Regression 3: Discussion on OLS Assumptions

Instructor: Yuta Toyama

Last updated: 2021-05-18

Introduction

Roles of OLS Assumptions and How to defend them

- Assumption 2: ϵ_i has zero conditional mean $E[\epsilon_i|X_{i1},\ldots,X_{iK}]=0$
  - 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.**
- Assumption 4: No perfect multicollinearity
- We need to think whether these assumptions are valid given the research setting.
- Question
  - What if these assumptions are vilated?
  - How to defend these assumptions?

Contents

Takeaway for Causal Analysis

- Suppose that you want to know the causal effect of D on Y in the following linear model

$$y_i = \alpha_0 + \alpha_1 D_i + \beta' x_i$$

- **The variation of the variable of interest D** is important in the following senses.

- 1: **Exogenous** variation after conditioning on x_i
  - i.e., uncorrelated with error term
  - **mean independence assumption** (no bias)

- 2: **Enough** variation after conditioning on x_i
  - a key for **precise estimation** (smaller standard error)
  - related to multicolinearity

Endogeneity

Endogeneity problem

- When $Cov(x_k, \epsilon) = 0$ does not hold, we have **endogeneity problem (内生性問題)**
  - We call such x_k an **endogenous variable (内生変数)**.
- There are several cases in which we have endogeneity problem

1. Omitted variable bias (欠落変数バイアス)
2. Measurement error (観測誤差)
3. Simultaneity (同時性)
4. Sample selection (サンプルセレクション)

Omitted Variable Bias (欠落変数バイアス)

- 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.

- β_1: the effect of the schooling on the wage **holding other things fixed**.
- Issue: 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.

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

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

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.

- 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

- Can discuss the direction of the bias

Summary: Exogeneity of X

- Mean independence is a key for unbiased estimation.

- However, this is hard to argue, as we have to discuss about **unobserved** factors.

- Moreover, there is **no formal test for exogeneity assumption**.
  - Question: Examine the correlation between the residual $\hat{\epsilon}_i$ and explanatory variables X_{ik} . Would this work?

- How to avoid this issue?
  - 1: Add control variables
  - 2: Natural experiment

Adding Control Variables

- Consider the model with an interest in α_1.

$$y_i = \alpha_0 + \alpha_1 D_i + \beta' x_i + \epsilon_i,\ E[\epsilon_i | D_i, x_i] = 0$$

- Idea: Adding more variables into x means
  - controlling for the factors that are correlated with treatment variable D_i.
  - avoiding omitted variables
  - mean independence assumption of D_i and ϵ_i more likely hold.

- Should we add variables as much as possible? Not necesarily.
  - Issue 1: More controls lead to less precise estimation. See this later.
  - Issue 2: **Bad control problem**

Bad Control Problem

- Consider the model

$$wage_i = \alpha_0 + \alpha_1 college_i + \alpha_2 occupation_i + \epsilon_i,\ E[\epsilon_i | D_i, x_i] = 0$$

- You are interested in α_i, effects of going to a college on wage **after controlling for occupation**.
- However, occupation is certainly affected by college choice.
- The estimated α_1 cannot capture the effect of attending a college on wage through occupation choice.
- Here, the variable $occupation_i$ is called a **bad control**. You should not include this to estimate α_1.

A Guidance on Variable Choice

	Affect y_i	**Not affect y_i**
Affect X_i or simultaneously determined with X_i	Must to avoid omitted variable bias.	Should not include, as it increases the variance. But the bias does not change.
X_i affects the variable	No. Bad control problem	(same as above)
Not correlated with X_i	Should include to decrease the variance. But no bias even without it.	(same as above)

Natural Experiment (自然実験)

- Natural experiment refers to the situation where **the variable of interest is determined randomly as if it were in experiment**.
- It is however not the actual experiment.
- Some examples:
  - Weather
  - Policy assignment is often determined by lottery (e.g., millitary draft)
  - Birth-related events (twin, exact date, etc)
  - and many more!!
- Economists put effort to find such situation to establish causal estimates.

Multicollinearity issue

Perfect Multicollinearity

- Perfect multicolinearity: One of the explanatory variable is a linear combination of other variables.
- 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 $x_2 = 2x_1$.
- Cannot 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.

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.
- 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.

Multiple Dummy Variables

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

$$
\begin{aligned}
freshman_i &= \begin{cases} 1 & if\ freshman \\ 0 & otherwise \end{cases} \\
sophomore_i &= \begin{cases} 1 & if\ sophomore \\ 0 & otherwise \end{cases} \\
junior_i &= \begin{cases} 1 & if\ junior \\ 0 & otherwise \end{cases} \\
senior_i &= \begin{cases} 1 & if\ senior \\ 0 & otherwise \end{cases}
\end{aligned}
$$

and

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

Imperfect Multicollinearity.

- Imperfect Multicollinearity: Correlation between explanatory variables is high.
- Although we can estimate the model by OLS, it affects the precision of the estimate, that is standard errors.
- To see this, we consider the following simple model (with homoskedasticity)

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

Sampling variance of the OLS Estimator

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

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

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

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

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

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

Four factors that decrease the variance.

1. N is large
2. $\hat{V}(x_{1i})$ is large
    - more variation in x_{1i}!
3. R_1^2 is small.
    - R_1^2 measures how well x_{1i} is explained by other variables in a linear way.
    - The extreme case is $R_1^2 = 1$ (i.e., x_{1i} is the linear combination of other variables)
4. Smaller variance of the error term σ^2.
    - This reflects how much the variation of y_i is explained.
    - More control variables lead to lower variance of the error term.
    - But remember the above point 3!!

Summary: Enough variation of X.

- With more variation in X, can precisely estimate the coefficient.
- The variation of the variable **after controlling for other factors** is also crucial
- If you include many control variables to deal with the omitted variable bias, you may end up having no independent variation of X.

Robustness Analysis

How to defend your analysis? Robustness Analysis

- Exogeneity assumption (mean independence assumption) is hard to argue.
- In a good empirical analysis, do **robustness analysis (頑健性分析)** to see how robust your results are against concerns.
- Deryugina "Some Tips For Robustness Checks And Empirical Analysis In General" provides an overview.
- Two major appraoches
  - **Sensitivity analysis (感度分析)** for control variables.
  - **Placebo test (プラシーボテスト)** -> See this in an empirical application.

Sensitivity Analysis

- Step 1: Consider a specification of the model with control variables that you think are reasonable and estimate it. (baseline specification)
- Step 2: Add additional controls to the above and re-estimate it.
- Step 3: See how the estimated coefficient of interest (typically treatment variable) changes. If it does not change that much, your result is robust (or endogeneity concern is small).

Why is this a good way to discuss exogeneity?

- The concern on exogeneity is the correlation between D_i and the error term.
- If you add control variables and the estimated coefficient does not change, it **suggests that the effect of omitted variables are likely to be small.**
- However, this procedure is not formal. It is rather a practical technique.
- See Altonji, Elder, and Taber (2005) and Oster (2019) for a more formal discussion.