Recitation One
Econometrics - Fall 2018

Nathaniel Mark

October 16, 2018

1 Introduction

Hello Everyone! My name is Nate – I am one of those odd TAs that enjoys teaching, so please let me know if you have any questions. Additionally, this is my third time teaching econometrics – so I have plenty of experience with the material. Lastly, this recitation is **for YOU**, so let me know if there is anything you would like me to cover in recitations. You can do so by emailing me at ndm2125@columbia.edu or you can leave anonymous feedback by filling in the survey at the top of my teaching website at www.nathanieldmark.weebly.com.

Additionally, *all* of my recitation notes can be found on that website.

2 Topic One: Random Variables

There is a lot of variation in the real world. Take height. Some people are short, some are tall. How do we describe this variation? One way is to think of each person's height coming from a random variable. So, when you are born, we take a random draw from this random variable and that will be your height. This is often how we think about data in econometrics. In statistics, we called this the "population distribution".

What is a random variable? Prof. Bai said it is a "mapping from events to the real line". What does he mean by that? He means that a random variable is a collection of possible numbers ("events") that can occur and the "probability" that each of those numbers occur.

A *parameter* is a value that describes the random variable.

A random variable is *discrete* if there are a finite number of possible events that the random variable describes. Then the random variable is a collection of all of those finite events and the probabilities that they occur.

Example: For discrete random variables, these definitions is pretty clear. For example: A *Bernoulli* random variable with *parameter* p represents:
2 possible events: X=0 or X=1.
Probabilities associated with those events: P(X=0)=p and P(X=1)=1-p.

Example: For discrete random variables, these definitions is pretty clear. For example: A *Bernoulli* random variable with *parameter* p represents the following "mapping":
2 possible events: X=0 or X=1.
probabilities associated with those events: P(X=0)=p and P(X=1)=1-p.

A random variable is continuous if all possible numbers in some range are possible events.

With continuous random variables, the interpretation of random variables are a little more confusing. With continuous random variables, the probability of any given number occurring is zero (think: what is the probability that a given human is 5.4235239482930583... feet tall?). So, instead of probabilities, continuous random variables are defined by a "probability density function" (PDF). The PDF again assigns a number to all possible events. This number is, in a sense, a measure of the *relative probability* or *relative likelihood* of different events with respect to each other. If the probability density function for height is =.05 for 5 feet and .01 for 4 feet, we can say that it is *5 times more likely for 5 feet to occur than for 4 feet to occur.*

3 Topic Two: What is important from probability overview?

- Understand what Bernoulli, Binomial, Normal, and Exponential random variables are.

  - What does their probability density function look like? What are their expectations? Variance?
  - What kind of data do we use them to model?

- Definition of expectation and rules of expectation
- Definition of variance and rules of variance
- Law of total expectation
- Joint v. Marginal distributions
- Conditional Expectation (*important*)
- Definition of IID

4 Topic Three: Estimators

As I discussed in the beginning, one way to describe outcomes in the real world is to think that there is some underlying random variable that describes those outcomes. For example, we can think of some random variable that a person's income is drawn from. The point of econometrics is to estimate what this underlying random variable looks like. To do so, we estimate its *parameters*.

For example, we might want to estimate the *expected income or 22 year olds* or the *variance of mortage interest rates*.

What is an estimator?

An estimator is a function with :

- inputs: Data
- output: An estimate

Example 1: We estimate expected income by the average of incomes.
input: $Y_1, Y_2, Y_3, ..., Y_N$ are incomes of people 1,2,3...
The estimator is the function $\frac{\sum Y_i}{N}$.

Example 2: We estimate the change in expected value of income, given a one unit change in age.
input: $Y_1, Y_2, Y_3, ..., Y_N$ are incomes of people 1,2,3..N; $X_1, X_2, X_3, ..., X_N$ are ages of people 1,2,3..N
The estimator is the function $\frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$.

Properties of estimators:
Before we have seen our data, we can think of $Y_1, ..., Y_N$ as random variables. Then, the estimator $\frac{\sum Y_i}{N}$ is also a random variable.

Once we observe our specific data values, then we can plug in observe y_1, y_2, y_3 to our estimator to get the *estimate* $\frac{\sum y_i}{N}$.

There are three properties of estimators that we talk about a lot in this class (especially the first two):

Say θ is the **true value** of the parameter we are trying to estimate. θ is simply a number. An estimator, denoted $\hat{\theta} = g(X_1, ...X_N)$ is simply a function of the data.

1. Consistency
   Definition: An estimator $\hat{\theta}$ is *consistent* if $\hat{\theta} \to^p \theta$. $\to^p$ means "converges in probability"
   More detailed definition: An estimator $\hat{\theta}$ is *consistent*
   if for all $\epsilon, lim_{N\to\infty} P(|\hat{\theta} - \theta| < \epsilon) = 0$
   Intuitive definition: An estimator $\hat{\theta}$ is *consistent* if $\hat{\theta}$ gets arbitrarily close to θ as N gets larger and larger.

2. Unbiasedness
   Definition: An estimator $\hat{\theta}$ is *unbiased* if $E[\hat{\theta}] = \theta$.

3. Efficiency
   Definition: An estimator $\hat{\theta}$ is more efficient than another estimator $\tilde{\theta}$if $Var(\hat{\theta}) < Var(\tilde{\theta})$. (assuming both are unbiased)

EXAMPLES:
Unbiased but not consistent estimator for $E[Y_i]$: Y_1
Consistent but biased estimator for $E[Y_i]$: $\frac{1+\sum Y_i}{N}$

5 Topic Four: Ordinary Least Squares / Regression

In economics, a question we often want to answer is *what is the relationship between variable A and variable B?* For example, *what is the effect of getting*

older on your expected income? One way to answer this is to assume that B is a linear function of A. Using our example, we assume that the relationship takes the form:

$$Income_i = \beta_0 + \beta_1 Age_i + \epsilon_i$$

β_1 is the increase in expected income by increasing age by one year. We are assuming that there is a linear line that describes the relationship between income and age.

What estimator do we use to estimate β_0 and β_1? In other words, how do we find the line that best fits the data? *Ordinary least squares* (aka regression) says the line that best fits the data is the line that minimizes the sum of squared errors. That is, it is the line that minimizes the sum of squared distances between $Income_i$ and $\beta_0 + \beta_1 Age_i$. Therefore, our estimator of β_0 and β_1 are defined as the possible values that minimize that sum of squared distances:

$$min_{\beta_0,\beta_1} \sum (Income_i - (\beta_0 + \beta_1 Age_i))^2$$

After some algebra, we find that the solution to this minimization process is:

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

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