Introductory Econometrics I

Econometrics Basics: A Review

Yingjie Feng

School of Economics and Management

Tsinghua University

April 19, 2024

Main Tasks

- Usually, there are three tasks in standard econometric analysis
  - Identification
    - What quantities are you interested in?
  - Estimation
    - What method do you want to use to estimate the quantity of interest?
  - Inference
    - How do you characterize the uncertainty of the estimate?
- We will review them using the example in Q1 of Problem Set 1

Identification

- Regression on a binary variable

$$y = \beta_0 + \beta_1 x + u, \quad x \in \{0, 1\}$$

- Identification: How do you interpret β_0 and β_1
  - Identification is discussed at the *population* level.
  - This is related to the key **zero conditional mean assumption**

$$\mathbb{E}[u|x] = 0$$

  - In this case

$$\mathbb{E}[u|x=0] = \mathbb{E}[u|x=1] = 0$$
$$\Leftrightarrow \quad \beta_0 = \mathbb{E}[y|x=0], \quad \beta_1 = \mathbb{E}[y|x=1] - \mathbb{E}[y|x=0]$$

  - But maybe this is not what we want. The model of interest is

$$y = \beta_0' + \beta_1' x + u', \quad \mathbb{E}[u'|x] \neq 0$$

Identification

- Regression on a binary variable

$$y = \beta_0 + \beta_1 x + u, \quad x \in \{0, 1\}$$

- Identification: How do you interpret β_0 and β_1
  - Example: effectiveness of a vaccine
  - Treatment: $x = 1$ if vaccinated; $x = 0$ if not
    - Treatment group *vs* Control group
  - Outcome: $y = 1$ if infected with Covid; $y = 0$ if not
  - Do you want to directly compare $\mathbb{E}[y|x = 1]$ and $\mathbb{E}[y|x = 0]$?
    - Remember we want a causal interpretation!
    - Gold standard: randomized experiment
    - Observational data: think about the vaccine assignment

Identification

- Regression on a binary variable

$$y = \beta_0 + \beta_1 x + u, \quad x \in \{0, 1\}$$

- Identification: How do you interpret β_0 and β_1
  - Example: effectiveness of a vaccine
  - The discussion can be formalized in a **potential outcome** framework
  - The notion of *cause* and *effect* will be precisely defined by comparing *potential* outcomes
  - We will come back to this example at the end of Chapter 7

Identification

- Regression on a binary variable

$$y = \beta_0 + \beta_1 x + u, \quad x \in \{0, 1\}$$

- Identification: How do you interpret β_0 and β_1
  - Note: precisely speaking, the "no perfect collinearity" assumption is also a key condition for identification
    - It relates to the question, "Are the parameters uniquely determined by population distributions?"
    - Think about the moment condition of OLS

$$\mathbb{E}[y - \beta_0 - \beta_1 x] = 0, \qquad \mathbb{E}[x(y - \beta_0 - \beta_1 x)] = 0$$

Estimation

- Regression on a binary variable

$$y = \beta_0 + \beta_1 x + u, \quad x \in \{0, 1\}$$

- Estimation: how you estimate β_0 and β_1
  - Now we need a sample $\{(y_i, x_i) : 1 \leq i \leq n\}$ (assume i.i.d.)
  - You have shown OLS gives the following estimators

$$\hat{\beta}_1 = \bar{y}_1 - \bar{y}_0, \quad \hat{\beta}_0 = \bar{y}_0$$
$$\bar{y}_1 = \frac{1}{n_1} \sum_{i:x_i=1} y_i, \quad \bar{y}_0 = \frac{1}{n_0} \sum_{i:x_i=0} y_i$$

  - You have shown some nice properties such as

$$\mathbb{E}[\hat{\beta}_1] = \beta_1 \quad (\text{"unbiasedness"}), \qquad \hat{\beta}_1 \to_{\mathbb{P}} \beta_1 \quad (\text{"consistency"})$$

    - ★ What conditions do you need?

Inference/Uncertainty Quantification

- Regression on a binary variable

$$y = \beta_0 + \beta_1 x + u, \quad x \in \{0, 1\}$$

- Inference: How do you characterize the uncertainty of $\hat{\beta}_0$ and $\hat{\beta}_1$?
  - Variance: measure of "variability" of the estimator

$$\mathbb{V}[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{(assumes } \mathbb{V}[u|x] = \sigma^2\text{)}$$

  - Standard error: estimate of $\sqrt{\mathbb{V}[\hat{\beta}_1]}$

$$se(\hat{\beta}_1) = \hat{\sigma} \Big/ \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{1/2}$$

Inference/Uncertainty Quantification

- Regression on a binary variable

$$y = \beta_0 + \beta_1 x + u, \quad x \in \{0, 1\}$$

- Inference: How do you characterize the uncertainty of $\hat{\beta}_0$ and $\hat{\beta}_1$?
  - Even better, we characterize the whole distribution of the estimator

$$\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{n-2} \qquad \text{(assume } u|x \sim \mathsf{N}(0, \sigma^2))$$

  - Or in large samples ($n \to \infty$)

$$\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \to_d \mathsf{N}(0, 1)$$

Inference/Uncertainty Quantification

- Regression on a binary variable

$$y = \beta_0 + \beta_1 x + u, \quad x \in \{0, 1\}$$

- Inference: How do you characterize the uncertainty of $\hat{\beta}_0$ and $\hat{\beta}_1$?
  - Then we can construct confidence intervals, e.g.,

$$\mathbb{P}\Big(\beta_1 \in [\hat{\beta}_1 - 1.96 \cdot se(\hat{\beta}_1),\ \hat{\beta}_1 + 1.96 \cdot se(\hat{\beta}_1)]\Big) \approx 95\%$$

  - Or test hypotheses, e.g.,

$$\mathrm{H}_0 : \beta_1 = 0 \quad vs \quad \mathrm{H}_1 : \beta_1 \neq 0$$

    - ★ Find a critical value c, reject H_0 if $|t\text{-stat}| > c$
    - ★ Other tests (one-sided tests, joint hypothesis tests) are also possible

What is next?

- OLS further issues
  - Units of measurements, functional forms, etc.
- Regression with qualitative information
  - We will also give a brief introduction to program evaluation
- Deviate from Gauss-Markov Assumptions
  - Relax MLR.5 (Homoskedasiticity): heteroskedasticity
  - Relax MLR.2 (Random sampling): serial correlation or within-cluster correlation
  - Relax MLR.4 (Zero conditional mean): instrumental variable
  - Relax MLR.1 (Linear in parameters): nonlinear models