of approximately 1.83 thousand dollars in annual income.

7 Model Evaluation and Diagnostics

After estimating the model, it is important to:

- **Assess the goodness-of-fit:** Check the R-squared value to see how well the independent variables explain the variability in Income.
- **Evaluate significance:** Look at the p-values associated with each coefficient to determine if they are statistically significant.
- **Perform diagnostic checks:** Plot residuals to assess if the assumptions of linearity, homoscedasticity, independence, and normality of errors are satisfied.