# Econometrics

## Lecture 4

Nathaniel Higgins

ERS and JHU

26 September 2011

# Homework

- Any questions?

- We might think that $y$ is related to $x$ by a certain function
- We are interested in how $y$ is related to $x$ *on average*
- This notion is captured in what is called the conditional expectation function

$$E(y|x) = \beta_0 + \beta_1 x$$

- The conditional expectation function gives the expected value of $y$ (the variable that is being affected), *conditional on $x$*
- The term *conditional on $x$* means “given what we know about $x$”

# Regression

## Conditional expectation

- The term *conditional on x* means “given what we know about $x$”
- Forget about regressions for just a moment. Suppose we know that the expected value of income in the U.S. for an individual is $30k
- If you pick a random person out of the population (totally random!) and you have to guess their income, what would your guess be? Probably $30k
- Now suppose that before you guessed, you were given one extra fact about the person: they live in Malibu. Would you guess $30k? Probably not. You would guess higher
- The first guess (before you knew anything about the randomly drawn person) is akin to $E(y)$
- The second guess (after you knew that they lived in Malibu) is akin to $E(y|x)$, where $x$ represents where the individual lives

- Get it?
- The conditional expectation is what we expect given all the knowledge and information that we have
- This is a very useful concept
- Remember that this was the representation of the conditional expectation function that we had

$$E(y|x) = \beta_0 + \beta_1 x$$

- If we had all the information — if nothing was left out of the equation explaining $y$ — then this would be enough
- For every value of $x$ it would be straightforward to know what $y$ was. If $x = 5$, then $y = \beta_0 + \beta_1 * 5$. Period.
- But this is unrealistic. We don't know everything about how $y$ is determined. So there is some extra randomness, which we add into the model with $u$

$$E(y|x) = \beta_0 + \beta_1 x + u$$

- We start with a variable that is “caused” or “determined.” We call it $y$.
- We think that $y$ is determined in part by some variable $x$.
- We create some model of how $y$ is determined by a function of $x$

$$y = f(x)$$

  - We typically estimate $f(x)$ by a linear function (because they are simple and work pretty darn well)
  - The typical example is $f(x) = \beta_0 + \beta_1 x$
- We get some data and estimate the relationship so that we get a prediction of what $y$ is for every value of $x$

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

- 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 |
|---|---|
| $y_1$ | $x_1$ |
| $y_2$ | $x_2$ |
| $y_3$ | $x_3$ |
| $\vdots$ | $\vdots$ |
| $y_n$ | $x_n$ |

# Regression

## Summary

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

# Regression
## Summary

| y | x | |
|---|---|---|
| $y_1$ | $x_1$ | $\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$ |
| $y_2$ | $x_2$ | $\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$ |
| $y_3$ | $x_3$ | $\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$ |
| $\vdots$ | $\vdots$ | |
| $y_n$ | $x_n$ | $\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$ |

# Regression

## Summary

$$\begin{array}{c|cccc}
y & x & & & \\
\hline
y_1 & x_1 & \rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow & & \widehat{y_1} \\
y_2 & x_2 & \rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow & & \widehat{y_2} \\
y_3 & x_3 & \rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow & & \widehat{y_3} \\
\vdots & \vdots & & & \vdots \\
y_n & x_n & \rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow & & \widehat{y_n} \\
\hline
\end{array}$$

# Regression
## Summary

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

# Regression

## Summary

| y | x | | | "errors" |
|---|---|---|---|---|
| $y_1$ | $x_1$ | $\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$ | $\widehat{y_1}$ | $(y_1 - \widehat{y_1})^2$ |
| $y_2$ | $x_2$ | $\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$ | $\widehat{y_2}$ | $(y_2 - \widehat{y_2})^2$ |
| $y_3$ | $x_3$ | $\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$ | $\widehat{y_3}$ | $(y_3 - \widehat{y_3})^2$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | |
| $y_n$ | $x_n$ | $\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$ | $\widehat{y_n}$ | $(y_n - \widehat{y_n})^2$ |

# Regression

## Summary

| y | x | | | "errors" |
|---|---|---|---|---|
| $y_1$ | $x_1$ | $\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$ | $\widehat{y_1}$ | $(y_1 - \widehat{y_1})^2$ |
| $y_2$ | $x_2$ | $\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$ | $\widehat{y_2}$ | $(y_2 - \widehat{y_2})^2$ |
| $y_3$ | $x_3$ | $\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$ | $\widehat{y_3}$ | $(y_3 - \widehat{y_3})^2$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | |
| $y_n$ | $x_n$ | $\rightarrow (\widehat{\beta_0}, \widehat{\beta_1}) \rightarrow$ | $\widehat{y_n}$ | $(y_n - \widehat{y_n})^2$ |

$$\sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

# Regression

- We select the values of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ that minimize the squared difference between $y$ and $\widehat{y}$

$$(y - \widehat{y})^2$$

- That's it!
- Now we have two things:
  1. Point estimates of the true parameters $\beta_0$ and $\beta_1$
  2. Predictions of $y$ for every value of $x$ in our data
- Before we explore anything else, let's try some basic regressions

# Try it

- Load up some data from Wooldridge

**Stata code**

```
use ceosal1.dta
* What is in the data?

* What are the basic statistics of the data?
```

# Try it

- Load up some data from Wooldridge

**Stata code**

```
use ceosal1.dta
* What is in the data?
describe
* What are the basic statistics of the data?
```

# Try it

- Load up some data from Wooldridge

**Stata code**

```
use ceosal1.dta
* What is in the data?
describe
* What are the basic statistics of the data?
sum
```

# Try it

**Stata code**

```
* We are interested in the relationship
* between CEO salary (salary) and return on
* investment (roe)
* Let’s look at how they appear to be related
* by examining a simple scatter of the data
scatter salary roe
* What is the first thing that you notice?
```

# Try it

### Stata code

```
* Eliminate some of the outliers
* so that we can get a better look
scatter salary roe if (salary < 5000)
```

# Try it

### Stata code

```
* Now that we have a rough idea of what
* the data look like, let’s run a model
reg salary roe
```

# Try it

## Stata code

```
* Let's draw the model on top of the data
* First, let's get the predictions of the
* model
predict yhat
* Now plot yhat against roe
```

# Try it

### Stata code

```
* Let's draw the model on top of the data
* First, let's get the predictions of the
* model
predict yhat
* Now plot yhat against roe
scatter salary roe if salary < 5000 ///
|| scatter yhat roe
```

# Try it

### Stata code

```
* One last thing.  We might want to see a line
* (rather than a scatter) of the predicted
* values
scatter salary roe if salary < 5000 ///
|| line yhat roe
```

# Summing up

- What is the interpretation of our estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$?
- Now, that's all well and good. But you might have noticed there are some other things printed to the screen when you type `reg salary roe`.
- We move next to these other things.
- Look again at the results on your screen

# The results on your screen

```
                                          ___  ____  ____  ____  ____(R)
                                         /__    /   ____/   /   ____/
                                        ___/   /   /___/   /   /___/
                                          Statistics/Data Analysis

. reg salary roe

      Source |       SS       df       MS              Number of obs =     209
-------------+------------------------------           F(  1,   207) =    2.77
       Model |  5166419.04     1  5166419.04           Prob > F      =  0.0978
    Residual |   386566563   207  1867471.32           R-squared     =  0.0132
-------------+------------------------------           Adj R-squared =  0.0084
       Total |   391732982   208  1883331.64           Root MSE      =  1366.6

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         roe |   18.50119   11.12325     1.66   0.098    -3.428196    40.43057
       _cons |   963.1913   213.2403     4.52   0.000     542.7902    1383.592
------------------------------------------------------------------------------
```

# The results on your screen

- There is a lot there. Some of the stuff is ignorable for now. But let's start with a few things you can immediately understand
- Look in the upper-right. `Number of obs` is *n*: the number of observations
- We also see the coefficients corresponding to `roe` ($\widehat{\beta_1}$) and `_cons` ($\widehat{\beta_0}$)
- These are the things we have covered so far. The next thing we want to look at are the numbers under the column heading `SS` in the upper-left of the results

# Sum of squares

## TSS

- `SS` stands for sum of squares
- There are 3 different "SS" entries:
  1. Model
  2. Residual
  3. Total
- Let's deal with the Total SS first
  - The Total SS is the sum of squared differences between $y$ and its mean $\bar{y}$
  - This is called the Total SS (or TSS) because it represents all the variation of $y$
  - This variation in $y$ is what the model is trying to explain!
  - When we run our econometric model, we are trying to explain movements in $y$ with movements in $x$ $\rightarrow$ we cannot explain more than the total variation in $y$ that we observe
- TSS $= \sum(y_i - \bar{y}_i)^2$

# Sum of squares

## TSS

- You can picture the variation in $y$ as being represented by a circle — the bigger the circle, the more variation there is to explain
- FIGURE

# Sum of squares

## MSS/SSE

- Just to confuse you, Stata and Wooldridge don't agree on what to call the next thing
- Stata: Model SS (MSS); Wooldridge: Explained SS (SSE)
- Either way, they mean the same thing: the sum of the difference between $\hat{y}$ and its mean $\bar{\hat{y}}$
- The variation in our predicted $y$'s is less (often much less) than the variation in the original $y$'s
- This is because our model doesn't predict all of the movement in $y$ — our model isn't all powerful
- Picture the MSS (or SSE) as a circle also

# Sum of squares
## R-squared

- The overlap of these two circles represents the proportion of the total variation in $y$ that is explained by our model (explained by $x$)
- There is a single number that summarizes this information:

$$R^2 = \frac{MSS}{TSS} \text{ or } \frac{SEE}{TSS}$$

- Try calculating for yourself

Stata

Wooldridge

# Try it

## R-squared

**Stata code**

```
* Notice that lots of good stuff is saved
* whenever we run a regression.  Check out the
* bottom of the help browser after typing:
help reg
```

# Try it

## R-squared

**Stata code**

```
* Notice that lots of good stuff is saved
* whenever we run a regression.  Check out the
* bottom of the help browser after typing:
help reg
scalar TSS = e(mss) + e(rss)
scalar rSq = e(mss)/TSS
```

# Sum of squares

## R-squared

- You should notice that the number you got for $R^2$ is the same as what is printed in the upper right-hand column of your Stata results after you executed the command
  `reg salary roe`

# A word about functional form

- Everything we have learned so far has been about linear models

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

- We say that the model above is *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*

# A word about functional form

- We don't know what to do with a model like that yet
- There is a thing called nonlinear least squares (NLLS, as opposed to OLS). But we aren't there yet.
- There is a small set of nonlinear functional forms that you *can* work with
- . . . because you can transform them into linear functional forms
- Consider an exponential function: check out the Stata example
- (you've heard of "exponential growth" right?)

# Stata example

## Stata code

```
* Going to create some data with 100 obs
set obs 100
* Generate some random x data
gen x = 3*runiform()
* Generate random unobservable data
gen u = rnormal()
* Generate a nonlinear y
gen y = exp(x) + u
* Look what happens when you regress y on x
* when the relationship is nonlinear
scatter y x || lfit y x
```

# Stata example

### Stata code

```
* How to solve the problem?  Transform to
* a linear model
gen logY = log(y)
* Now regress the new variable on x
scatter logY x || lfit logY x
```

# Why do we like OLS so much?

## Properties of the OLS estimator

- It is linear (simplicity; works well in lots of situations)
- Unbiased
- As it happens, the OLS estimator is the *best* linear unbiased estimator that we could possibly come up with!
- There is an acronym for this: BLUE
- The linearity of OLS is inherent; the unbiasedness and the “bestness” of OLS are based on some assumptions that we should state and discuss

# Why do we like OLS so much?

## Assumptions

- Assumptions necessary to insure that OLS is BLUE
- Unobservables are conditionally independent from $x$

$$E(u|x) = 0$$

- $y$ and $x$ are randomly sampled (“i.i.d”)
- $y$ and $y$ have finite variance

# Why do we like OLS so much?

## Unbiasedness

- If we make these assumptions, we know that OLS is *unbiased*
- Recall what *unbiased* means:
  - An estimator is unbiased if it is equal to the true value, *on average*
  - Remember, our estimate of $\beta_1$, which we call $\widehat{\beta_1}$ is itself a random variable, so it doesn't take on the same value every time
  - If $\beta_1$ is a single true number, and $\widehat{\beta_1}$, is a random variable, then sometimes $\widehat{\beta_1}$ takes on values different from $\beta_1$. That is, sometimes $\widehat{\beta_1}$ is *wrong*
  - But if $\widehat{\beta_1}$ is right on average, then $\widehat{\beta_1}$ is unbiased

# Why do we like OLS so much?

## Unbiasedness

- The proof of unbiasedness is beyond the scope of this class
- If you plan to continue with econometrics at a high level, try to work through the proof on page 50 of Wooldridge
- A lot of the proof is "just math," so to speak, but there is a wee bit of intuitive content

# Why do we like OLS so much?

## Unbiasedness

- The formula for the expectation of $\widehat{\beta}$ works out to

$$E(\widehat{\beta}) = \beta + f(x, u)$$

- The expected value of our estimator is equal to the true value of the parameter, plus some other stuff
- The other stuff is a function of the $x$ data and the unobservables
- If the $x$ data and the unobservables are not related at all (remember the assumption that $E(u|x) = 0$?) then the term $f(x, u)$ is equal to zero and we are left with

$$E(\widehat{\beta}) = \beta$$

# Last thing

## Causality

- We have several other properties of the OLS estimator to talk about once we get to the multivariate model
- The last thing we need to discuss before we take that leap is *causality*
- We have been looking at a simple model that looks at how much of the movement in $y$ can be explained by $x$
- When we say that $y$ is *explained by* $x$, it is natural to interpret this as causality: $x$ does this, so $y$ does that
- The interpretation of the OLS model is not, however, always causal
- The assumption that $E(u|x) = 0$ is key to this issue

# Causality

## An illustrative example

- Do hospitals make people feel better (make them less sick)?
- Said another way: does a hospital visit increase health?
- Seems light a straightforward question with an intuitively straightforward answer

# Causality

## An illustrative example

- Suppose we set up a survey; we have survey administrators set up at Dupont circle
- Survey administrators ask people how they are feeling; let's say they report how they are feeling on a 100 point scale (`health`)
- Survey administrators also ask people how many times they have visited the hospital in the last month (`hospital`)
- We run our favorite model (which is easy to select, since we only know 1 right now!)

$$\text{health} = \beta_0 + \beta_1 \text{hospital} + u$$

- We find that $\widehat{\beta_1} < 0$!
- What is going on?

# Causality

## An illustrative example

- Why do people go to the hospital?

# Causality

## An illustrative example

- Why do people go to the hospital?
- Because they are sick!

# Causality

## An illustrative example

- Why do people go to the hospital?
- Because they are sick!
- So let's say that folks who go to the hospital (because they are sick) go to the hospital feeling like they about a 20 (on a 100 point scale)
- They visit the hospital and the hospital makes them feel better
- Say the hospital makes them feel 20 points better (not bad!)
- So what's the average health of a hospital-goer now? A: 40
- What do you think the average answer of a non-hospital-goer would be?

# Causality

## An illustrative example

- The folks who didn’t go to the hospital were fine to begin with
- Suppose they had an average score of 60

# Causality

## An illustrative example

- The folks who didn't go to the hospital were fine to begin with
- Suppose they had an average score of 60
- If we run a regression with this data, we are comparing sick people to healthy people in order to try to find out if hospitals make people feel better
- So you see the problem intuitively
- *This* is the embodiment of a situation when $E(u|x) \neq 0$

# Causality

## An illustrative example

- The folks who didn't go to the hospital were fine to begin with
- Suppose they had an average score of 60
- If we run a regression with this data, we are comparing sick people to healthy people in order to try to find out if hospitals make people feel better
- So you see the problem intuitively
- *This* is the embodiment of a situation when $E(u|x) \neq 0$
- The unobservable term $u$ contained a key piece of information that wasn't in our data: how the person was feeling when they went to the hospital in the first place, i.e. whether or not they were sick
- The unobservables were related to the $x$ variable (number of hospital visits last month)

# Causality

## An illustrative example

- The folks who didn't go to the hospital were fine to begin with
- Suppose they had an average score of 60
- If we run a regression with this data, we are comparing sick people to healthy people in order to try to find out if hospitals make people feel better
- So you see the problem intuitively
- *This* is the embodiment of a situation when $E(u|x) \neq 0$
- The unobservable term $u$ contained a key piece of information that wasn't in our data: how the person was feeling when they went to the hospital in the first place, i.e. whether or not they were sick
- The unobservables were related to the $x$ variable (number of hospital visits last month)
- We can't have this if we want to give the econometric model a causal interpretation

# Causality

## An illustrative example

- Experimental procedure
  - Suppose our population of interest is Washington DC . . .

# Causality

## An illustrative example

- Experimental procedure
  - Suppose our population of interest is Washington DC . . .
  - . . . so we drive around DC in a van with a sliding door, and we snatch people off the street

- Experimental procedure
  - Suppose our population of interest is Washington DC . . .
  - . . . so we drive around DC in a van with a sliding door, and we snatch people off the street
  - Then we take everyone to a warehouse, and we sort them into two categories, *randomly*: those who will go to the hospital, and those who will not

# Causality

## An illustrative example

- Experimental procedure
  - Suppose our population of interest is Washington DC . . .
  - . . . so we drive around DC in a van with a sliding door, and we snatch people off the street
  - Then we take everyone to a warehouse, and we sort them into two categories, *randomly*: those who will go to the hospital, and those who will not
  - Then, after the ”treatment” group has gone to the hospital and the ”control” group has not, we take our measurements
  - We ask all the subjects to rank how good they feel, on a scale of 1 to 5

# Wrapping-up

- There are a few topics covered in Ch 2 of Wooldridge that we will postpone until we cover multiple regression
- Variance of the estimators (standard error of the estimators)
- Homoskedasticity and its evil twin heteroskedasticity
- Regression through the origin
- We have covered everything else in Ch 2 and are ready for multiple regression

# Next time

- Multiple regressors (multiple different $x$'s)
- Chapter 3 material
- Homework: Read 3.1, 3.2, 3.3