Introduction

What is econometrics

Econometrics is “doing economics with data”

Or more formally, a set of statistical tools for analyzing economic relationships (and other relationships, econometrics isn’t just for economists)

In this class, we’re going to focus on one very useful set of tools, known broadly as **linear regression**

By the end of this course, you'll know how to

- Estimate and analyze causal effects and other relationships using real-world data
- Test hypothesis about these relationships
- Predict economic and other variables
- Think about the difference between causal effects and statistical relationships

This introductory lecture will give you a feel for what you can do with econometrics

But first, a bit more about the class

Software

Most of the work we'll do in this class will involve using statistical software to work with actual data

We will do most of our work in R, an open-source program for statistics that works on all operating systems

You will also want to download R studio, a program that makes it easier to write R code

Let me know if you run into any trouble installing these programs

Many employers want to hire people that know how to use R, so learning some of the basics will make you more competitive on the job market

On the other hand, R can also be a little bit of a pain

If you prefer, you can also do (most) of your work in gretl, another free econometrics program that is easier to use

Gretl can be downloaded from https://gretl.sourceforge.net

(I don't care which one you use, I think they're both cool)

Textbooks

I have created a fairly comprehensive set of lecture slides, so there is no required text

If (like me) you are the kind of student that likes to have a text, I have several favorites:

Econometrics texts:

- *Introductory Econometrics: Intuition, Proof and Practice* by Zax. This is a great book if you want to do a deep dive on the theory behind the topics we'll discuss in this course. My lectures are inspired by, and loosely based around, this book

- *Introductory Econometrics* by Wooldridge. This is a great book if you want to go beyond the topics covered in this course, or if you want even more examples on the topics that we will cover
- *Mastering Metrics* by Angrist and Pischke. This is a great book if you want to learn more about some of the applied topics that we will cover at the end of the semester

I didn't put any of these in the bookstore because it's often cheaper to find a used copy online

Software texts

I will show you the basics of statistical computing, but if you want to learn more, here are some nice resources:

- *Introduction to Econometrics with R* by Hanck, Arnold, Gerber and Schmelzer discusses how to use R for common econometric methods, and can be read for free online
- *An Introduction to R* by Venables, Smith, et al. is a classic (and shortish) introduction to basic R, and can be downloaded for free

- *R for Data Science* by Wickham, Cetinkaya-Rundel, and Grolemund is a good introduction to some modern, and advanced, R programming, and is also free to read online
- *Using gretl for Principles of Econometrics* by Adkins is a nice guide that gives more information on using Gretl, and is a free download

An example

Let’s use an example to motivate what econometrics is all about, and why you want to learn it

Along the way, we’ll also learn the basics of how our software works

The question we'll ask is an important, and long-standing, one in labor economics:

> Does having more education increase one's wages?

Why is this an important question?

Isn't it obvious that education and wages are related?

In economics, there are two main theories of education:

1. The human capital investment model: education increases your skills, making you more productive and increasing your wages
2. The Spence signalling model: education does not affect producivity, instead people “jump through the hoops” of getting an education to prove to employers that they are talented

The first theory says that education should affect wages, while the second says that it should not

Let’s investigate this using some data on hourly wages (measured in cents), education, and a bunch of other variables

We’ll load these data, along with some useful “package” that extend R’s capabilities.

```
library(tidyverse)
card <- read_csv("card.csv")
```

Next, let's look at some summary statistics (this is always an important step)

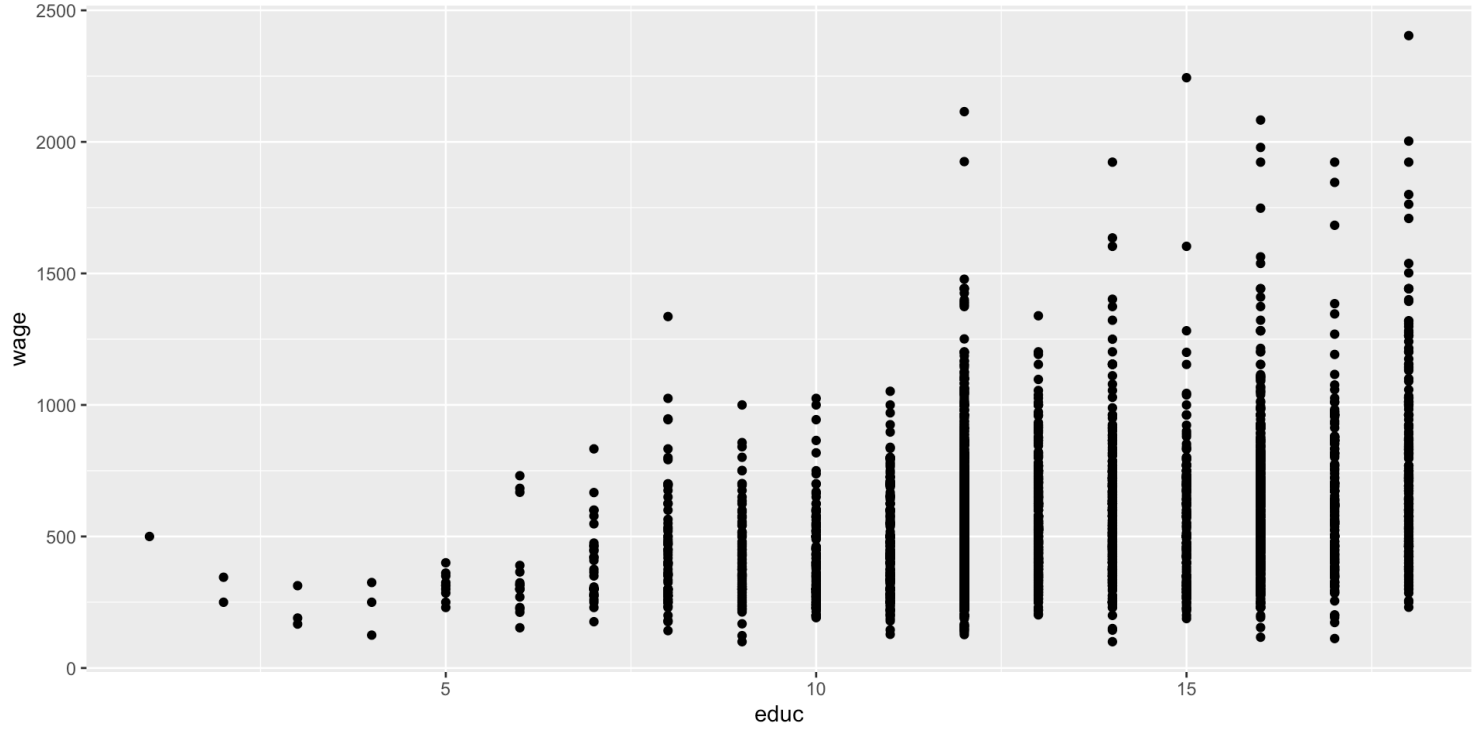
```
1 summary(card)
```

```
       id            nearc2            nearc4             educ      
 Min.   :   2   Min.   :0.0000   Min.   :0.0000   Min.   : 1.00  
 1st Qu.:1276   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:12.00  
 Median :2541   Median :0.0000   Median :1.0000   Median :13.00  
 Mean   :2582   Mean   :0.4409   Mean   :0.6821   Mean   :13.26  
 3rd Qu.:3859   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:16.00  
 Max.   :5225   Max.   :1.0000   Max.   :1.0000   Max.   :18.00  

      age            fatheduc       motheduc         weight       
 Min.   :24.00   Min.   : 0    Min.   : 0.00   Min.   :  75607  
 1st Qu.:25.00   1st Qu.: 8    1st Qu.: 8.00   1st Qu.: 122798  
 Median :28.00   Median :10    Median :12.00   Median : 365200  
 Mean   :28.12   Mean   :10    Mean   :10.35   Mean   : 321185  
 3rd Qu.:31.00   3rd Qu.:12    3rd Qu.:12.00   3rd Qu.: 406024  
 Max.   :34.00   Max.   :18    Max.   :18.00   Max.   :1752340  
                 NA's   :690   NA's   :353                      
```

To get a preliminary sense of how education and wages are related, let's plot them. We can do this using the `ggplot` command, which makes nice plots (but takes some learning):

```
1 ggplot(data=card, aes(x=educ, y=wage)) + geom_point()
```



It does appear that those with more education earn higher wages, but let's formalize this relationship

Specifically, let's use a linear model:

$$wage = \alpha + \beta educ + \varepsilon$$

Here, $wage$ is the **dependent variable** and $educ$ is the called an **independent variable**, **explanatory variable** or **regressor**

α is the intercept and β is the slope coefficient

ε is called the **error term,** and it reflects the fact that wages depend on other things besides education

Regression is a way of using data to *estimate* the parameters α and β in our model

Our estimated regression will look take the form:

$$\widehat{wage} = a + b \cdot educ$$

The "hat" on wage means that this equation is a **prediction** of the wage for someone with a particular amount of education

$$\widehat{wage} = a + b \cdot educ$$

a is our estimate of the true intercept α (we'll say more about what "true" means later)

We can interpret a as the wage that we'd predict for someone with no education

b is our estimate of the true slope coefficient β

We can interpret b as how an extra year of eduation will increase the predicted wage:

$$b = \frac{\Delta \widehat{wage}}{\Delta educ}$$

We can easily estimate the regression using the `lm` (linear model) command:

```r
educ.mod.1 <- lm(wage ~ educ, data=card)
summary(educ.mod.1)
```

```
Call:
lm(formula = wage ~ educ, data = card)

Residuals:
    Min      1Q  Median      3Q     Max 
-576.09 -173.36  -34.12  127.82 1686.25 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  183.949     23.104   7.962 2.38e-15 ***
educ          29.655      1.708  17.368  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 250.7 on 3008 degrees of freedom
```

According to our model,

$$\widehat{educ} = 183.95 + 29.66 \cdot educ$$

The value of $b = 29.66$ means that every additional year of education increases the predicted wage by about $.30

This suggests that wages and education are *related*, but that's not the question

The human capital model says that education has a **causal effect** on the wage

The signalling model says that there is only a "statistical association" because talented people get more education to signal their abilities

One problem is that people with more education have, by definition, spent less time in the workforce

This means that they have less experience, which probably also affects wages

On one hand, education might *increase* wages. On the other hand, more educated people have less experience, which might *decrease* wages

The overall relationship that we see in the data between education in wages reflects a combination of both of these factors

This might cause our regression to *understate* the effect of education on wages

We can address this by *controlling* for experience, or *holding experience constant*

In other words, looking at how wages differ between people with different amounts of education, but the *same* experience

This is an example of the idea of *ceteris parabus*, which means "all else constant"

Now, our regression becomes

$$\widehat{wage} = a + b_1 \cdot educ + b_2 \cdot exper$$

We can interpret b_1 as how the predicted wage changes when education increases by one year, *holding education constant*:

$$b_1 = \left.\frac{\Delta \widehat{wage}}{\Delta educ}\right|_{exper}$$

This is called a **multivariate** regression, and we'll see the details of how it works later

For now, let's see how it affects our estimates

```r
educ.mod.2 <- lm(wage ~ educ + exper, data=card)
summary(educ.mod.2)
```

```
Call:
lm(formula = wage ~ educ + exper, data = card)

Residuals:
    Min      1Q  Median      3Q     Max 
-647.16 -161.20  -29.57  124.79 1612.62 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -375.588     37.802  -9.936   <2e-16 ***
educ          55.055      2.140  25.722   <2e-16 ***
exper         25.142      1.383  18.174   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Controlling for experience *increases* the slope coefficient on education

Since more educated people have less experience (and experience probably increases wages), this is exactly what we expect

When we control for experience by holding it constant, the apparent relationship betwen education and wages increases

Under the signalling theory, we might think that people with more education have higher abilities

To account for this, we want to try to *control* for ability—i.e., see how wages differ between people with different education but the *same* underlying ability

Of course, “ability” is an ambiguous concept, but our data contains variables that might be considered measures of ability

For starters, we can control for individual's IQ test scores, as well as their scores on a test that measures "knowledge of the world of work":

```r
educ.mod.3 <- lm(wage ~ educ + exper + IQ + KWW, data=card)
summary(educ.mod.3)
```

```
Call:
lm(formula = wage ~ educ + exper + IQ + KWW, data = card)

Residuals:
    Min      1Q  Median      3Q     Max 
-618.80 -151.35  -30.79  120.55 1619.31 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -445.4814    58.0504  -7.674 2.57e-14 ***
educ          38.4039     3.4914  11.000  < 2e-16 ***
exper         23.3413     1.9970  11.688  < 2e-16 ***
IQ             1.2912     0.4287   3.012  0.00263 ** 
KWW            5.4820     0.8971   6.111 1.19e-09 ***
---
```

Now the coefficient on education *decreases* from 55 to 38

Again, this is exactly what we expect

People with higher ability (as measured by IQ and KWW) probably have more education, but would earn higher wages regardless of their education

When we hold IQ and KWW constant, the apparent relationship between education and wages shrinks

We might also be concerned that people inherit some of their abilities from their parents

To try to address this, let’s control for parental education and whether individual’s grew up in a household where someone had a library card

```r
educ.mod.4 <- lm(wage ~ educ + exper + IQ + KWW + motheduc + fatheduc + lib
summary(educ.mod.4)
```

```
Call:
lm(formula = wage ~ educ + exper + IQ + KWW + motheduc + fatheduc +
    libcrd14, data = card)

Residuals:
    Min      1Q  Median      3Q     Max
-644.00 -151.36  -30.76  120.65 1588.56

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -485.8863    68.0075  -7.145 1.37e-12 ***
educ          39.7635     4.0080   9.921  < 2e-16 ***
exper         28.0601     2.3735  11.822  < 2e-16 ***
IQ             1.1266     0.5002   2.252  0.02444 *
KWW            3.4802     1.0579   3.290  0.00102 **
```

Now the coefficient shrinks a little more, down to 39.76

This means that—holding experience, IQ, KWW, parental education, and having a library card constant—each year of education increases the predicted hourly wage by about $40

The key point is that, even after controlling for all of these factors, the relationship between education and wages is still positive (this result is also “statistically strong” in a sense that we’ll clarify later)

This suggests that the relationship between education and wages that we have estimated might be **causal** and not just a "statistical association"

We can interpret this as evidence in favor of the human capital theory over the signalling theory (most studies of this phenomenon reach similar conclusions)

Of course, it's always possible that there are other factors that we should be controlling for, but aren't

Later, we'll discuss other techniques that we can use to try to be sure that the relationships that we're estimating are actually causal (although the truth is that you can never be sure)