# Econometrics

## Lecture 2

Nathaniel Higgins

JHU

14 September 2015

# Plan for the lecture

- Administrative details
- Probability basics (review)
- Linear regression modeling with a single variable (the "univariate" model)
- Linear regression modeling with two variables (the "bivariate" model
- Selected homework solutions (if time)

# Administrative stuff

- Everybody have access to the student discounted version of Leada? You should not be paying full price.

# Random variables

- What is a random variable (and why should you care?)
- A random variable is a variable that takes on different values, depending on a random process
- Another way of saying the same thing: The values are *a result* of a consistent random process
- The value that a random variable takes on cannot be known before the “draw” (or “experiment”)
- *But*, we usually assume that we know *something* about random variables: we know how the values are distributed. What does this mean?

# Random variables

- The *why you care* part is pretty easy in one respect, and hard in another
- Short term: you care because econometrics uses random variables all the time, so you need to familiarize yourself with the concept
- Longer term: the reason why it makes sense to think about economic/social variables are *random* — in the technical sense — takes some experience to appreciate

# Random variables

- The *why you care* part is pretty easy in one respect, and hard in another
- Short term: you care because econometrics uses random variables all the time, so you need to familiarize yourself with the concept
- Longer term: the reason why it makes sense to think about economic/social variables are *random* — in the technical sense — takes some experience to appreciate
- For now, just realize that:
  - part of what we're going to do as econometricians is "model" things — come up with some sort of an equation that describes them
  - we always have the option to model things as random or deterministic
  - most interesting variables will be model as (at least partially) random

# Random variables

- . . . back to the question of what it means to know how a variable is distributed . . .
- Knowing how a variable is distributed is just knowing how often some values come up relative to others
- This makes random variables different from other (non-random) variables

# Relationships between random variables

- Random variables are part of every econometrics equation (model)
- Think of an equation like

$$mass = density \times volume$$

- anything random about that?

# Relationships between random variables

- Random variables are part of every econometrics equation (model)
- Think of an equation like

$$mass = density \times volume$$

- anything random about that?
- Nope. It's a fixed, physical relationship

# Relationships between random variables

- Random variables are part of every econometrics equation (model)
- Think of an equation like

$$mass = density \times volume$$

- anything random about that?
- Nope. It's a fixed, physical relationship
- But most data we care about doesn't work like this
- If we have fixed relationships (mass = density $\times$ volume) then we don't need econometrics
- econometrics is about relationships that are
  1. less than perfectly predictable (but somewhat predictable)
  2. based on relationships between multiple random variables

# Econometric relationships

- A more common representation of an econometric relationship would be

$$y = 2x + u$$

- where $u$ is a random variable, $y$ is earnings, and $x$ is years of education (and "2" is a number describing the relationship between $x$ and $y$)

# Econometric relationships

- A more common representation of an econometric relationship would be

$$y = 2x + u$$

- where $u$ is a random variable, $y$ is earnings, and $x$ is years of education (and "2" is a number describing the relationship between $x$ and $y$)
- What does $u$ represent?

# Econometric relationships

- A more common representation of an econometric relationship would be

$$y = 2x + u$$

- where $u$ is a random variable, $y$ is earnings, and $x$ is years of education (and "2" is a number describing the relationship between $x$ and $y$)
- What does $u$ represent?
- $u$ represents everything about the determination of $y$ that is not captured by $y = 2x$
- Every time $y$ is not equal to $2x$, it is equal to $2x + \textit{something}$
- In econometrics, we view that *something* as a random variable

# Econometrics: bridge from statistics

- Viewing that *something* as a random variable is what enables us to use all our statistics to estimate the relationship between $y$ and other stuff
- it's what enables us to test hypotheses
- Without $u$, there is no econometrics as we know it

# Econometric models

- What else is random?

# Econometric models

- What else is random?
- $y$ is random

# Econometric models

- What else is random?
- $y$ is random
- Just know that for now. If you don't quite understand why on an intuitive level, that's OK
- On a less-than-intuitive level, we know $y$ is a random variable because it is a linear combination of random variable(s) (i.e. because $y$ is *built* using random variables, it must be a random variable)

# Econometric models

- What else is random?
- $y$ is random
- Just know that for now. If you don't quite understand why on an intuitive level, that's OK
- On a less-than-intuitive level, we know $y$ is a random variable because it is a linear combination of random variable(s) (i.e. because $y$ is *built* using random variables, it must be a random variable)
- Suppose $y$ is annual earnings of individuals in the U.S.
- Let's build a model of earnings. But for now, let's build an uber-simple model of earnings
- If you had to predict what earnings would be, and you didn't know anything about econometrics yet, but you did know a thing or two about random variables and statistics, how would you do it?

- Hold that thought

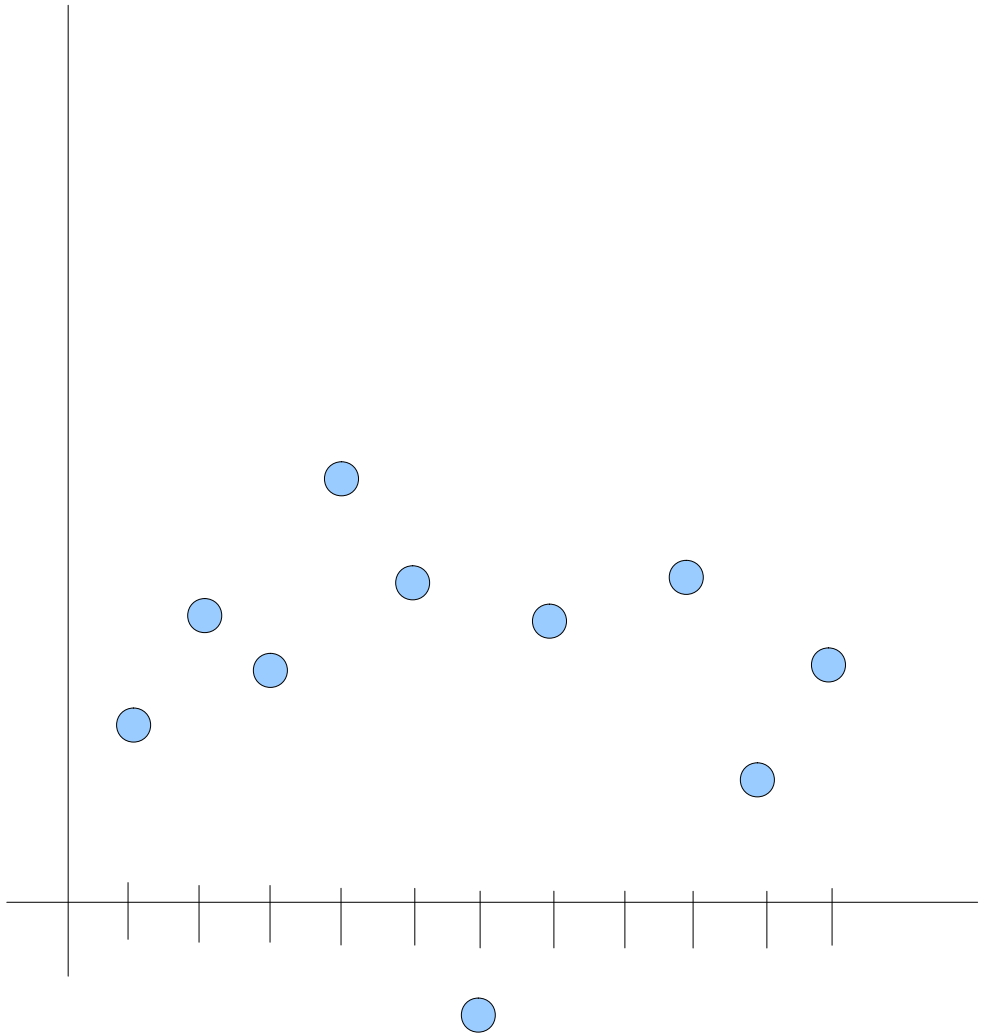
- Hold that thought
- Time to start thinking about models

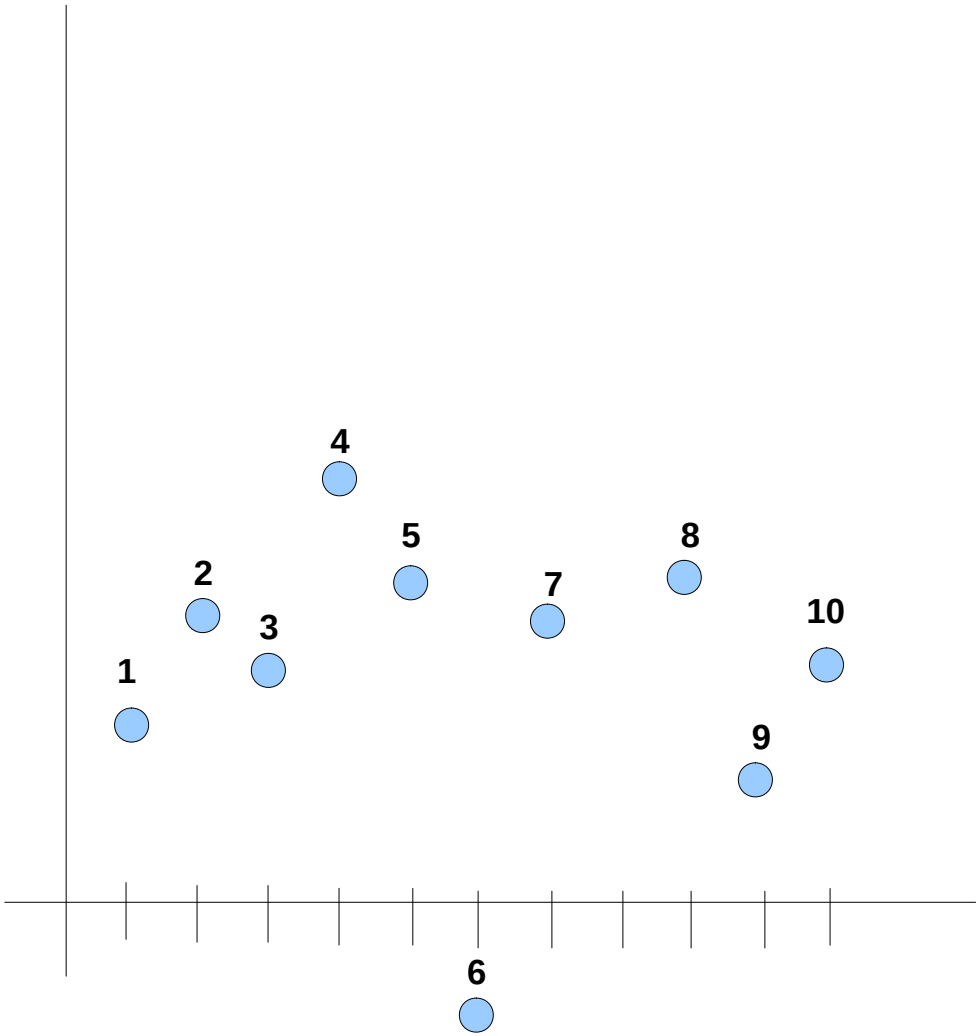
- Hold that thought
- Time to start thinking about models
- Suppose you have a variable $y$ (you can continue to think of it as earnings, if you want). You want to model $y$. What does that *mean*?

- Hold that thought
- Time to start thinking about models
- Suppose you have a variable $y$ (you can continue to think of it as earnings, if you want). You want to model $y$. What does that *mean*?
- Think of *modeling* $y$ as: predicting the value of $y$

- Hold that thought
- Time to start thinking about models
- Suppose you have a variable $y$ (you can continue to think of it as earnings, if you want). You want to model $y$. What does that *mean*?
- Think of *modeling* $y$ as: predicting the value of $y$
- So you have some data, say 100 observations of $y$, and your goal is to predict the value of the next observation (the $101^{st}$). What do you do?

- Think graphically




1
2
3
4
5
6
7
8
9
10

- What is the next point going to be? (don't say it out loud yet)
- What is your best guess of the value of point 11?

- What is the next point going to be? (don't say it out loud yet)
- What is your best guess of the value of point 11?
- If we put some numbers to it . . .

| obs | value |
|---|---|
| 1 | 2.5 |
| 2 | 3.5 |
| 3 | 3 |
| 4 | 5 |
| 5 | 4 |
| 6 | -1 |
| 7 | 3.75 |
| 8 | 4 |
| 9 | 2 |
| 10 | 3 |
| 11 | ? |

- Now you can say it out loud
- What is your best guess of the value of point 11?

- Now you can say it out loud
- What is your best guess of the value of point 11?
- You could guess whatever you want
- or you could come up with a systematic response — a system that you could use to answer this question in many other circumstances

- Now you can say it out loud
- What is your best guess of the value of point 11?
- You could guess whatever you want
- or you could come up with a systematic response — a system that you could use to answer this question in many other circumstances
- My systematic prediction (and most people's) is to choose the mean
- I'm going to introduce some standard notation here, so we'll go slow

- Prediction of $y_{11}$:

$$\widehat{y_{11}}$$

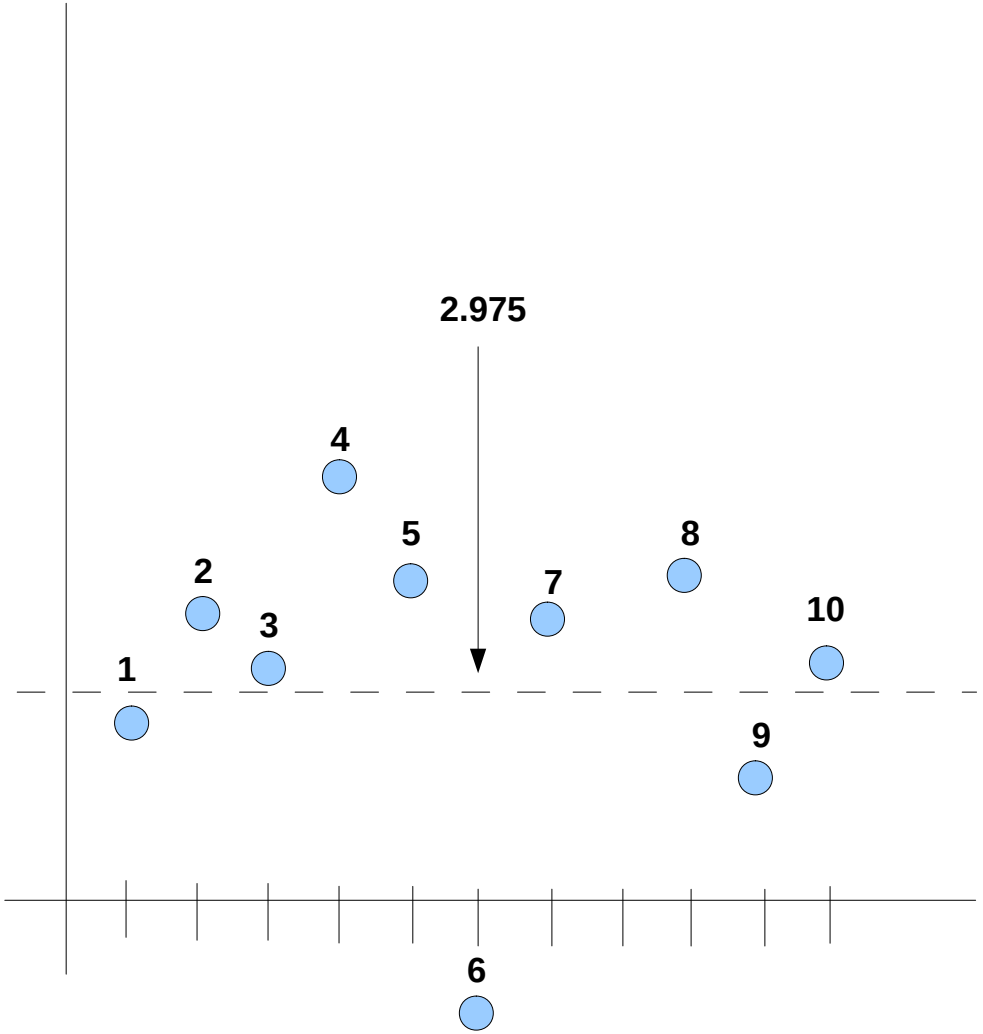
- Prediction of $y_{11}$:

$$\widehat{y_{11}} = \bar{y}$$

- Prediction of $y_{11}$:

$$\begin{aligned}\widehat{y_{11}} &= \bar{y} \\ &= \frac{1}{10}\sum_{i=1}^{10} y_i\end{aligned}$$

- What will this look like graphically?


2.975
4
5
8
2
7
10
3
1
9
6

- A straight line

- A straight line
- A *constant* value

- A straight line
- A *constant* value
- Our best prediction of a value of $y$ **if we don't know anything else**, is (usually) the mean of $y$

- A straight line
- A *constant* value
- Our best prediction of a value of $y$ **if we don't know anything else**, is (usually) the mean of $y$
- This isn't too surprising to anybody
- What's unique is to think in terms of a *model*

- So how to think about this in terms of a model?

- So how to think about this in terms of a model?
- We want to write a model where $y$ is equal to an unknown constant

$$y = \beta_0$$

- (it is standard in econometrics to use $\beta$ to represent unknown parameters)
- (it is standard to use subscripts to enumerate multiple unknown parameters, starting with zero: $\beta_0, \beta_1, \ldots, \beta_k$)

- If this were our model . . .

$$y = \beta_0$$

- . . . then this is our idea of *truth*
- This is the nature of a *model*
  - When we model something seriously, we have to take this notion of truth seriously, too
  - "Garbage in, garbage out"

- If this were our model . . .

$$y = \beta_0$$

- . . . then this is our idea of *truth*
- This is the nature of a *model*
  - When we model something seriously, we have to take this notion of truth seriously, too
  - "Garbage in, garbage out"
- So what's wrong with this model?

- The model $y = \beta_0$ implies that $y$ is always equal to $\beta_0$
- But we know this isn't true. We know that $y$ is sometimes different from $\beta_0$
- How to say this in a model?

- The model $y = \beta_0$ implies that $y$ is always equal to $\beta_0$
- But we know this isn't true. We know that $y$ is sometimes different from $\beta_0$
- How to say this in a model?

$$y = \beta_0 + u$$

- where $u$ is all the **unobservable** stuff that makes each observation of $y$ different from $\beta_0$
  - all the "unknown heterogeneity"

- The model $y = \beta_0$ implies that $y$ is always equal to $\beta_0$
- But we know this isn't true. We know that $y$ is sometimes different from $\beta_0$
- How to say this in a model?

$$y = \beta_0 + u$$

- where $u$ is all the **unobservable** stuff that makes each observation of $y$ different from $\beta_0$
  - all the "unknown heterogeneity"
- Thinking like an econometrician means that we view $\beta_0$ as unknown — our goal is to estimate it

- The standard way to denote an estimated parameter is to put a "hat" over the parameter
- So when we write $\beta_0$ we mean the real, true, fixed, and unknown $\beta_0$
- And when we write $\widehat{\beta_0}$ we mean our best estimate of what the real, true, fixed, and unknown $\beta_0$ is

# Linear regression

- When we write the model $y = \beta_0 + u$, we are saying that we think there is some *fixed* $\beta_0$ out there, but that we don't know what it is
- This is our *econometric model*

# Linear regression

- When we write the model $y = \beta_0 + u$, we are saying that we think there is some *fixed* $\beta_0$ out there, but that we don't know what it is
- This is our *econometric model*
- Next question: how do we pick our best estimate?
- We need a method — we need to create an *estimator*

# Linear regression

- When we write the model $y = \beta_0 + u$, we are saying that we think there is some *fixed* $\beta_0$ out there, but that we don't know what it is
- This is our *econometric model*
- Next question: how do we pick our best estimate?
- We need a method — we need to create an *estimator*
- An estimator is a procedure — a series of rules/steps/logic — that we apply to our model and the data that yield an *estimate*

# Linear regression

- When we write the model $y = \beta_0 + u$, we are saying that we think there is some *fixed* $\beta_0$ out there, but that we don't know what it is
- This is our *econometric model*
- Next question: how do we pick our best estimate?
- We need a method — we need to create an *estimator*
- An estimator is a procedure — a series of rules/steps/logic — that we apply to our model and the data that yield an *estimate*
- We will discuss three in this class (these are the three that dominate econometrics): (1) method of least squares; (2) method of moments; (3) method of maximum likelihood

# Linear regression

- Wooldridge does a nice job laying out the method of moments
  - Read Wooldridge's explanation of MM and we'll discuss it briefly
- I will complement what you can find in Wooldridge by walking you through the method of least squares, or *ordinary least squares* (OLS)

# Linear regression

- Wooldridge does a nice job laying out the method of moments
  - Read Wooldridge's explanation of MM and we'll discuss it briefly
- I will complement what you can find in Wooldridge by walking you through the method of least squares, or *ordinary least squares* (OLS)
- When using the method of least squares, we minimize the sum of squared errors
- In the context of our model, we ask ourselves, "what is the number $\widehat{\beta_0}$ that minimizes the sum of squared errors?" (given our model)
- The number that does this is our estimate, which we will denote $\widehat{\beta_0}^{OLS}$ (To the blackboard!)

# Timeout

- At this point in the class, we derived the OLS estimator in the univariate case ($y = \beta_0 + u$) and in the bivariate case ($y = \beta_0 + \beta_1 x + u$) using two methods:
  1. The method of moments
  2. The method of least squares
- Consult the document "Deriving-OLS.pdf"