Recitation 1

Seung-hun Lee

Columbia University

September 27th, 2021

Some logistics

- Name: Seung-hun Lee
- Office hours: Monday 10:30am - 12:30pm at Online (Link to Zoom)
- Recitation: Monday 9-10am at 315 Hamilton
  - 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
  - Recitation notes will be posted before class (8PM on Sundays)
  - Recitation slides will be posted after class with annotations
- Contact: sl4436@columbia.edu

Some logistics

- **Goal**: The goals of this recitation are threefold (in 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
     $\rightarrow$ so please visit the TAs often for help
- **Recitation**: I will spend time reviewing class materials, solving some unassigned questions in the problem sets, showing STATA demo along the way. (If you think there is a better way, do let me know.)
- **Questions**: You are more than welcome to ask questions. Do make use of recitation, office hours and Ed Discussions

Preview: What is econometrics?

- Econometrics is ultimately about making a **quantitative** statement about two or more random events - either correlational or causational (ideally the latter).
- The methods we learn in this course is aimed to help you find a **clean, reliable** ways to make that numerical statement.
  - 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*
  - Dependent variable is binary: Nonlinear methods
  - Multiple observations across multiple time periods: Panel method
  - Have variables that we can use as a ‘proxy’: Instrumental variable method
  - Experimental context: Difference-in-differences, Regression discontinuity
  - Observe one entity over multiple periods: Time series method
  - If we need to deal with Big Data methods: LASSO

Review: Probability

- Suppose that you are throwing a fair dice twice, where all the possible outcome for each throw is

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

- Defining key terms
  - 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.
  - They are called **mutually exclusive** if occurence of one event prevents another event from occuring.
  - If there are no other possible event, such an event is considered an **exhaustive event**

Review: Probability

- 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.

Properties of probability

- $0 \leq \Pr(A) \leq 1$
- 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 (r.v.)** X : A function where the sample space acts as a domain and the set of numbers is the range.
  - It numerically describes the outcome of an experiment.
  - 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.

Review: Probability Density Functions

- Let X be a r.v. $f(x)$ 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 a probability of any value less than or equal to x occurring. (Denoted as $\Pr(X \leq x)$)
- We can study the probability of a multiple r.v. X and Y
  - $f(x, y) = \Pr(X = x, Y = y)$ is a **probability mass function**
  - The **marginal probability density function** of x is defined by

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

(we fix $X = x$ and sum over all possible values of Y)

Review: Conditional PDF

- In econometrics, we are frequently interested in the behavior of one variable while conditioning that the other variable takes certain value

- A **conditional PDF**, denoted as $f(x|y)$, 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 **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

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

Review: How to characterize the distribution?

- 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**: $var(X) = E[(X - E(X))^2]$
  - **Covariance**: $cov(X, Y) = E[(X - E(X))(Y - E(Y))]$
  - **Correlation Coefficient**: $corr(X, Y) = \frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y)}}$
- There are nice properties involving variances and expected values that makes calculation simpler. (Refer to the lecture notes)

Review: Useful distributions

- **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

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

a standard normal distribution has mean 0 and variance. The PDF for the standard normal distribution can be written as

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

Review: Useful distributions

- **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

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

- 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

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

Review: Statistical Inference

- **Statistical Inference** refers to any process of using data analysis to make guesses 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.
  - **Estimation**: process of guessing the statistic of interest (sample mean and sample variance).
  - **Hypothesis testing**: You test the null hypothesis (H_0) is tested against an alternative hypothesis (H_1).
  - **Confidence interval**: How accurately your statistic of interest is calculated. Typically, researchers use 95% or 99% confidence interval.

Review: Statistical Inference (example)

- Suppose you are interested in the effect of class sizes on test scores.
- Find a small classroom is (say, any class with fewer than 20 students)
- Then, calculate the sample mean and the sample variance of test scores of each type of classroom.
- You now calculate the test statistic:

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

- Your hypothesis would be ”the mean test score of small classroom is different from others”. We can write

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

Review: Statistical Inference (example)

- To carry out the test, we can do as follows
  - Calculate the test statistics and directly compare with the **critical value** derived from assuming that the null hypothesis distribution is correct
    - If we have a standard normal and use 5% significance level, we compare the test statistic against the critical value of 1.96
  - You get the **confidence interval** (usually 95%) to see if this interval includes 0, the value claimed by the null hypothesis.
    - If the confidence interval includes 0, then null hypothesis cannot be rejected. Otherwise, null hypothesis is rejected.
  - Other way is to see the **p-value**, which is roughly defined as the probability of finding a more extreme result than the observed data.
    - Typically, we want to see if the p-value is less than 0.05
    - Even better if less than 0.01

Review: Desirable properties in statistical Inference

- **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 the number of observation increases (central limit theorem)

Ordinary Least Squares: Population vs sample linear regression models

- 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 the true values of the population parameters - β_0 and β_1
- An alternative way to approach the problem is to use the **sample linear regression model** (or just model)
$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + u_i$$
where $\hat{\beta}_0, \hat{\beta}_1$ are estimates of β_0, β_1

Ordinary Least Squares: Definition

- 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

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