Introductory Econometrics
Lecture 2: Introduction to linear regression model
Suggested Solution

by Hieu Nguyen

Fall 2024

1.

Derive the 'summation formula' for the OLS estimator $\hat{\beta}_1$ in the linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Solution: Our Simple Regression Model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

which can be rearranged as:

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$$

Given a sample of data (x_i, y_i) where $i = 1, ..., n$, we can try to estimate β_0 and β_1 by minimising the sum of squared residual:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Find estimates of (β_0, β_1) to minimise the sum of squared residual:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{argmin} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

To get the value of β_0 and β_1 that makes the $\sum_{i=1}^{n} \epsilon_i^2$ minimum, we can take derivative with respect to each of them. We get:

First order condition with respect to $[\beta_0]$

$$\begin{aligned}
-2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\
\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\
\sum_{i=1}^{n} y_i - n \cdot \hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i &= 0 \\
\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i &= n \cdot \hat{\beta}_0 \\
\frac{1}{n} \cdot \sum_{i=1}^{n} y_i - \frac{1}{n} \cdot \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \hat{\beta}_0 \\
\bar{y} - \hat{\beta}_1 \bar{x} &= \hat{\beta}_0 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned}$$

First order condition with respect to $[\beta_1]$

$$\begin{aligned}
2 \cdot \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot (-x_i) &= 0 \\
\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot x_i &= 0 \\
\sum_{i=1}^{n} (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) \cdot x_i &= 0 \\
\sum_{i=1}^{n} (y_i - \bar{y}) x_i + \hat{\beta}_1 (\bar{x} - x_i) \cdot x_i &= 0 \\
\sum_{i=1}^{n} (y_i - \bar{y}) x_i + \hat{\beta}_1 \sum_{i=1}^{n} (\bar{x} - x_i) \cdot x_i &= 0 \\
\hat{\beta}_1 \sum_{i=1}^{n} (\bar{x} - x_i) \cdot x_i &= - \sum_{i=1}^{n} (y_i - \bar{y}) x_i \\
-\hat{\beta}_1 \sum_{i=1}^{n} (\bar{x} - x_i) \cdot x_i &= \sum_{i=1}^{n} (y_i - \bar{y}) x_i \\
\hat{\beta}_1 \sum_{i=1}^{n} (-1) \cdot (\bar{x} - x_i) \cdot x_i &= \sum_{i=1}^{n} (y_i - \bar{y}) x_i \\
\hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x}) \cdot x_i &= \sum_{i=1}^{n} (y_i - \bar{y}) x_i \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (y_i - \bar{y}) \cdot x_i}{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot x_i}
\end{aligned}$$

Because we have:

$$\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} (y_i x_i - y_i \bar{x} - \bar{y} x_i + \bar{y} \bar{x})$$

$$= \sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i - \bar{y} \sum_{i=1}^{n} x_i + n \cdot \bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - \bar{x} \cdot n \cdot \frac{1}{n} \sum_{i=1}^{n} y_i - \bar{y} \sum_{i=1}^{n} x_i + n \cdot \bar{x}\bar{y}$$

$$= \sum_{i=1}^{n} x_i y_i - \bar{x} \cdot n \cdot \bar{y} - \bar{y} \sum_{i=1}^{n} x_i + n \cdot \bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y}$$

$$= \sum_{i=1}^{n} (x_i y_i - x_i \bar{y}) = \sum_{i=1}^{n} x_i (y_i - \bar{y})$$

The same way, we also have:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}(x_i-\bar{x})\cdot(x_i-\bar{x}) = \sum_{i=1}^{n}(x_i^2-\bar{x}\cdot x_i-\bar{x}\cdot x_i+\bar{x}^2)\\
&= \sum_{i=1}^{n}(x_i^2-2\bar{x}\cdot x_i+\bar{x}^2) = \sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}^2 = \sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+n\cdot\bar{x}^2\\
&= \sum_{i=1}^{n}x_i^2-2\bar{x}\cdot n\cdot\frac{1}{n}\sum_{i=1}^{n}x_i+n\cdot\bar{x}^2 = \sum_{i=1}^{n}x_i^2-2\bar{x}\cdot n\cdot\bar{x}+n\cdot\bar{x}^2 = \sum_{i=1}^{n}x_i^2-2n\cdot\bar{x}^2+n\cdot\bar{x}^2\\
&= \sum_{i=1}^{n}x_i^2-n\cdot\bar{x}^2 = \sum_{i=1}^{n}x_i^2-n\cdot\bar{x}\cdot(\frac{1}{n}\cdot\sum_{i=1}^{n}x_i) = \sum_{i=1}^{n}x_i^2-n\cdot\bar{x}\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}x_i^2-\bar{x}\cdot\sum_{i=1}^{n}x_i\\
&= \sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}\bar{x}\cdot x_i = \sum_{i=1}^{n}(x_i^2-\bar{x}\cdot x_i) = \sum_{i=1}^{n}(x_i\cdot x_i-\bar{x}\cdot x_i) = \sum_{i=1}^{n}x_i(x_i-\bar{x})\end{aligned}$$

If we use the above two equations into the result of $\hat{\beta}_1$, we will have:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})\cdot x_i}{\sum_{i=1}^{n}(x_i-\bar{x})\cdot x_i} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})\cdot(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(x_i-\bar{x})} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})\cdot(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

The above function is as similar as calculating sample covariance of y_i and x_i, divided by sample variance of x_i. It is easier to remember.

2.

Suppose you are a university director considering canceling the entrance exams and searching for new criteria for selecting good students for your school. Assume you consider using the grades from high school (in the US called the American College Testing [ACT] scores, 1–36).

To decide, some data about current students seem to be helpful. Let us say you have a sample of 15 students. The following table contains the ACT scores and the GPA (Grade Point Average, 0–4) for these 15 students. Grade Point Average represents the student's performance at the university; it is based on a four-point scale and has been rounded to one digit after the decimal.

Student	GPA	ACT
1	2.8	21
2	3.4	24
3	3.0	26
4	3.5	27
5	3.6	29
6	3.0	25
7	2.7	25
8	3.7	30
9	3.2	23
10	3.0	28
11	3.5	30
12	2.5	20
13	3.8	32
14	2.6	26
15	2.7	23

(a)

Estimate the relationship between GPA and ACT using OLS; that is, obtain the intercept and slope estimates in the equation: GPA $= \beta_0 + \beta_1$ACT $+ \varepsilon$. Use Excel for the computation:

1. Compute the intercept and slope estimates using the summation formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$.
2. Compute the intercept and slope estimates using the matrix formula for $\hat{\beta}$.

Solution: (a) Please see the attached Excel file. You might want to check out Excel commands `=MMULT()` and `=MINVERSE()`.

We use formulas:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

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

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Resulting estimated equation (rounded): $\widehat{\text{GPA}} = 0.67 + 0.09\text{ACT}$.

(b)

Comment on the direction of the relationship. Does the intercept have a meaningful interpretation here? Explain. How much higher is the GPA predicted if the ACT scores increase by 5 points?

Solution: The positive $\hat{\beta}_1 = 0.09$ suggests a positive relationship between the ACT scores and GPA. The higher the ACT scores are, the higher the predicted GPA is. The interpretation of $\hat{\beta}_1 = 0.09$ is that an increase of the ACT scores by 1 unit (1 point) is associated with by 0.09 higher GPA.

In this regression example, the estimated intercept $\hat{\beta}_0 = 0.67$ does not have any useful interpretation or meaning because it is impossible to have ACT $= 0$ in reality (the scale is 1–36). Mathematically, our estimation predicts $\hat{\text{GPA}} = 0.67$ in case ACT $= 0$. The 'issue' of including β_0 into the regression model equation and interpreting the estimated intercept $\hat{\beta}_0$ was discussed in detail during the seminar.

The impact of an increase of the ACT scores by 5 points is an increase of GPA by $5 \cdot 0.09 = 0.45$.

(c)

Find and list the fitted values and the residuals of the model.

Solution: Please, see the attached Excel file again. We use formulas from the lecture #2 slides:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

(d)

What is the predicted value of GPA when ACT $= 20$?

Solution: When ACT $= 20$, the predicted $\hat{\text{GPA}} = 0.67 + 0.09 \cdot 20 = 2.47$.

3. BONUS EXERCISE (very similar to 2.) to take home and master OLS in Excel:

This exercise illustrates the crucial distinction between the stochastic error term and the residual. Usually, we can never observe the error term, but we can get around this difficulty if we assume values for the true parameters. Calculate values of the error term and the residual for each of the following six observations given that the true β_0 equals 0, the true β_1 equals 1.5.

	y_i	x_i
1	2	1
2	6	4
3	3	2
4	8	5
5	5	3
6	4	4

Solution: Please see the attached Excel file Seminar 02 JEM062 solution.xlsx, lists 3 and 4. You may need Excel commands `=MMULT()` and `=MINVERSE()` again.

We use formulas:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i,$$

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

Resulting estimated equation (rounded): $\hat{y} = 1.32 + 0.48x$.