Regression 3: Discussions on OLS Assumptions

Instructor: Yuta Toyama

Last updated: 2020-03-30

Section 1

Introduction

Introduction

- Remember that we have four assumptions in OLS estimation

1. Random sample: $\{Y_i, X_{i1}, \dots, X_{iK}\}$ is i.i.d. drawn sample - i.i.d.: identically and independently distributed
2. ϵ_i has zero conditional mean

$$E[\epsilon_i | X_{i1}, \dots, X_{iK}] = 0$$

   - This implies $Cov(X_{ik}, \epsilon_i) = 0$ for all k. (or $E[\epsilon_i X_{ik}] = 0$)
   - **No correlation between error term and explanatory variables.**
3. Large outliers are unlikely:
   - The random variable Y_i and X_{ik} have finite fourth moments.
4. No perfect multicollinearity:
   - There is no linear relationship betwen explanatory variables.

- The OLS estimator has ideal properties (consistency, asymptotic normality, unbiasdness) under these assumptions.
- In this chapter, we study the role of these assumptions.
- In particular, we focus on the following two assumptions
  1. No correlation between ϵ_{it} and X_{ik}
  2. No perfect multicollinearity

Section 2

Endogeneity

Endogeneity problem

- When $Cov(x_k, \epsilon) = 0$ does not hold, we have **endogeneity problem**
  - We call such x_k an **endogenous variable**.
- There are several cases in which we have endogeneity problem
  1. Omitted variable bias
  2. Measurement error
  3. Simultaneity
  4. Sample selection
- Here, I focus on the omitted variable bias.

Omitted variable bias

- Consider the wage regression equation (true model)

$$\begin{aligned} \log W_i = & \beta_0 + \beta_1 S_i + \beta_2 A_i + u_i \\ E[u_i|S_i, A_i] = & 0 \end{aligned}$$

  where W_i is wage, S_i is the years of schooling, and A_i is the ability.
- What we want to know is β_1, the effect of the schooling on the wage **holding other things fixed**. Also called the returns from education.
- An issue is that we do not often observe the ability of a person directly.

- Suppose that you omit A_i and run the following regression instead.

$$\log W_i = \alpha_0 + \alpha_1 S_i + v_i$$

  - Notice that $v_i = \beta_2 A_i + u_i$, so that S_i and v_i is likely to be correlated.
- The OLS estimator $\hat{\alpha}_1$ will have the bias:

$$E[\hat{\alpha}_1] = \beta_1 + \beta_2 \frac{Cov(S_i, A_i)}{Var(S_i)}$$

- You can also say $\hat{\alpha}_1$ is not consistent for β_1, i.e.,

$$\hat{\alpha}_1 \xrightarrow{p} \beta_1 + \beta_2 \frac{Cov(S_i, A_i)}{Var(S_i)}$$

- This is known as **omitted variable bias formula**.
- Omitted variable bias depends on
  1. The effect of the omitted variable (A_i here) on the dependent variable: β_2
  2. Correlation between the omitted variable and the explanatory variable.
- This is super-important: You can make a guess regarding the direction and the magnitude of the bias!!
- This is crucial when you read an empirical paper and do am empirical exercise.
- Here is the summary table
  - x_1: included, x_2 omitted. β_2 is the coefficient on x_2.

	$Cov(x_1, x_2) > 0$	$Cov(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Correlation v.s. Causality

- Omitted variable bias is related to a well-known argument of "Correlation or Causality".
- Example: Does the education indeed affect your wage, or the unobserved ability affects both the ducation and the wage, leading to correlation between education and wage?
- See my lecture note from Intermediate Seminar (Fall 2018) for the details.

Section 3

Multicollinearity issue

Perfect Multicollinearity

- If one of the explanatory variables is a linear combination of other variables, we have perfect multicolinearity.
- In this case, you cannot estimate all the coefficients.
- For example,

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 \cdot x_2 + \epsilon_i$$

  and $x_2 = 2x_1$.
- These explanatory variables are collinear. You are not able to estimate both β_1 and β_2.
- To see this, the above model can be written as

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 \cdot 2x_1 + \epsilon_i$$

- this is the same as

$$y_i = \beta_0 + (\beta_1 + 2\beta_2) x_1 + \epsilon_i$$

- You can estimate the composite term $\beta_1 + 2\beta_2$ as a coefficient on x_1, but not β_1 and β_2 separately.

Some Intuition

- Intuitively speaking, the regression coefficients are estimated by capturing how the variation of the explanatory variable x affects the variation of the dependent variable y
- Since x_1 and x_2 are moving together completely, we cannot say how much the variation of y is due to x_1 or x_2, so that β_1 and β_2.

Example: Dummy variable

- Consider the dummy variables that indicate male and famale.

$$male_i = \begin{cases} 1 & if \ \ male \\ 0 & if \ \ female \end{cases}, \ female_i = \begin{cases} 1 & if \ \ female \\ 0 & if \ \ male \end{cases}$$

- If you put both male and female dummies into the regression,

$$y_i = \beta_0 + \beta_1 famale_i + \beta_2 male_i + \epsilon_i$$

- Since $male_i + famale_i = 1$ for all i, we have perfect multicolinarity.

- You should always omit the dummy variable of one of the groups in the linear regression.
- For example,

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

- In this case, β_1 is interpreted as the effect of being famale **in comparison with male**.
  - The omitted group is the basis for the comparison.

- You should the same thing when you deal with multiple groups such as

$$freshman_i = \begin{cases} 1 & if \ \ freshman \\ 0 & otherwise \end{cases}$$

$$sophomore_i = \begin{cases} 1 & if \ \ sophomore \\ 0 & otherwise \end{cases}$$

$$junior_i = \begin{cases} 1 & if \ \ junior \\ 0 & otherwise \end{cases}$$

$$senior_i = \begin{cases} 1 & if \ \ senior \\ 0 & otherwise \end{cases}$$

and

$$y_i = \beta_0 + \beta_1 freshman_i + \beta_2 sophomore_i + \beta_3 junior_i + \epsilon_i$$

Imperfect multicollinearity.

- Though not perfectly co-linear, the correlation between explanatory variables might be very high, which we call imperfect multicollinearity.
- How does this affect the OLS estimator?
- To see this, we consider the following simple model (with homoskedasticity)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, V(\epsilon_i) = \sigma^2$$

- You can show that the conditional variance (not asymptotic variance) is given by

$$V(\hat{\beta}_1|X) = \frac{\sigma^2}{N \cdot \hat{V}(x_{1i}) \cdot (1 - R_1^2)}$$

where $\hat{V}(x_{1i})$ is the sample variance

$$\hat{V}(x_{1i}) = \frac{1}{N} \sum (x_{1i} - \bar{x_1})^2$$

and R_1^2 is the R-squared in the following regression of x_2 on x_1.

$$x_{1i} = \pi_0 + \pi_1 x_{2i} + u_i$$

- You can see that the variance of the OLS estimator $\hat{\beta}_1$ is small if
  1. N is large (i.e., more observations!)
  2. $\hat{V}(x_{1i})$ is large (more variation in x_{1i}!)
  3. R_1^2 is small.
- Here, high R_1^2 means that x_{1i} is explained well by other variables in a linear way. g– The extreme case is $R_1^2 = 1$, that is x_{1i} is the linear combination of other variables, implying perfect multicolinearity!!

Lesson for an empirical analysis

- We often say **the variation of the variable of interest is important in an empirical analysis**.
- This has two meanings:
  1. **exogenous** variation (i.e., uncorrelated with error term)
  2. large variance
- The former is a key for **mean independence assumption**.
- The latter is a key for **precise estimation** (smaller standard error).

- If we have more variation, the standard error of the OLS estimator is small, meaning that we can precisely estimate the coefficient.
- The variation of the variable **after controlling for other factors that affects** y is also crucial (corresponding to $1 - R_1^2$ above).
  - If you do not include other variables (say x_2 above), you will have omitted variable bias.
- To address research questions using data, it is important to find a good variation of the explanatory variable that you want to focus on. This is often called **identification strategy**.
  - Identification strategy is context-specific. To have a good identification strategy, you should be familiar with the background knowledge of your study.