Introduction to Econometrics: Recitation 1

Seung-hun Lee*

September 15th, 2022

1 Logistics

1.1 Recitation

- Location: 602 Northwest Corner Building
- Time: Thursdays 1PM-2PM
  - Roughly 30-40 minutes will be spent on reviewing materials from the professor's lecture for the week. The rest will be spent on Stata demonstrations. On weeks without Stata demonstrations, 60 minutes will be fully devoted to the review.
  - Depending on the room availability, I will stay an extra 10-20 minutes to answer your questions about problem sets, concepts covered, and etc.
- I will focus on reviewing the concepts covered on the two classes before the recitation - so my recitations cover materials from the Tuesday and Thursday regular classes in that week.
- As you will notice in this semester, new methods arise in an attempt to fix drawbacks of the previous methods. I will attempt to establish the relationship among econometric methods by pointing out how one method makes up for flaws in the previous methods and so on.
- I TA-ed the same class multiple times. The materials are expected to be similar. For those who want to take a look, go to here (F2021) and here (F2019).

*Contact me at sl4436@columbia.edu if you spot any errors or have suggestions on improving this note.

1.2 Office Hours

- Location: Zoom (Click here to join) - I will be logging onto Zoom from the Lehman Library room 327, so you have an option to see me in person there as well.
- Time: Fridays 4PM - 6PM (If you can't make it on this time, send me an email)

1.3 What you should expect from me and the recitations

- Post recitation notes by noon on Thursdays for you to download.
- Slides will be posted AFTER the recitation (before midnight on Thursdays).
- While I am the one teaching the course, it will not be complete without you. Do not hesitate to ask questions, make suggestions at any time. I am here to help you all in any way I can.

2 What is econometrics trying to achieve?

Econometrics is a field in economics that tries to answer real life questions, from finding how additional schooling affects our wage in the real market to how certain policies affect certain outcomes (e.g. Did a particular welfare program improve health?). Ultimately, it is about making a **quantitative** statement about two or more random events. We can make simple correlational statements using econometrics (e.g. wages and schooling is positively correlated). However, we want to go one extra mile and identify *causal relationships* - as in how much does our wage rise with one additional year of schooling? In order to achieve this goal, we collect data from a suitably defined population (more on this later) 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?

2.1 Correlational vs causal relations

Suppose you have two random variables - call them X and Y. You want to identify if X *causes* Y, beyond being simply correlated. A **correlation** between X and Y is a statistical mea-

sure that describes how the two variables move together. It captures *any* type of statistical dependence that moves the two variables together. It can actually be causal in some cases but in most cases, it is not. For instance, it may be the case that X and Y are correlated because X actually causes Y (think about more schooling and higher wages). However, Y can cause X, which makes them correlated but not X causing Y . It may be the case that X and Y are jointly moving because there is another factor Z that affects both, making the connection X and Y indirect.

For us to be able to claim **causal** relationship, it must be that cleanly changing variables X leads to changes in Y. More formally, if we are able to observe that exogenous changes in X affects Y, the two variables are in a causal relationship in that X causes Y. This is actually much more difficult to achieve than it seems. Changes in X may be a combination of many things - changes from X alone and changes from other factors that may indirectly affect X. Because real life events can rarely be controlled, we need to be able to tell the two apart. This is where randomized control trials and natural experiments can come in handy. Both methods, if correctly designed and the conditions are met, can help us tell whether changes in X affects Y, and by how much. Methods we use in econometrics are handy in that we can express concisely the relationship between X and Y variables in a single equation.

2.2 Suitably defined population

In econometrics, we need to be clear about the population that we are interested in. When we say we are interested in the relationship between schooling and wages, whose effects are we interested in? Are we interested in this for the entire US population, high school graduates, or college graduates? This matters because it determines what sampling methods we need to use in order to ensure that our observations in the data are comparable (or at least get close to it). In terms of sampling methods, we can use complete randomization, stratified randomization, or cluster randomization depending on our population characteristics.

This also raises another issue: We are almost surely never going to get the data from the entire population. Gathering data from the entire population is logistically (and maybe even 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, and we will do various diagnostic testing to see if they can be reasonably close to the true value.

2.3 What methods to use?

In econometrics, we will use many estimation methods to obtain the sample analogue of the parameter of interest. The most commonly used methods are:

- Ordinary least squares (OLS): Used in cross-sectional analysis where we observe multiple observations in one time period. 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

Also, depending on the type of variables we use in our exercise, we have the following methods

- 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

3 Randomized control trials

In **randomized control trials**, we randomly categorize some individuals under treatment group and controlled group and run various tests as if this is a laboratory experiment. While this is difficult to pull of due to logistical and ethical concerns, many economists have successfully ran field experiments to answer interesting questions in economics (e.g. Does health interventions, such as deworming treatment, increase educational attainment for children in Kenya?[1]).

[1]Edward Miguel, and Michael Kremer (2004) "Worms: Identifying impacts on education and health in the presence of treatment externalities", *Econometrica* 72(1), 159-217

Randomized control trial serves as a benchmark for good program evaluation, where we assess the effectiveness of a policy in achieving its intended goals. As such, the topics we cover in this course will be largely based on randomized control trials and how to analyze/critique them.

3.1 Potential outcomes framework

Let Y_i be the observed outcome for individual $i \in \{1, ..., N\}$. For each individual i, she can either be treated or be in the control group (not both). This will be captured by a variable W_i which equals 1 if i is treated, 0 if otherwise. Since each N individuals can be assigned to the treatment or not, we can define $\mathbf{W}$, an N-tuple vector of treatment assignment for all individuals. While it is realistic to think that individual treatment could be affected by how many others have taken the treatment, we will assume this away for the entire course until otherwise specified. This is known as **stable unit treatment value assumption (SUTVA)**. Mathematically, we can write $Y_i(\mathbf{W}) = Y_i(w)$

A **potential outcome**, written as $Y_i(w)$ where $w = 1$ if individual i is in the treatment group and $w = 0$ if not, is the outcome for the treated ($Y_i(1)$) and the untreated individual ($Y_i(0)$). Since an individual cannot be treated and untreated at the same time, we can only observe at most one of $Y_i(1)$ or $Y_i(0)$ for a given individual. This is known as a **fundamental problem of missing data** and raises problems in estimating the effect of a treatment.

We can write the relation between the observed outcome Y_i and the pair of potential outcomes $Y_i(1)$ and $Y_i(0)$ as follows.

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

In here, we know W_i and Y_i for everyone regardless of treatment assignment. The problem is that we cannot see $Y_i(0)$ for the treated group and $Y_i(1)$ for the untreated group.

3.2 Treatment effects

What we are interested in is whether the treatment W_i affects the outcome of interest Y_i, or gains from the treatment that can be written as

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

However, because of the fundamental problem of missing data, we cannot capture the exact gain at all. Thus, the closest we can get to is the **average treatment effect (ATE)** and the **average treatment effect on the treated (ATT)** which are defined as

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

Furthermore, to characterize these concepts with something we can estimate, we need to set some assumptions. The strongest assumption possible is the **randomized assignment**, which states that $Y_i(1)$ and $Y_i(0)$ are independent of W_i. Or

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

and similarly for $E[Y_i(0)]$. With this assumption and the definition of potential outcomes framework, we can write

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

In this framework, we can see that by letting $W_i = 1$, we get $E[Y_i(1)] = E[Y_i|W_i = 1]$. If $W_i = 0$, then $E[Y_i(0)] = E[Y_i|W_i = 0]$. This implies that under random assignment, we can identify the average treatment effect as

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

or by using the difference in the subsample means. Later, you will learn that this can be obtained by an OLS regression where you have W_i as your single independent variable