Recitation 1: Understanding Stata and Randomized Control Trials

Seung-hun Lee

Columbia University
Undergraduate Introduction to Econometrics Recitation

September 14th, 2021

Logistics of the recitation

Recitation Logistics

- Location: 602 Northwest Corner Building
- Time: Thursdays 1PM-2PM
  - 30-40 minutes will be spent on reviewing materials from the lecture, the rest will be spent on Stata demonstrations.
  - Pending room availability, I will stay an extra 10-20 minutes to answer your questions
  - Recitation notes to be posted by noon on Thursdays for you to download.
  - Slides will be posted AFTER the recitation (before midnight on Thursdays).
- Office Hours
  - Zoom (Click here to join) and Lehman 327 (So technically hybrid)
- Further materials
  - You can go here for my old recitation materials

Econometrics and RCT

What is econometrics trying to achieve?

- Econometrics is a field in economics that tries to answer real life questions.
  - Ultimately, it is about making a **quantitative** statement about two or more random events
  - We can make **correlational** statements, but we want to identify *causal relationships*

- In order to achieve this goal, we collect data from a suitably defined population and use various methods to estimate a parameter that implies correlational/causal relationship.

- To fully understand what econometrics is trying to achieve, we need to ask ourselves these three questions
  - What is the difference between correlational and causal relation?
  - What is the suitably defined population?
  - What are the methods that we need to use in econometrics?

Correlation vs Causation

- Suppose you have two random variables X and Y. You want to identify if X *causes* Y
- A **correlation** between X and Y is a statistical measure that describes how the two variables move together
  - It captures *any* type of statistical dependence that moves the two variables together: Causation, but also others too!
  - Not causal 1: Y can cause X
  - Not causal 2: X and Y are jointly moving because there is Z that affects both
- A **causal** relationship: *Cleanly (exogenously)* changing variables X leads to changes in Y
  - Much more difficult: Changes in X may be a combination of many things
  - Changes from X alone and changes from other factors that may indirectly affect X
  - RCT: Isolates clean changes in X that can help us tell whether changes in X affects Y, and by how much
  - Econometrics: We can express concisely the relationship between X and Y variables in a single equation.

Suitably defined population?

- When we say we are interested in the relationship between schooling and wages, whose effects are we interested in?
  - The entire US population, high school graduates, or college graduates?
  - Determines sampling methods we use to obtain a representative and comparable sample
  - Complete randomization, stratified randomization, or cluster randomization

- Note: We are almost surely never going to get the data from the entire population.
  - Gathering data from the entire population is logistically (and maybe ethically) difficult.
  - The estimate we are obtaining through any econometric exercise is thus a *sample analogue* of the actual value we are trying to get
  - We will do diagnostic tests to see if they can be reasonably close to the true value.

What methods?

- In econometrics, we will use many estimation methods to obtain the sample analogue of the parameter of interest.
  - Ordinary least squares (OLS): Suitable for randomized control trials or in any case where the treatment assignment is as good as random.
  - Panel estimation: If data has multiple individuals and multiple time periods.
  - Instrumental variable methods: When we have proxy variables relevant to variable of interest and is reasonably exogenous
  - Difference-in-differences: When we study ‘before & after’ events with multiple entities
  - Regression discontinuity: In treatment with a cutoff determining treatment assignment
  - Time series: When we observe one entity over multiple periods

- Depending on the type of variables we use in our exercise, we have:
  - Univariate regression: One variable (besides an overall constant) controlled for
  - Multivariate regression: Multiple variables (besides an overall constant) controlled for
  - Nonlinear regression: Binary dependent variables
  - Big data methods

Understanding RCTs

- In **randomized control trials**, we randomly categorize some individuals under treatment group and controlled group and run various tests
  - Benchmark for good program evaluation

- Potential outcomes framework
  - Y_i : Observed outcome for individual $i \in \{1, ..., N\}$
  - i: Either in treatment or control group (not both)$\rightarrow$ $W_i = 1$ if i is treated, 0 if otherwise
  - **W**: an N-tuple vector of treatment assignment for all individuals
  - Key assumption: Others' treatment assignment has no effect on my treatment (stable unit treatment value assumption (SUTVA))
  - Potential outcome $Y_i(w)$: Outcome for treated ($Y_i(1)$) and the untreated individual ($Y_i(0)$)
    - ★ Fundamental problem of missing data: Individual i cannot have both $Y_i(1)$ and $Y_i(0)$ - at most one of them

Potential vs observed outcome

- We can bridge the two with this relation

$$
\begin{aligned}
Y_i &= Y_i(1)W_i + Y_i(0)(1 - W_i) \\
&= Y_i(0) + W_i(Y_i(1) - Y_i(0))
\end{aligned}
$$

  - We know W_i and Y_i for everyone regardless of treatment assignment
  - We cannot see $Y_i(0)$ for the treated group and $Y_i(1)$ for the untreated group

Treatment effect

- If we want to see if the treatment has any effect, we would ideally see

$$Y_i(1) - Y_i(0)$$

- But they cannot be obtained b/c fundamental problem of missing data
  - Alternative is average treatment effect or average treatment effect on the treated

$$\text{ATE} = E[Y_i(1) - Y_i(0)]$$
$$\text{ATT} = E[Y_i(1) - Y_i(0)|W_i = 1]$$

  - We also need assumptions about our treatment: Randomized assignment is one of them

$$E[Y_i(1)] = E[Y_i(1)|W_i = 1] = E[Y_i(1)|W_i = 0]$$
$$E[Y_i(0)] = E[Y_i(0)|W_i = 1] = E[Y_i(0)|W_i = 0]$$

Obtaining ATE

- With this assumption and the definition of potential outcomes framework

$$
\begin{aligned}
E[Y_i|W_i] &= E[Y_i(1)W_i + Y_i(0)(1-W_i)|W_i] \\
&= E[Y_i(1)|W_i]W_i + E[Y_i(0)|W_i](1-W_i) \\
&= E[Y_i(1)]W_i + E[Y_i(0)](1-W_i)
\end{aligned}
$$

  - $W_i = 1$: Get $E[Y_i(1)] = E[Y_i|W_i = 1]$.
  - $W_i = 0$: Get $E[Y_i(0)] = E[Y_i|W_i = 0]$.

- Under random assignment, we can identify the average treatment effect as

$$
\text{ATE} = E[Y_i(1)] - E[Y_i(0)] = E[Y_i|W_i = 1] - E[Y_i|W_i = 0]
$$

- Econometrically: OLS with W_i as independent variable (Problem set 2!)