Introduction to Econometrics: Recitation 1

Seung-hun Lee

September 27th, 2021

1 Logistical Matters

- **Basics**:
  - Name: Seung-hun Lee
  - Office hours: Monday 10:30am - 12:30am at Online (link to Zoom)
  - Recitation: Monday 9:00am - 10:00am at Hamilton 315
    - If these times do not work for you, reach out to other TAs (they are very talented). If you do want to reach out to me for personal matters, send me an email. Otherwise, use Ed Discussions
  - Contact: sl4436@columbia.edu
- **Goal**: The goals of this recitation are threefold, in the order of importance:
  1. To make sure that you are comfortable with the key concepts in class
  2. To suggest key methods to approach various questions
  3. To introduce you to STATA, an "industry standard" for those studying applied econometrics

  To this end, the TAs for this course will help you along the way. So visit the teaching assistants as often as possible. We will be ready to answer your questions any way we can
- **Recitation**: In terms of teaching, I will prepare the lecture notes for each recitation. They will be uploaded by no later than 8PM on Sundays. I will use slides as well for the recitation, but that will be uploaded after each recitation with the annotations. I will spend time reviewing class materials, solving some unassigned questions in the problem sets, showing STATA demonstrations along the way. If you think there is a better way for you to learn from me, do let me know. Ed Discussions is the preferred medium, but my email is also open.

- **Questions**: You are more than welcome to ask questions. Do make use of recitation and office hours to get answers to your questions. If you want to ask questions online, I highly recommend you to use Ed Discussions. The rationale is to make sure that everyone is on the same page when it comes to key questions that arise during the semester. TA's and instructors will regularly check and answer them.
- **...and lastly, what is econometrics?**: (Very, very personal) Econometrics is ultimately about making a quantitative statement about two or more random events. They can be a statement about correlation or causation, ideally the latter. For instance, you may be interested in whether higher GPAs (X) in college leads to higher wages after graduation(Y). The methods we learn in this course is aimed to help you find a clean, reliable ways to make that numerical statement. You may not understand the items below now, but you can go back to these after the course is over and remind yourself what you have learned this semester.
  - The ideal case: Ordinary least squares
    - $\rightarrow$ *Unfortunately, they may not always be applicable because of the data structure, measurement error in variables, and unobservable variables determining our outcome*
  - If our dependent variable is binary: Nonlinear methods
  - If we have multiple observations across multiple time periods: Panel method
  - If we have variables that we can use as a 'proxy': Instrumental variable method
  - If we have an experimental structure: Difference-in-differences, Regression discontinuity
  - If we observe one entity over multiple periods: Time series method
  - If we need to deal with Big Data methods: LASSO

2 Review of Key Concepts from Statistics

2.1 Probability

Suppose that you are throwing a fair dice twice, and that you keep track of what number you get for each throw. The possible outcome for each throw is

$$\{(1,1),(1,2),...,(1,6),(2,1),...,(2,6),...,(6,6)\}$$

The collection of every possible outcome is defined as a **sample space**, or a **population**. An **event** refers to a subset of the sample space. For instance, the event that you would be interested in might be the appearance of an odd number on the first throw and so on. They are called **mutually exclusive** if occurrence of one event prevents another event from occurring. An example of an mutually exclusive event to the aforementioned example would

be an appearance of an even number on the first throw. When we consider the union of the two events, there is no other possible event, which makes the union an **exhaustive event**.

A **probability** of an event A, denoted as $P(A)$ or $\Pr(A)$, is the proportion of times the event A occurs in repeated trials of an experiment. They satisfy

- $0 \leq \Pr(A) \leq 1$
- If $A_1, .., A_n$ are mutually exclusive, $\Pr(A_1 \cup ... \cup A_n) = \Pr(A_1) + ... + \Pr(A_n)$
- If $A_1, .., A_n$ are exhaustive, then $\Pr(A_1 \cup ... \cup A_n) = 1$

A **random variable** X is a function defined such that the sample space acts as a domain and the set of numbers is the range. The random variable numerically describes the outcome of an experiment. For instance, X could be defined as the sum of the two numbers that appear on each throw. The random variables can be either **discrete** if it takes only takes finite or countably infinite values. They are **continuous** if they can take any value from some interval of numbers.

For the sake of discussion, let X be a continuous random variable. Then $f(x)$ is a **probability density function** (PDF) if

- $f(x) \geq 0$ for all $x \in X$
- $\int_{-\infty}^{\infty} f(x)dx = 1$, $\int_a^b f(x)dx = P(a \leq x \leq b)$

A **cumulative density fuction** (CDF) $F(x)$ is defined as $\Pr(X \leq x)$, or probability of any value less than or equal to x occurring.

There are cases where we are concerned with multiple random variables. Let X, Y be discrete random variables. We call $f(x, y)$ a **discrete joint PDF** (or joint probability mass function) if

$$f(x, y) = \Pr(X = x, Y = y)$$

The **marginal probability density function** of x is defined by

$$f(x) = \sum_{y \in Y} f(x, y) \text{ or if continuous, } \int_{y \in Y} f(x, y)dy$$

As we will see later on, problems like this is usually addressed by drawing a table of joint probability distribution.

It is possible that our interest could be on a behavior of one variable while conditioning that the other variable takes certain value. For this, we use the concept of **conditional PDF**, denoted as $f(x|y)$. It calculates the probability that the random variable X takes the value x

while the sample space is effectively reduced to Y taking the value y. Mathematically, it is defined as

$$\Pr(X = x|Y = y) = f(x|y) = \frac{f(x,y)}{f(y)} = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)}$$

The two random variables are said to be **statistically independent** if the following is satisfied

$$\Pr(X = x|Y = y) = \Pr(X = x) \text{ (or } f(x|y) = f(x))$$

which implies that the joint PDF can be expressed as

$$f(x,y) = f(x)f(y)$$

In many cases, we are interested in some key properties of the distribution. Some of them are (X, Y are discrete random variables)

- **Expected Value**: $E(X) = \sum_{x \in X} xf(x)$
- **Variance**: $var(X) = E[(X - E(X))^2] = E(X^2) - [E(X)]^2$
- **Covariance**: $cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$
- **Correlation Coefficient**: $corr(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y)}}$

Also, it is quite handy to be familiar with the following properties.

Property 2.1. *Note: Small case letters denote a constant, capital case denote a random variable*

- $E(aX + b) = aE(X) + b$
- $E(g(X)) = \sum_{x \in X} g(X)f(x)dx$
- $var(aX + b) = a^2 var(X)$
- $var(aX + bY) = a^2 var(X) + b^2 var(Y) + 2ab \times cov(X,Y)$
- *If X, Y are independent, then $Cov(X,Y) = 0$, implying that $E(XY) = E(X)E(Y)$*
- $cov(a + bX, c + dY) = bd \times cov(X,Y)$

2.2 Some useful distributions

Here, I will go over some distributions that will be repeatedly used throughout the course. You do not have to memorize the PDFs of all these distributions (but it always helps if you do, even if not for the exam). The goal here is to help you be familiar with these distributions whenever the use of these becomes necessary.

- **Normal distribution**: A distribution of random variable X is said to be normal with mean μ and variance σ^2 if we write the PDF as

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2}}$$

  Even better, we can standardize this random variable to have mean 0 and variance 1 by defining a random variable $Z = \frac{X-\mu}{\sigma}$ (You can verify that Z has mean 0 and variance 1 using the properties above). The new PDF for the standard normal distribution can be written as

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

  They are used in cases where we are working with a large number of samples

- **Chi-squared(χ^2) distribution**: If Z_1,...,Z_n are independent standard normal distribution, we can define a new random variable $Z = \sum_{i=1}^{n} Z_i^2$ as a chi-squared distribution with degrees of freedom n.

- t **distribution**: If Z is a standard normal variable and X is a chi-squared distribution with k degrees of freedom, then t distribution with k degrees of freedom is defined by

$$t_k = \frac{Z}{\sqrt{X/k}}$$

  We normally use them in small sample cases. Note that as the number of sample rises, this distribution approximates to normal distribution.

- F **distribution**: Let X_1 and X_2 be chi-squared distribution with degrees of freedom k_1 and k_2 respectively. Then F distribution with (k_1, k_2) degrees of freedom is defined by

$$F_{k_1,k_2} = \frac{X_1/k_1}{X_2/k_2}$$

  They are usually carried out to test multiple hypotheses at the same time or testing the whole model.

2.3 Statistical Inference

Statistical Inference refers to any process of using data analysis to make judgements on some parameters of a population using a randomly sampled observation from the larger population. To conduct a statistical inference, one must first identify a statistically **testable question (hypothesis)**, collect and organize the data, carry out an estimation, test the hypothesis, and come to a conclusion using confidence intervals or other methods.

Once the question is identified and the data is obtained, **estimation** of the statistic of interest should follow. Usually, this means obtaining sample mean and sample variance. The next step is **hypothesis testing**, where the null hypothesis (H_0) is tested against an alternative hypothesis (H_1). Typically, claims related to status quo is put as null hypothesis and the new discovery one is trying to sell is classified as alternative hypothesis. Lastly, one constructs a **confidence interval** either support the null hypothesis or reject it. Typically, researchers use 95% or 99% confidence interval. Confidence interval is used to gauge how accurately your test statistic is calculated. If the confidence interval includes the null hypothesis value (zero, in most cases), then your test statistic is calculated inaccurately and you fail to reject the null hypothesis.

For instance, suppose that the question of interest is the effect of class sizes on test scores. Then, one needs to define what a small classroom is. Let's suppose that any class with fewer than 20 students is a small classroom. Then, calculate the sample mean and the sample variance of test scores of each type of classroom. Let $\bar{Y}_s, S_s^2$ denote sample mean and sample variance of test scores from the small classrooms. ($\bar{Y}_b, S_b^2$ are sample mean and sample variance from large classrooms). Once those are calculated, a test statistic is required to test the following null and alternative hypothesis:

$$H_0 : E(Y_b) - E(Y_s) = 0 \text{ vs. } H_1 : E(Y_b) - E(Y_s) \neq 0$$

$$\text{Test statistic: } \frac{\bar{Y}_b - \bar{Y}_s}{\sqrt{\frac{S_b^2}{n_b} + \frac{S_s^2}{n_s}}}$$

Once test statistic is obtained, it should be compared with a critical value, which differs depending on the significance level and the distribution of the null hypothesis. If the null hypothesis is standard normal and we are using a 5% significance level, we would be using a critical value 1.96. If the test statistics is larger than 1.96, we reject the null. Otherwise, we accept the null.

Another way to conduct this test is to construct a confidence interval of 95% to see if this interval includes the value of $\bar{Y}_b - \bar{Y}_s$ claimed by the null hypothesis, 0. If the confidence interval includes 0, then null hypothesis cannot be rejected. Otherwise, null hypothesis is rejected.

Yet another way to conclude whether null hypothesis should be rejected or not is to see the **p-value**, which is roughly defined as the probability of finding a more extreme result than the observed data, assuming that H_0 is true. The calculation of the p-value depends on how

H_1 is defined. For instance, a p-value of 0.061 would imply that there is a 0.061 probability that the next draw would become more extreme than the current draw. If the significance level is chosen to be 5%, the null hypothesis would not be rejected. If it is selected to be 10%, the null hypothesis would be rejected.

Note that the above is a **two-sided test**. One can always have a **one-sided test** by setting the alternative hypothesis as

$$H_1 : E(Y_b) - E(Y_s) > 0 \text{ or } E(Y_b) - E(Y_s) < 0$$

Before discussing the desirable properties of the estimators, some concepts need to be clarified

- **i.i.d.**: Independent and Identical Distribution. In a random sampling, If $Y_1, .., Y_n$ are from the same population and are selected at random, they are identically distributed. Moreover, if a selection of Y_1 has no impact on the selection of Y_2 and so on, they are independently distributed.
- **Sampling Distribution**: It is a probability distribution of a test statistic derived from the random sample you are working on

To ensure that the estimators obtained from sampling has good qualities, the following standards can be used:

- **Unbiasedness:** $E(\bar{Y}) = \mu_y$, where μ_y is the true parameter value.
- **Efficiency:** $\bar{Y}$ is the efficient estimator if compared against any other estimator $\hat{Y}$, it is the case that $var(\bar{Y}) \leq var(\hat{Y})$
- **Consistency:** $\bar{Y}$ is consistent if $\bar{Y}$ converges to μ_y in probability.
- **Asymptotic Normality:** The estimator is asymptotically normal if it becomes normally distributed as n increases (central limit theorem)

3 Least Squares Estimation

3.1 Derivation of Least Squares Estimator

Suppose that the **population linear regression model** (also known as data generating process in some books) is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

However, we do not know what the true values of the population parameters - β_0 and β_1 in this case- are, an alternative way to approach the problem is to use the **sample linear regression model** (or just model)

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + u_i$$

The ideal estimator minimizes the squared sum of residuals. Mathematically, this can be obtained by solving the following minimization problem and the first order conditions

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$
$$[\hat{\beta}_0] : -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$
$$[\hat{\beta}_1] : -2 \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

The resulting **least squares estimators** are

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \;\; \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Proof for deriving $\hat{\beta}_0, \hat{\beta}_1$

From $[\hat{\beta}_0] : -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$ Note that

$$-2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \implies \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$
$$\implies \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0 \implies \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i$$
$$\implies \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i$$

Using the fact that $\sum_{i=1}^{n} Y_i = n\bar{Y} \iff \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}$, I get

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

As for $[\hat{\beta}_1]: -2\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$ divide both sides by -2 and rearrange

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \implies \sum_{i=1}^{n} X_i Y_i = (\bar{Y} - \hat{\beta}_1 \bar{X}) \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2$$

$$\implies \hat{\beta}_1 \left(\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i \implies \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

Note that

$$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}_i) = \sum_{i=1}^{n} X_i Y_i - \bar{X} \sum_{i=1}^{n} Y_i - \bar{Y} \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \bar{X}\bar{Y}$$

$$= \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} - \bar{Y} \sum_{i=1}^{n} X_i + n\bar{X}\bar{Y} = \sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i$$

and similarly, $\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$. So $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$