Econometrics

Lecture 3

Nathaniel Higgins

JHU

Plan for the lecture

- Finish up derivations from last time
- Introduce the idea of maximum likelihood
- Quick review; some terminology
- How to think about the regression equation
- Quick intro to multiple regression using the Ballantine
- Replication project — interim assignment #1

Stuff on the chalkboard

- We started with work on the whiteboard
- Worked through the least-squares derivation of the univariate and bivariate models
- For a full review of the method of moments (MM) and least squares (OLS), see the document “Deriving-OLS.pdf” (posted to website)

Stuff on the chalkboard

- Summing-up where we've been ...
- We introduced the idea of *modeling*, i.e. writing down an equation that describes a variable y
- When we write down a model, our goal is to estimate the parameters of the model
- We estimate the parameters of the model using three ingredients: (1) the model; (2) data; (3) a procedure (or "recipe")
- So far, we have reviewed two procedures: MM and OLS
- Now let's introduce maximum likelihood estimation (MLE) ...

Maximum likelihood

- *Maximum likelihood* is a process — and estimator — that we use to derive estimates of β_0, β_1, etc.
- In short, we select the β parameters that are *most likely*, given the data we observe

Maximum likelihood

- *Maximum likelihood* is a process — and estimator — that we use to derive estimates of β_0, β_1, etc.
- In short, we select the β parameters that are *most likely*, given the data we observe
- So, in terms of our model, our estimates of the parameters are the estimates that maximize something called the *likelihood function*
- Before we apply the ML estimator to our simple univariate or bivariate models, let's work through it on a conceptual basis. This will make the future math much more interpretable.

Maximum likelihood

- Four balls in a hat
- Pull balls out of the hat N times (generate N pieces of data)
- We get data: $(y_1, y_2, \ldots, y_N)$
- Suppose (just to have a concrete example) that the data looked like this:

$$(red, red, red, black, red, black, black, red, red, red)$$

- $N = 10$ (10 draws)
- Number of red balls drawn $= 7$
- Number of black balls drawn $= 3$
- We want to know how many red balls are in the hat (the number of red balls is equal to β, the parameter to be estimated)

Maximum likelihood

- If there had been one red ball and three black balls in the hat:
- The probability of observing the data we observed would be:

$$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{4}, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}) \tag{1}$$

- Or, more compactly:

$$\left(\frac{1}{4}\right)^7 \times \left(\frac{3}{4}\right)^3$$

- This is the probability of the first draw being a red ball, the second draw being a red ball, etc., so that we get *exactly* the draws in the exact order that we got them in (1) above
- (Note that this is not quite the same thing as the probability of getting 7 red balls and 3 black balls)

Maximum likelihood

- If, instead, if there had been two red balls and two black balls in the hat:
- The probability of observing

$$(red, red, red, black, red, black, black, red, red, red)$$

would be:

$$(\frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4})$$

or

$$\left(\frac{2}{4}\right)^7 \times \left(\frac{2}{4}\right)^3$$

Maximum likelihood

- If there were three red balls and one black ball in the hat then the probability of getting the data we observed would be:

$$(\frac{3}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{4}, \frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{3}{4})$$

or

$$\left(\frac{3}{4}\right)^7 \times \left(\frac{1}{4}\right)^3$$

Maximum likelihood

- Compare the probability of observing

$$(red, red, red, black, red, black, black, red, red, red)$$

- If there had been one red ball in the hat:

$$\left(\frac{1}{4}\right)^7 \times \left(\frac{3}{4}\right)^3 = 0.0000257492$$

- If there had been two red balls in the hat:

$$\left(\frac{2}{4}\right)^7 \times \left(\frac{2}{4}\right)^3 = 0.0009765625$$

- Or if there had been three red balls in the hat:

$$\left(\frac{3}{4}\right)^7 \times \left(\frac{1}{4}\right)^3 = 0.002085686$$

Maximum likelihood

- If there had only been $\beta = 1$ red ball in the hat, the chance of observing the data that we did is about 0.003%
- If there had been $\beta = 2$ red balls in the hat, the chance of observing the data that we did is approximately 0.1%, i.e. a little better
- Finally, if there had been $\beta = 3$ red balls in the hat, the chance of observing the data that we did is about 0.2%, twice as likely as if there had been two red balls!

Maximum likelihood

- What is our *maximum likelihood* estimate of β?
- Of the choices 1, 2, or 3 red balls, which one is most likely to have produced the data we observed?
- We pick the choice that results in the highest likelihood

$$\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{4}, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) = 0.0000257492$$

$$\left(\frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}\right) = 0.0009765625$$

$$\left(\frac{3}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{4}, \frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{3}{4}\right) = 0.002085686$$

- We choose $\widehat{\beta}^{MLE} = 3$

- OK. Leave the balls-in-a-hat example behind.
- Speaking generally:
- The maximum likelihood concept is very "portable" — the basic concept works the same in all cases
- The question we seek to answer is always the same: What βs make the data that we observed the most likely?
- We select the parameters β that are most likely to have generated the data we observe $(y_1, y_2, \ldots, y_N)$ and call those βs our estimates $\widehat{\beta}^{MLE}$
- To do this we only need to be able to form the likelihood function

- The likelihood function always takes the same basic form
- To form the likelihood we need to be able to express the probability of each "event," or piece of data y_i occurring
- We need to be able to write down the probability that each observation turns out as it does, that: $Y_i = y_i$

- We write this probability in terms of our data and call it $p(y_i|x_i, \beta)$
  - This is what we did in the balls-in-a-hat example when we wrote down the probability that a given draw $y_i = 1$ (red ball) $= \frac{\beta}{4}$ and the probability that a given draw $y_i = 0$ (black ball) $= 1 - \frac{\beta}{4}$
  - Those two things were expressions of $p(y_i = 1|x_i, \beta)$ and $p(y_i = 0|x_i, \beta)$, respectively

- The likelihood function always takes the same basic form
- To form the likelihood we need to be able to express the probability of each "event," or piece of data y_i occurring
- We need to be able to write down the probability that each observation turns out as it does, that: $Y_i = y_i$
- We write this probability in terms of our data and call it $p(y_i|x_i, \beta)$
  - This is what we did in the balls-in-a-hat example when we wrote down the probability that a given draw $y_i = 1$ (red ball) $= \frac{\beta}{4}$ and the probability that a given draw $y_i = 0$ (black ball) $= 1 - \frac{\beta}{4}$
  - Those two things were expressions of $p(y_i = 1|x_i, \beta)$ and $p(y_i = 0|x_i, \beta)$, respectively

- We write down each and every likelihood function in exactly the same way:

$$L(y_1, \ldots, y_N) = \prod_{i=1}^{N} p(y_i | x_i, \beta)$$

- How do we write down the probability of each occurrence in a simple model?

$$L(y_1, \ldots, y_N) = \prod_{i=1}^{N} p(y_i | x_i, \beta)$$

- Start with $p(y_i | x_i, \beta)$ — what are the probabilities that we'd like to express?

- Back to the balls-in-a-hat example
  - 4 balls
  - Some red, some black
  - True parameter: β = number of black balls
  - We want to estimate β based on data (draws from the hat)
  - There are 3 candidates for $\widehat{\beta}^{ML}$: $\widehat{\beta}^{ML} = 1, \widehat{\beta}^{ML} = 2, \widehat{\beta}^{ML} = 3$
- Our job is to choose one
- There are lots of ways we could do this (exercise: think of how you would do this using an ordinary least squares regression)
- Using the logic of maximum likelihood, we choose our estimate of β to be that which makes observing the data that we did the most likely outcome

- In our case, we evaluated our data (always the *same* data) using the three candidate values of $\widehat{\beta}^{ML}$
- It turned out that the value of β that made observing our data most likely was 3
- So we select $\widehat{\beta}^{ML} = 3$

- In our case, we evaluated our data (always the *same* data) using the three candidate values of $\widehat{\beta}^{ML}$
- It turned out that the value of β that made observing our data most likely was 3
- So we select $\widehat{\beta}^{ML} = 3$
- Q: How did we figure out which value made our data most likely?

- In our case, we evaluated our data (always the *same* data) using the three candidate values of $\widehat{\beta}^{ML}$
- It turned out that the value of β that made observing our data most likely was 3
- So we select $\widehat{\beta}^{ML} = 3$
- Q: How did we figure out which value made our data most likely?
- A: We evaluated the *likelihood function*

- In our case, we evaluated our data (always the *same* data) using the three candidate values of $\widehat{\beta}^{ML}$
- It turned out that the value of β that made observing our data most likely was 3
- So we select $\widehat{\beta}^{ML} = 3$
- Q: How did we figure out which value made our data most likely?
- A: We evaluated the *likelihood function*
- Each data point contributes to the likelihood function. We build the likelihood function by multiplying together the likelihood of observing each piece of data

- In our case, we evaluated our data (always the *same* data) using the three candidate values of $\widehat{\beta}^{ML}$
- It turned out that the value of β that made observing our data most likely was 3
- So we select $\widehat{\beta}^{ML} = 3$
- Q: How did we figure out which value made our data most likely?
- A: We evaluated the *likelihood function*
- Each data point contributes to the likelihood function. We build the likelihood function by multiplying together the likelihood of observing each piece of data
- Example: The likelihood of observation 1 being a black ball if there are 3 black balls in the hat is (3/4)

Maximum likelihood
Introduction to a concept

- In our case, we evaluated our data (always the *same* data) using the three candidate values of $\widehat{\beta}^{ML}$
- It turned out that the value of β that made observing our data most likely was 3
- So we select $\widehat{\beta}^{ML} = 3$
- Q: How did we figure out which value made our data most likely?
- A: We evaluated the *likelihood function*
- Each data point contributes to the likelihood function. We build the likelihood function by multiplying together the likelihood of observing each piece of data
- Example: The likelihood of observation 1 being a black ball if there are 3 black balls in the hat is (3/4)
- Total likelihood is (the likelihood of observation 1) $\times$ (the likelihood of observation 2) $\times \ldots \times$ (the likelihood of observation N)

- This concept expands easily to other problems
- No matter what the model is, we always form the likelihood the same way:

$$L = (\text{likelihood of obs 1}) \times (\text{likelihood of obs 2}) \times \ldots$$
$$\ldots \times (\text{likelihood of obs N})$$

- The only thing that changes is how we determine the likelihood of observation *i*
- Let's take a look at how we would determine the likelihood of observation *i* in a bivariate regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- We're using the ML procedure to evaluate candidate $\widehat{\beta_0}$ and $\widehat{\beta_1}$
- We ask ourselves, with candidates $\widehat{\beta_0}$ and $\widehat{\beta_1}$, if we observe x_i, what y_i would we expect to observe?

- We're using the ML procedure to evaluate candidate $\widehat{\beta_0}$ and $\widehat{\beta_1}$
- We ask ourselves, with candidates $\widehat{\beta_0}$ and $\widehat{\beta_1}$, if we observe x_i, what y_i would we expect to observe?
- Consider just the first observation:
- If we observe x_1 and are evaluating candidate parameter estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$, what is the probability of observing the y_1 that we actually observed?
- (We want to express the probability of observing the *gdp_pc* of Albania — \$4,319.68 — given an average temperature of 15.03 degrees C, and parameters $\widehat{\beta_0} = 20094.04$ and $\widehat{\beta_1} = -606.65$)

- If average temperature is 15.03 degrees C, $\beta_0 = 20094.04$ and $\beta_1 = -606.65$, what does our model say that y (*gdp_pc*) will be?

$$gdp\_pc = \beta_0 + \beta_1 temp + u$$

- If average temperature is 15.03 degrees C, $\beta_0 = 20094.04$ and $\beta_1 = -606.65$, what does our model say that y (*gdp_pc*) will be?

$$gdp\_pc = \beta_0 + \beta_1 temp + u$$

- We start with 20094.04, then add -606.65*15.03 = 10976.09 = $\widehat{y}_{albania}$

- If average temperature is 15.03 degrees C, $\beta_0 = 20094.04$ and $\beta_1 = -606.65$, what does our model say that y (*gdp_pc*) will be?

$$gdp\_pc = \beta_0 + \beta_1 temp + u$$

- We start with 20094.04, then add -606.65*15.03 = 10976.09 = $\widehat{y}_{albania}$
- But we *didn't* observe the gdp per capita of Albania to be exactly $10,976.09 — in fact, we observed it to be $4,319.677
- What gives?

- If average temperature is 15.03 degrees C, $\beta_0 = 20094.04$ and $\beta_1 = -606.65$, what does our model say that y (*gdp_pc*) will be?

$$gdp\_pc = \beta_0 + \beta_1 temp + u$$

- We start with 20094.04, then add -606.65*15.03 = 10976.09 = $\widehat{y}_{albania}$
- But we *didn't* observe the gdp per capita of Albania to be exactly $10,976.09 — in fact, we observed it to be $4,319.677
- What gives?
- What is the probability of this happening (observing $4,319.677 when the model says we should observe $10,976.09)?

- Only one source of unresolved randomness here . . .

- Only one source of unresolved randomness here . . .
- There is only way that our model allows for a difference between the following two numbers:
  1. $\beta_0 + \beta_1 x_1$
  2. y
- What is it?

- Only one source of unresolved randomness here . . .
- There is only way that our model allows for a difference between the following two numbers:
  1. $\beta_0 + \beta_1 x_1$
  2. y
- What is it?
- u!

- Only one source of unresolved randomness here . . .
- There is only way that our model allows for a difference between the following two numbers:
  1. $\beta_0 + \beta_1 x_1$
  2. y
- What is it?
- u! u is the difference between what we actually observe — y — and what the model says y should be

- Only one source of unresolved randomness here . . .
- There is only way that our model allows for a difference between the following two numbers:
  1. $\beta_0 + \beta_1 x_1$
  2. y
- What is it?
- u! u is the difference between what we actually observe — y — and what the model says y should be
- With that in mind, let's restate our question

- Original question: What is the probability of observing \$4,319.677 when the model says we should observe \$10,976.09?
- Restated question: What is the probability of observing $u = 4,319.677 - 10,976.09 = -6,656.41$?

- Original question: What is the probability of observing \$4,319.677 when the model says we should observe \$10,976.09?
- Restated question: What is the probability of observing $u = 4,319.677 - 10,976.09 = -6,656.41$?
- Now this should start to look a lot more like a standard problem from statistics

- Original question: What is the probability of observing \$4,319.677 when the model says we should observe \$10,976.09?
- Restated question: What is the probability of observing $u = 4,319.677 - 10,976.09 = -6,656.41$?
- Now this should start to look a lot more like a standard problem from statistics
- In order to say anything about probabilities, we need to assume something about how u is distributed

- Original question: What is the probability of observing \$4,319.677 when the model says we should observe \$10,976.09?
- Restated question: What is the probability of observing $u = 4,319.677 - 10,976.09 = -6,656.41$?
- Now this should start to look a lot more like a standard problem from statistics
- In order to say anything about probabilities, we need to assume something about how u is distributed
- The standard assumption is that u is distributed normally (there are reasons for this that we won't get into at this point)

- Original question: What is the probability of observing \$4,319.677 when the model says we should observe \$10,976.09?
- Restated question: What is the probability of observing $u = 4,319.677 - 10,976.09 = -6,656.41$?
- Now this should start to look a lot more like a standard problem from statistics
- In order to say anything about probabilities, we need to assume something about how u is distributed
- The standard assumption is that u is distributed normally (there are reasons for this that we won't get into at this point)
- If we assume that u is distributed normally, we can form the likelihood function

- We want to know

$$p(-6,656.41) =$$

- We want to know

$$p(-6,656.41) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(u_i)^2\right)$$

- We want to know

$$p(-6,656.41) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(u_i)^2\right)$$

$$p(-6,656.41) =$$

- We want to know

$$p(-6,656.41) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(u_i)^2\right)$$

$$p(-6,656.41) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

- We want to know

$$p(-6,656.41) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(u_i)^2\right)$$

$$p(-6,656.41) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

$$p(-6,656.41) =$$

- We want to know

$$p(-6,656.41) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(u_i)^2\right)$$

$$p(-6,656.41) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

$$p(-6,656.41) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(-6,656.41)^2\right)$$

- We want to know

$$p(-6,656.41) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(u_i)^2\right)$$

$$p(-6,656.41) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

$$p(-6,656.41) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(-6,656.41)^2\right)$$

- We form the likelihood function by multiplying all of the observations together

$$L(y_1, \ldots, y_N) = \prod_{i=1}^{N} p(u_i)$$

- We form the likelihood function by multiplying all of the

$$L(y_1, \ldots, y_N) = \prod_{i=1}^{N} p(u_i)$$

- We form the likelihood function by multiplying all of the

$$L(y_1, \ldots, y_N) = \prod_{i=1}^{N} p(u_i)$$

- We want to maximize the likelihood (make the number L as big as possible) by choosing the right values of β_0 and β_1
- Same as maximizing

$$\sum_{i=1}^{N} \log(p(u_i)) = \sum_{i=1}^{N} -\frac{1}{2}(y_i - \beta_0 - \beta_1 x_i)^2$$

- We form the likelihood function by multiplying all of the

$$L(y_1, \ldots, y_N) = \prod_{i=1}^{N} p(u_i)$$

- We want to maximize the likelihood (make the number L as big as possible) by choosing the right values of β_0 and β_1
- Same as maximizing

$$\sum_{i=1}^{N} \log(p(u_i)) = \sum_{i=1}^{N} -\frac{1}{2}(y_i - \beta_0 - \beta_1 x_i)^2$$

- Goal is to maximize the *negative* sum of squared errors

- We form the likelihood function by multiplying all of the

$$L(y_1, \ldots, y_N) = \prod_{i=1}^{N} p(u_i)$$

- We want to maximize the likelihood (make the number L as big as possible) by choosing the right values of β_0 and β_1
- Same as maximizing

$$\sum_{i=1}^{N} \log(p(u_i)) = \sum_{i=1}^{N} -\frac{1}{2}(y_i - \beta_0 - \beta_1 x_i)^2$$

- Goal is to maximize the *negative* sum of squared errors
- This is just like *minimizing* the sum of squared errors

- Goal is to maximize the *negative* sum of squared errors
- This is just like *minimizing* the sum of squared errors

- Goal is to maximize the *negative* sum of squared errors
- This is just like *minimizing* the sum of squared errors
- So without doing a single stitch of additional math, we know that we are going to get a formula for our maximum likelihood estimators $\widehat{\beta}^{MLE}$ that is identical to what we already found for $\widehat{\beta}^{OLS}$ and $\widehat{\beta}^{MM}$

- Moral of the story?
- We start with a very simple model relating one variable to another:

$$y = \beta_0 + \beta_1 x + u$$

- . . . we use 3 different sets of reasoning (and assumptions) to turn our data and this single model into estimates of β_0 and β_1:
  1. Minimize the sum of squared errors (a "goodness of fit" criteria)
  2. Make true in the data things we think should be true: (1) our predictions are right *on average* and (2) errors are random
  3. Make observing our data the most *likely* event

- Moral of the story?
- We start with a very simple model relating one variable to another:

$$y = \beta_0 + \beta_1 x + u$$

- . . . we use 3 different sets of reasoning (and assumptions) to turn our data and this single model into estimates of β_0 and β_1:
  1. Minimize the sum of squared errors (a "goodness of fit" criteria)
  2. Make true in the data things we think should be true: (1) our predictions are right *on average* and (2) errors are random
  3. Make observing our data the most *likely* event
- . . . and all three lead to the same estimator. Pretty cool.

Making this all make sense: putting our estimators to work

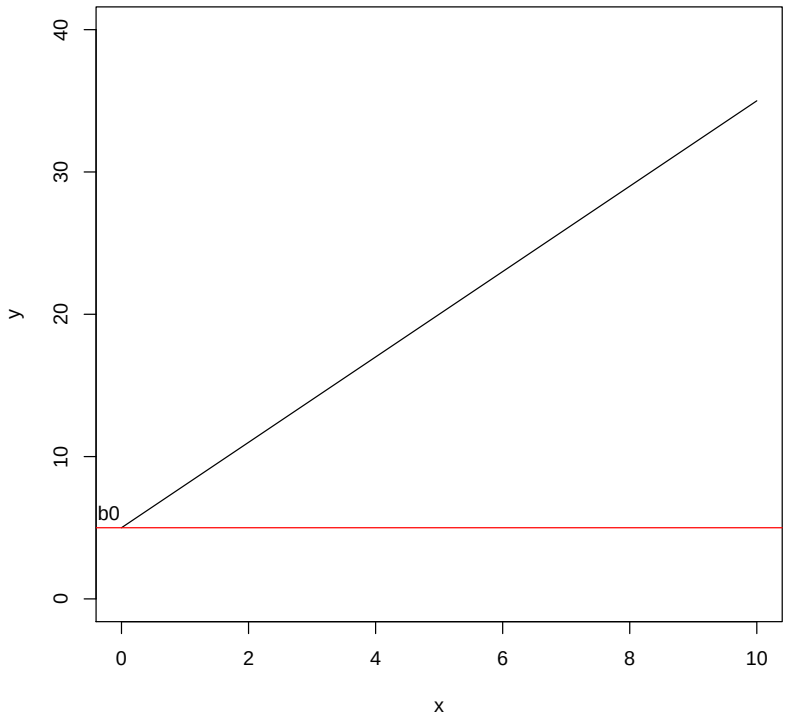
- So far we have gone over:
  1. The derivation of the univariate model estimator (β_0 only) using both the method of moments and least squares
  2. The derivation of the bivariate model estimator (β_0 and β_1) using method of moments least squares
  3. The derivation (most of it) of the bivariate model estimator using the method of maximum likelihood
- Goal now: translating the bivariate estimator into something that is interpretable
- But first, we need to remember why we're doing this in the first place

Bivariate regression model

- Lots of examples of relationships like this
- From Wooldridge:
  - crop yield and fertilizer
  - wage and years of education
  - crime rate and number of police
  - campaign donations and votes
- From (some of) your own specializations:
  - international aid and gdp growth (Easterly, 2003 JEP)
  - access to credit and poverty rate (Newman (1993) and Galor and Zeira (1993))
  - gdp and pollution (EKC; Frankel and Rose (2005))
  - minimum wage and unemployment (Card and Krueger 1994)
  - Walmart and job creation (Basker 2005)
- Note that in these published papers there are always more variables, but the relationship of interest is often between just two

Parameters of the bivariate model

- Realize that we are estimating a linear model. This makes β_0 the *intercept* (remember $y = mx + b$ from 8th grade?)
- β_0 is the intercept — this is the value of y when x is 0
- β_1 is the slope



Parameters of the bivariate model

- The "Change"-interpretation
  - β_0 is just a constant
  - β_0 is often not the parameter of interest, but it is hugely important that it be included in the model (for reasons we'll discover soon and throughout the course)
  - β_1 represents the relationship between two variables we are interested in (like the returns to education, i.e. earnings and education)
  - β_1 tells us how much y changes when we change x by some amount

- β_1 tells us how much y changes when we change x by some 1 unit
- Example

$$earnings = \beta_0 + \beta_1 education + u$$

- Education is measured in years and earnings is measured in $ per year

- β_1 tells us how much y changes when we change x by some 1 unit
- Example

$$earnings = \beta_0 + \beta_1 education + u$$

- Education is measured in years and earnings is measured in $ per year
- β_1 = extra dollars of earnings from 1 extra year of education

- β_1 tells us how much y changes when we change x by some 1 unit
- Example

$$earnings = \beta_0 + \beta_1 education + u$$

- Education is measured in years and earnings is measured in $ per year
- β_1 = extra dollars of earnings from 1 extra year of education
- Another way to refer to β_1 is the "marginal effect" of x on y
- The marginal change in dollars of earnings from a small change (a 1-unit or 1-year change) in education

What is random and what is fixed?

- What is random and what is fixed?

$$y = \beta_0 + \beta_1 x + u$$

$$earnings = \beta_0 + \beta_1 education + u$$

What is random and what is fixed?

- What is random and what is fixed?

$$y = \beta_0 + \beta_1 x + u$$

$$earnings = \beta_0 + \beta_1 education + u$$

- u

What is random and what is fixed?

- What is random and what is fixed?

$$y = \beta_0 + \beta_1 x + u$$

$$earnings = \beta_0 + \beta_1 education + u$$

- u is definitively not fixed (this is intuitive — u is a mashup of unobservables, which is easy to think of as "random")

What is random and what is fixed?

- What is random and what is fixed?

$$y = \beta_0 + \beta_1 x + u$$

$$earnings = \beta_0 + \beta_1 education + u$$

- u is definitively not fixed (this is intuitive — u is a mashup of unobservables, which is easy to think of as "random")
- What about y?

What is random and what is fixed?

- What is random and what is fixed?

$$y = \beta_0 + \beta_1 x + u$$

$$earnings = \beta_0 + \beta_1 education + u$$

- u is definitively not fixed (this is intuitive — u is a mashup of unobservables, which is easy to think of as "random")
- What about y?
- Since y is a function of u, y is random
  - last two digits of the amplitude of atmospheric noise is random
  - the number of ounces in a pint is not random
  - 16 + the last two digits of the amplitude of atmospheric noise is random (just like the coin-flipping example from last lecture)

What is random and what is fixed?

- What about x?

What is random and what is fixed?

- What about x?
- Depending on the situation it can be easier to think of x as either random or fixed, but it is safer to think of it always as random
- Why safer?

What is random and what is fixed?

- What about x?
- Depending on the situation it can be easier to think of x as either random or fixed, but it is safer to think of it always as random
- Why safer?
- Because if we can deal with the slight complications of thinking of x as random, then the same techniques work when x is fixed
- We will learn how to deal with random x — it is more conservative to think of x as random
- Example: is the number of years of education of a given individual totally explainable by observable factors? No? Then it's *random* for our current intents and purposes

What is random and what is fixed?

- What about x?
- Depending on the situation it can be easier to think of x as either random or fixed, but it is safer to think of it always as random
- Why safer?
- Because if we can deal with the slight complications of thinking of x as random, then the same techniques work when x is fixed
- We will learn how to deal with random x — it is more conservative to think of x as random
- Example: is the number of years of education of a given individual totally explainable by observable factors? No? Then it's *random* for our current intents and purposes
- Finally, what about β_0 and β_1?

What is random and what is fixed?

- β_0 and β_1 are the *true* parameters of the model
- The β's are fixed
- Why?

What is random and what is fixed?

- β_0 and β_1 are the *true* parameters of the model
- The β's are fixed
- Why?
- The β's are part of our model — they are things that we believe exist, but which we can never observe
- Just as when we take the sample mean of some data ($\bar{y}$) in an attempt to try to estimate the true mean of a population, we now try to estimate the true parameters of our model (β_0 and β_1)

The bivariate model: summing up

- y and x are observable data, which we think of as random
- u is unobservable data, which we think of as random
- The β's are parameters of the model, which we think of as fixed

The bivariate model: summing up

- y and x are observable data, which we think of as random
- u is unobservable data, which we think of as random
- The β's are parameters of the model, which we think of as fixed
- Our goal is to take some data and develop a procedure that we can use to take the data and turn it into estimates of β_0 and β_1
- How to translate data into estimates?

$$data + model \rightarrow \boxed{procedure} \rightarrow estimates$$

Terminology

- y: dependent variable, regressand (rare)
- x_k: independent variable, regressor, covariate, control variable
- β_0: intercept
- β_k: parameter associated with x_k (slope parameters)
- u: unobservables ("errors" (yuck), "disturbance" (double-yuck))

Regression modeling

- Modeling a relationship is like specifying how y is produced
- (I find it helpful to think of the term "Data Generating Process")
- We think that y is determined in part by some variable x
- We create some model of how y is determined (in part) by a function of x

$$y = f(x, \text{ other stuff})$$

- We typically estimate $f(x)$ by a linear function (because linear functions are simple and work pretty darn well)
- The workhorse model we employ is a linear model that relates y to x (or many x's)

$$y = \beta_0 + \beta_1 x + u$$

- A quick sidebar on linearity
- We frequently say that the function

$$y = \beta_0 + \beta_1 x + u$$

is linear. What do we mean?

- A quick sidebar on linearity
- We frequently say that the function

$$y = \beta_0 + \beta_1 x + u$$

  is linear. What do we mean?
- We mean that it is linear *in the parameters* — this does not mean that we cannot represent non-linear relationships between y and x with this model
- It means that each parameter is separate from the other parameters, and multiplies a unique term

- To understand what we mean when we say that a model is linear, it might be helpful to look at a model that is *not* linear
- What would a *nonlinear* model look like?

$$y = \left(\frac{\beta_1 x}{\beta_0 + x} \right)^u$$

- I just made that up. It doesn't mean anything, but it is *nonlinear*
- We'll talk more about functional form later — for now I just wanted to clear up what we mean when we say "linear"

Regression modeling

- Back to business . . .
- The regression model does a couple of things for us
  1. It tells us how y and x are related
  2. It allows us to predict y
- To predict y, we first estimate the parameters of the model, the $\widehat{\beta}$'s
- Once we have our estimated parameters, we "plug them in" to our model to obtain predicted y's

$$\widehat{y} = \widehat{f(x)} = \widehat{\beta_0} + \widehat{\beta_1} x$$

- That's all pretty straightforward. So, how do we get these predictions?

- We get these predictions by selecting our favorite values for $\widehat{\beta_0} and \widehat{\beta_1}$. How do we select our favorite values?
- Well, we look at our data and it looks like this

y	x
y_1	x_1
y_2	x_2
y_3	x_3
$\vdots$	$\vdots$
y_n	x_n

y	x	candidate parameters	candidate predictions	"errors"
y_1	x_1			
y_2	x_2			
y_3	x_3			
$\vdots$	$\vdots$			
y_n	x_n			

y	x	candidate parameters	candidate predictions	"errors"
y_1	x_1	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$		
y_2	x_2	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$		
y_3	x_3	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$		
$\vdots$	$\vdots$			
y_n	x_n	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$		

y	x	candidate parameters	candidate predictions	"errors"
y_1	x_1	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_1}$	
y_2	x_2	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_2}$	
y_3	x_3	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_3}$	
$\vdots$	$\vdots$		$\vdots$	
y_n	x_n	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_n}$	

y	x	candidate parameters	candidate predictions	"errors"
y_1	x_1	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_1}$	$(y_1 - \widehat{y_1})$
y_2	x_2	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_2}$	$(y_2 - \widehat{y_2})$
y_3	x_3	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_3}$	$(y_3 - \widehat{y_3})$
$\vdots$	$\vdots$		$\vdots$	
y_n	x_n	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_n}$	$(y_n - \widehat{y_n})$

y	x	candidate parameters	candidate predictions	"errors"
y_1	x_1	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_1}$	$(y_1 - \widehat{y_1})$
y_2	x_2	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_2}$	$(y_2 - \widehat{y_2})$
y_3	x_3	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_3}$	$(y_3 - \widehat{y_3})$
$\vdots$	$\vdots$		$\vdots$	
y_n	x_n	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_n}$	$(y_n - \widehat{y_n})$

Now we need some sort of rule

y	x	candidate parameters	candidate predictions	"errors"
y_1	x_1	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_1}$	$(y_1 - \widehat{y_1})$
y_2	x_2	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_2}$	$(y_2 - \widehat{y_2})$
y_3	x_3	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_3}$	$(y_3 - \widehat{y_3})$
$\vdots$	$\vdots$		$\vdots$	
y_n	x_n	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_n}$	$(y_n - \widehat{y_n})$

Now we need some sort of rule

- We use a rule to determine which of the (infinite number of) potential parameters to accept

y	x	candidate parameters	candidate predictions	"errors"
y_1	x_1	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_1}$	$(y_1 - \widehat{y_1})$
y_2	x_2	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_2}$	$(y_2 - \widehat{y_2})$
y_3	x_3	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_3}$	$(y_3 - \widehat{y_3})$
$\vdots$	$\vdots$		$\vdots$	
y_n	x_n	$\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$	$\widehat{y_n}$	$(y_n - \widehat{y_n})$

Now we need some sort of rule $\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$

- We use a rule to determine which of the (infinite number of) potential parameters to accept

- So far have been doing this:

$$y = \beta_0 + \beta_1 x + u$$

- we are bored with it now
- Almost never do bivariate regression in practical work

- So far have been doing this:

$$y = \beta_0 + \beta_1 x + u$$

- we are bored with it now
- Almost never do bivariate regression in practical work
- Now we are doing this:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- or this:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

Multivariate regression

- The difference between executing a bivariate and a multivariate regression in R (or Stata or whatever)?
- Nill

Multivariate regression

- The difference between executing a bivariate and a multivariate regression in R (or Stata or whatever)?
- Nill
- The difference between running the regression

$$birthWeight = \beta_0 + \beta_1 cigsPerDay + u$$

and

$$birthWeight = \beta_0 + \beta_1 cigsPerDay + \beta_2 methConsumption + u?$$

Multivariate regression

- The difference between executing a bivariate and a multivariate regression in R (or Stata or whatever)?
- Nill
- The difference between running the regression

$$birthWeight = \beta_0 + \beta_1 cigsPerDay + u$$

and

$$birthWeight = \beta_0 + \beta_1 cigsPerDay + \beta_2 methConsumption + u?$$

R code

```
lmer(birthWeight ~cigsPerDay)
lmer(birthWeight ~cigsPerDay + meth)
```

Multivariate regression

- So *doing* multivariate regression is easy (this is true)
- But it brings complications
- First things first: If our ONLY goal was to predict y (and not to interpret relationships between x and y), then you should throw in the kitchen sink, and forget about interpreting the coefficient estimates

Multivariate regression

- So *doing* multivariate regression is easy (this is true)
- But it brings complications
- First things first: If our ONLY goal was to predict y (and not to interpret relationships between x and y), then you should throw in the kitchen sink, and forget about interpreting the coefficient estimates
- So why include multiple x's at the same time? And why be careful about what x's we include/exclude in a model?

Multivariate regression

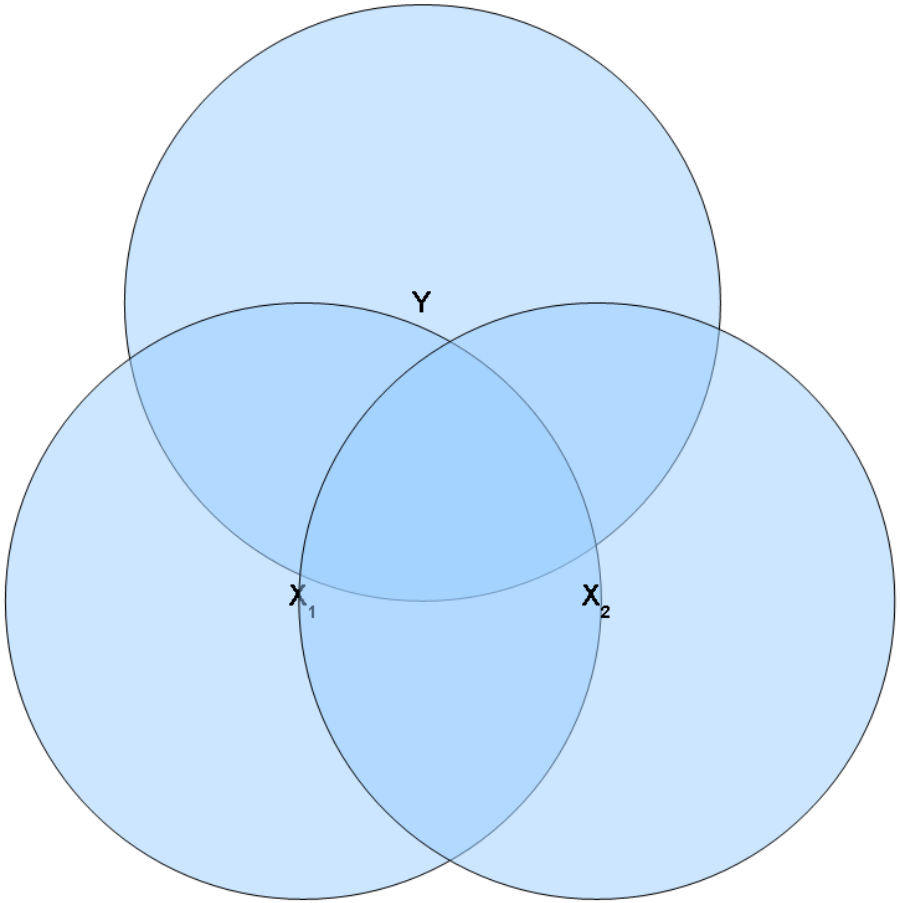
- So *doing* multivariate regression is easy (this is true)
- But it brings complications
- First things first: If our ONLY goal was to predict y (and not to interpret relationships between x and y), then you should throw in the kitchen sink, and forget about interpreting the coefficient estimates
- So why include multiple x's at the same time? And why be careful about what x's we include/exclude in a model?
- Because:
- We would like to be able to interpret $\widehat{\beta_k}$ as our best estimate of the relationship between y and x_k (we don't just want to *predict* poverty — we want to be able to say how poverty *changes*). Policy is all about *changes*.
- We would like to know how each x-variable *individually* relates to the single y-variable

- Another way of saying the same thing ...
- We would like to be able to interpret $\widehat{\beta_k}$ as our best estimate of the relationship between y and x_k, *holding other relevant factors constant*
- Including relevant (related) variables allows us to *hold constant* these other factors, so we can focus on the relationship between y and x_k

Multivariate OLS, graphically

- Ballantine time!

Multivariate OLS, graphically
Why is it called the Ballantine?

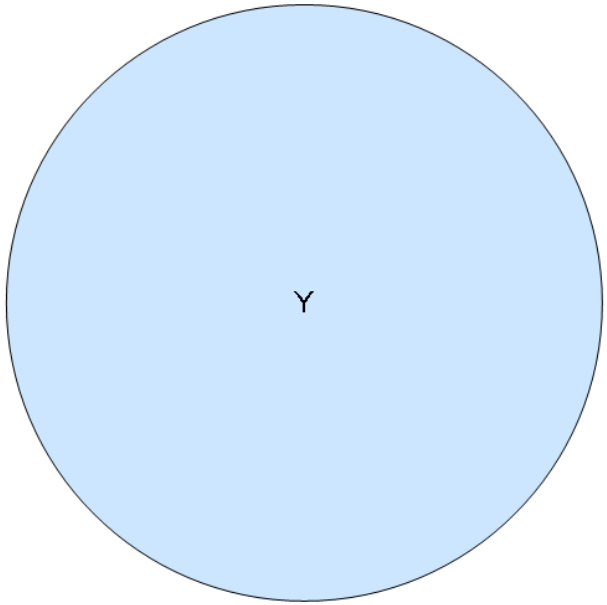


Multivariate OLS, graphically
Why is it called the Ballantine?

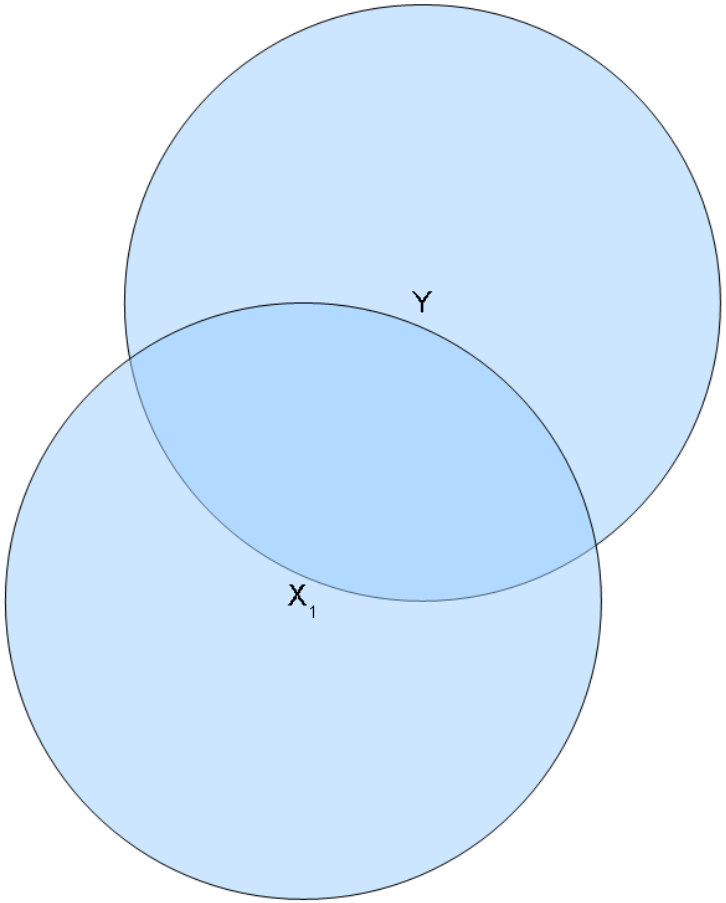
Multivariate OLS, graphically
What the Ballantine is for

- The Ballantine (which we are about to introduce) is useful for thinking about statistical relationships
- I find the Ballantine to be one of the best representations of variance, covariance, and statistical relation between a group of variables
- But it is important to realize that the Ballantine has limits. The Ballantine is a diagram that makes new concepts easier to digest. The Ballantine cannot be used to *prove* anything, and thinking too much using the Ballantine as your only source of intuition is dangerous
- The Ballantine is for: thinking about statistical co-movements
- The Ballantine is NOT for: thinking about the magnitude of systematic relationships

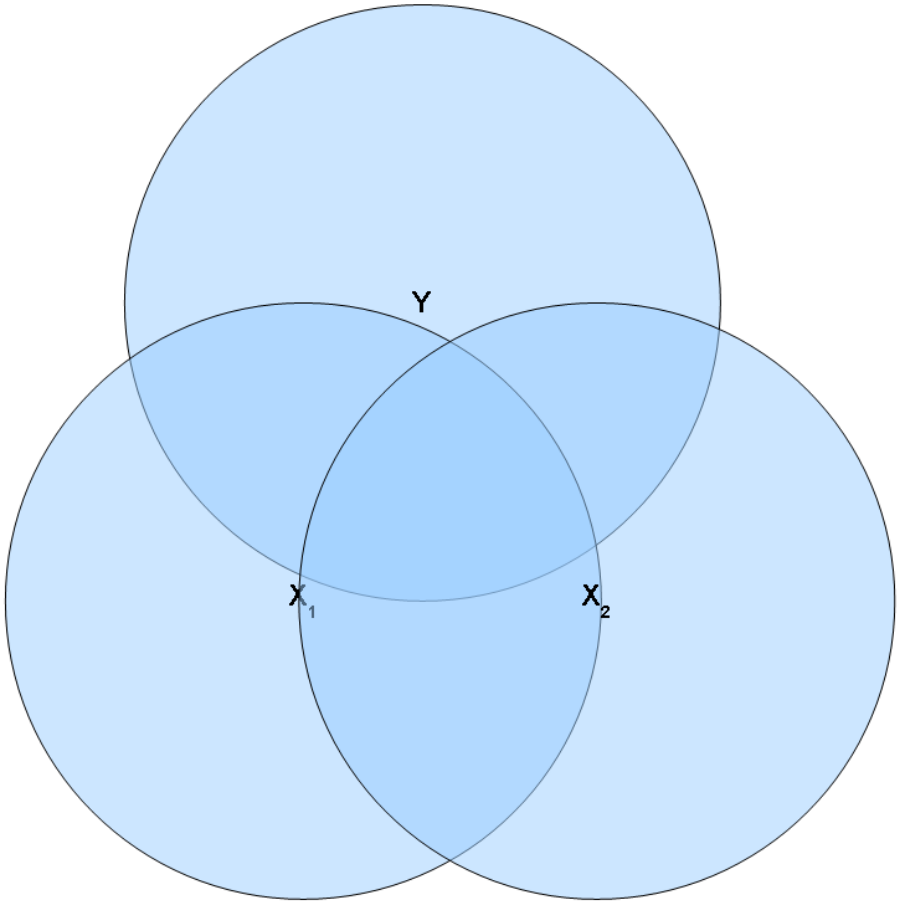
Multivariate OLS estimator
Graphically



Multivariate OLS estimator
Graphically



Multivariate OLS estimator
Graphically



Replication project
Interim assignment 1

- After today you will know everything you need to know to get through Horowitz
- After going to TA session and/or exploring R yourself (Leada!), you will know what you need to know to start the project
- Before that, get started on collecting data . . .

Replication project
Interim assignment 1

- 90% of econometric work is working with data
- Get to work collecting GDP and population data for all the countries in the sample. Collect 2010- most recently available data.
- Send me (Cc'd to Robert) three things:
  1. an R dset (a .RData-file) containing four variables: the name (or abbreviation) of each country (a "string" variable), and GDP in 2010, 2011, and 2012 (NOTE the update here to 2010-2012)
  2. the data in its original form you download from the internet (csv works really well)
  3. the r-file that accomplishes the importation of the data into R, and makes any changes in the way the data is organized
- Due next week (before class!)

Replication project
Interim assignment 1

- This is the smallest baby step you can take to get you going — GDP and population data is relatively easy data to find
- Don't hesitate to keep collecting more data (having read the first few sections of the paper, you will know what other data you need to get)
- You're going to have lots of questions/struggles. Let's get them out of the way early.