ECOM90025 Advanced Data Analysis

Tutorial 6

Zheng Fan

The University of Melbourne

Introduction

Zheng Fan

- Ph.D. student in Economics
- Email: fan.z@unimelb.edu.au
- Tutorial code and slides: github.com/zhengf1/ADA2022

Seek help?

- Ed discussion board
- Consultations: refer to Canvas for details

TE

Consider the treatment d takes value of 0 and 1. A simple model for a well-randomized experiment generates data y. So we observe independent pairs (y_i, d_i) for individual i. In addition, the allocation of d_i is purely random.

- Use n_0 (n_1) as the number untreated (treated) individuals.
- Define μ_1 and μ_0 as the mean of the treated and untreated outcomes, respectively.
- Define σ_1 and σ_0 as the standard deviation of the treated and untreated outcomes, respectively.

1. What is the mean and variance of the sample mean of the (un)treated outcomes?
2. What is the true value of the treatment effect (TE)?
3. What is the estimator for TE?
4. What is the variance of the estimator?
5. How to compute the standard error of the estimator?

Simulation

Assume that $\mu_0 = 0$, $\mu_1 = 1$, $\sigma_0 = 1$ and $\sigma_1 = 2$. Randomly draw $n_0 = 100$ and $n_1 = 50$ observations. Assume that the distribution of y_i is normal.

1. Simulate a data set and estimate the TE.
2. Carry out a Monte Carlo study to compare the simulated standard deviation to the theory.

Heterogeneity and ATE

The simulation study has two kinds of individuals, male and female. Use $g_i = 1$ for individual i being female and 0 for male.

Assume that

$$\mu_0^m = 0 \quad \mu_1^m = 1 \quad \mu_0^f = 0 \quad \mu_1^f = 2$$

where superscript ^{m} and ^{f} means male and female, respectively.

Randomly draw $n_0 = 100$ and $n_1 = 50$ observations (half female and half male). Assume that the distribution of y_i is normal. For simplicity, assume that the variance of each category is fixed at 1 (homoskedasticity).

Use Monte Carlo study to plot a histogram of TE

Regression for ATE: DGP

Simulate a data set from the previous question.

1. Your data should have 150 rows (observations $n_0 + n_1$) and 3 columns (y_i, d_i, g_i).
2. Half of g_i in the control and treatment groups is zero.
3. The data should be in the 'R' data frame format.

The data is constructed to be used in a regression to estimate the ATE

Regression Model

The DGP in Section **Heterogeneity and ATE** can be represented in a regression form as

$$y = \alpha_g + \gamma_g d + \epsilon$$

where $\sim N(0, 1)$.

- What is the ATE?

Regression Model

If, instead, we estimate a simple regression model

$$y = \alpha + \gamma d + u$$

while the true model has heterogeneity in gender.

Assume the true model is $y = \alpha_g + \gamma_g d + \epsilon$

- What is γ? Is it still ATE?

Regression Model

If, instead, we estimate a simple regression model

$$y = \alpha + \gamma d + u$$

while the true model has heterogeneity in gender.

Assume the true model is $y = \alpha_g + \gamma_g d + \epsilon$

- What is γ? Is it still ATE?

We can know

$$u = \epsilon + (\alpha_g - \alpha) + (\gamma_g - \gamma)d$$

$E(y \mid d = 1) - E(y \mid d = 0) = E(\alpha + \gamma + u) - E(\alpha + u) = \gamma.$
So, yes, γ is ATE.

Regression Model

If, instead, we estimate a simple regression model instead

$$y = \alpha_g + \gamma d + u$$

while the true model has heterogeneity in gender.

Assume the true model is $y = \alpha_g + \gamma_g d + \epsilon$

- What is γ? Is it still ATE?

Regression Model

If, instead, we estimate a simple regression model instead

$$y = \alpha_g + \gamma d + u$$

while the true model has heterogeneity in gender.

Assume the true model is $y = \alpha_g + \gamma_g d + \epsilon$

- What is γ? Is it still ATE?

We can know

$$u = \epsilon + (\gamma_g - \gamma)d$$

Similar to before. ATE is still γ.

Regression Model

Estimate three models

1. $y = \alpha + \gamma d + u$
2. $y = \alpha_g + \gamma d + u$
3. $y = \alpha_g + \gamma_g d + u$

And compute the ATE.

The end

Thanks for your attention! 

Tutorial code and slides: github.com/zhengf1/ADA2022