

ECON42720 Causal Inference and Policy Evaluation

2 Regression Recap

Ben Elsner (UCD)

About this Lecture

Linear regression is by far the most important **estimation technique in causal inference**

Yes, I know, **machine learning is all the rage these days**

- ▶ But a linear approximation is often the best we can do
- ▶ It's already hard enough to get the linear approximation right
- ▶ Fancy techniques are not always better

Resources

The material behind these slides can be found in any good econometrics textbook. For introductory econometrics, I recommend

- ▶ Wooldridge, J. Introductory Econometrics: A Modern Approach. 7th Edition. South-Western College Publishing, 2019.
- ▶ Stock, J. and M. Watson. Introduction to Econometrics. 3rd Edition. Pearson, 2017.

Regression recap

Why do we use linear regression?

- ▶ What we want to approximate: the conditional expectation function (CEF)

How does regression work?

- ▶ Interpretation of coefficients with and without controls
- ▶ Estimation with OLS and Inference

Undergrad Recap: Goal of Linear Regression

Quantify the **expected effect** of a **one unit change in X on Y**

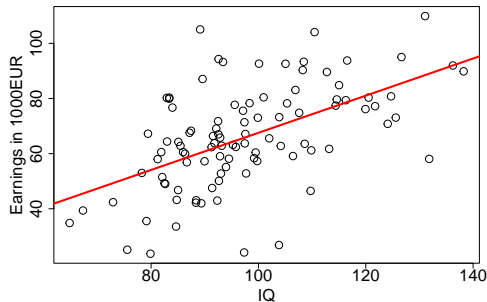
- ▶ If X goes up by one unit, by how many units does Y go up or down?
- ▶ **Causal interpretation:** If we/nature/an experimenter changes X by one unit, what is the expected effect on Y ?

This **effect** is equivalent to the **slope** coefficient β_1 in a **linear regression model**

$$Y = \beta_0 + \beta_1 X + u$$

Linear Regression

Regression analysis means that we **fit a straight line** through (X, Y) data points



In this example, the **regression line** is $\text{Earnings} = -3000 + 700 \text{ IQ}$

- An increase in the IQ by one point increases earnings on average by 700 EUR

What we are looking for: the conditional expectation function

In undergraduate econometrics, you probably learned about a **population regression model**

- ▶ the idea is that there is a **true relationship** between X and Y that we want to estimate
- ▶ we have a **sample** of n observations from this population and estimate $\hat{\beta}_0, \hat{\beta}_1$ using OLS

But is the **population regression model (PRM)** really what we are looking for?

- ▶ Yes and no. The PRM is an approximation of our object of interest
- ▶ This object is called the **conditional expectation function (CEF)**

Ingredients: Random Variables

Random variables are variables that take on different values with a certain probability

- ▶ x is a random variable that takes on values x_1, x_2, \dots, x_n with probabilities $f(x_1), f(x_2), \dots, f(x_n)$

Expected value: the average value of a random variable

$$\begin{aligned} E(x) &= x_1 f(x_1) + x_2 f(x_2) + \dots + x_n f(x_n) \\ &= \sum_{j=1}^n x_j f(x_j) \end{aligned}$$

Expected Value: Example

$x \in \{-1, 0, 2\}$ with probabilities $f(-1) = 0.3$, $f(0) = 0.3$, $f(2) = 0.4$. The expected value of X is

$$\begin{aligned} E(x) &= (-1)(0.3) + (0)(0.3) + (2)(0.4) \\ &= 0.5 \end{aligned}$$

Notation

We denote **random variables with lower case letters** x, y .

Typically, we **do not use indices when we talk about the population**. For example, the linear relationship between x and y in the population is

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

Realisations of a random variable are denoted with **lower case letters with indices** x_i with $i = 1, \dots, n$. We also use indices when

- ▶ Talking about **relationships in the sample**
- ▶ Talking about particular realisations of a random variable: $x_i = x$, for example $x_i = \textit{female}$ or $x_i = 5$

This can be confusing at the start but you'll get used to it!

The Conditional Expectation Function

We are interested in **explaining the relationship between x and y** in the population

A useful concept in this regard is the **Conditional Expectation Function (CEF)**:

$$E(y_i|x_i)$$

- ▶ what is the **population average of y_i for a given value of x_i**
- ▶ i.e. what if x_i takes value x ?

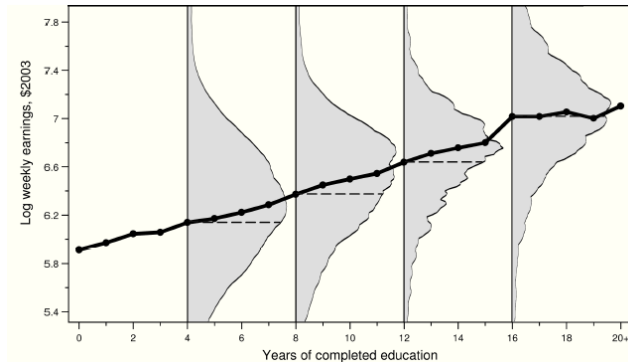
Example: x is a dummy that equals one if a person is female and zero otherwise. y is earnings.

The **CEF can take on two values:**

- ▶ Average earnings of women $E(y_i|x_i = 1)$
- ▶ Average earnings of men/other $E(y_i|x_i = 0)$

A Continuous CEF

Education vs earnings (from Angrist & Pischke, MHE)



At every level of completed education, we have a different expected value of earnings

- ▶ At each value x_i we have a distribution of y_i
- ▶ and the CEF comprises the averages of this distribution

Regression is a Linear Approximation of the CEF

The **population regression model** (PRM) is a linear approximation of the CEF

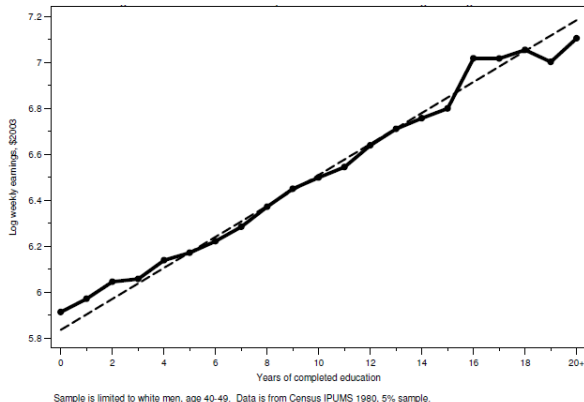
$$y = \beta_0 + \beta_1 x + u$$

- ▶ Our **ultimate goal** is to know the **CEF**
- ▶ But: with a linear regression, we can estimate the parameters of the PRM
- ▶ I.e. we **cannot estimate the CEF directly**

Why a **linear approximation is useful**:

- ▶ We typically have small sample sizes, so approximating a non-linear function is difficult
- ▶ We are often interested in marginal effects, so a linear approximation is often sufficient

PRM: Linear Approximation of the CEF



The dashed line is the Population regression model $y = \beta_0 + \beta_1 x + u$. The solid line is the CEF $E(y|x)$.

It can be shown that the PRM is the **best linear approximation of the CEF** (see Angrist & Pischke, MHE, ch. 3).

CEF and PRM: what's all this about?

The **CEF is the object of interest** in (most of) econometrics

- ▶ The PRM $y = \beta_0 + \beta_1 x + u$ is a **linear approximation** of the CEF.
- ▶ But it is an approximation, so it can be wrong.
- ▶ With data, we can draw inference on the PRM but not on the CEF

What does this mean for the empirical analysis?

- ▶ We need to think about the **relationship between x and y** in the population
- ▶ Linear approximations are more innocuous when we consider small changes in x

The Sample Regression Function

Now suppose we have a **sample of size n** that was **randomly sampled from the population**, $(y_1, x_1), \dots, (y_n, x_n)$

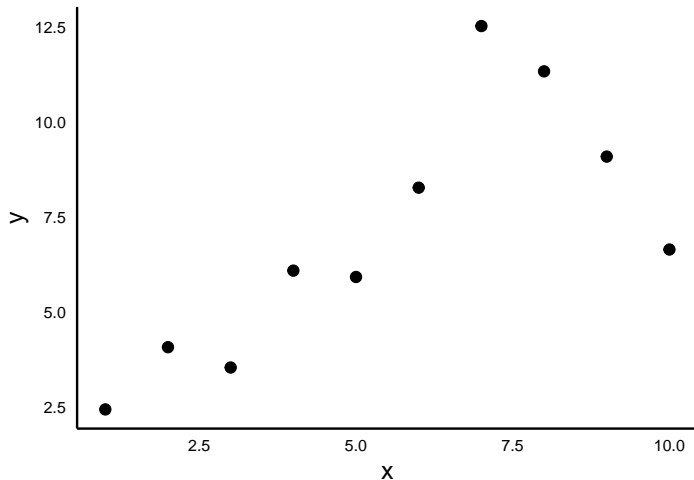
The **sample regression function** is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (1)$$

We can estimate the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ with Ordinary Least Squares (OLS)

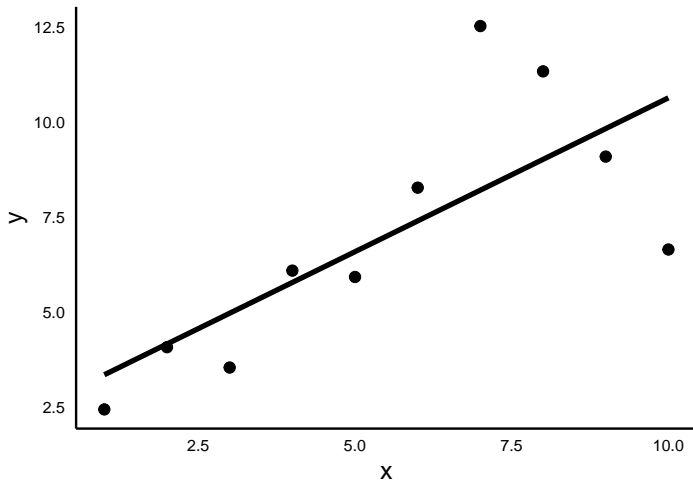
OLS: Intuition

Let's start with some data points



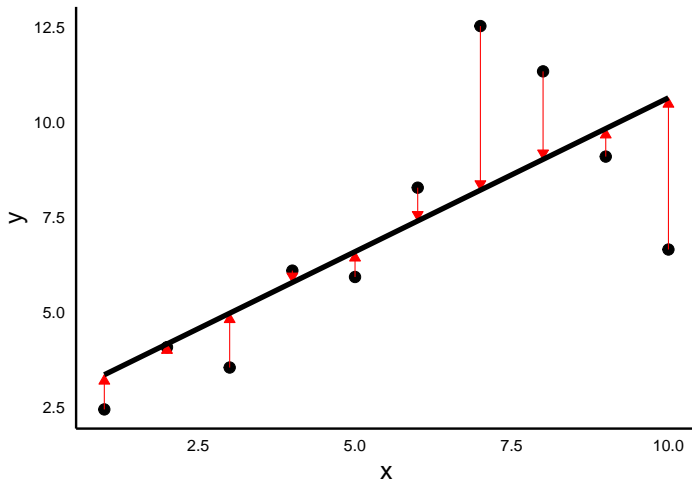
OLS: Intuition

Goal: fit a regression line through those points



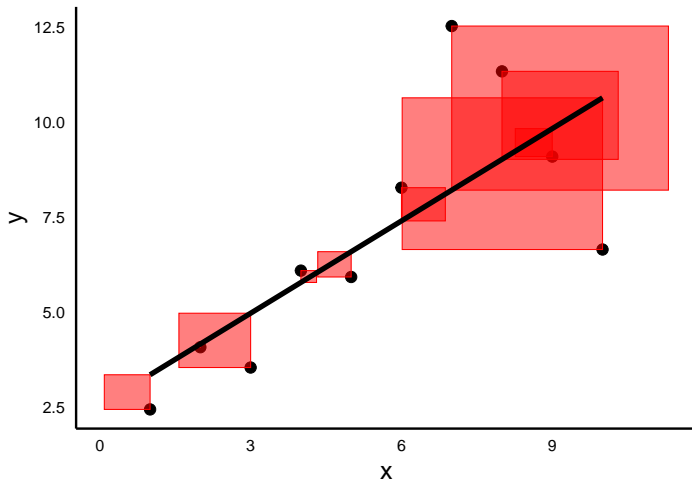
OLS: Intuition

The key ingredient of OLS are the residuals $\hat{u}_i = y_i - \hat{b}_0 - \hat{b}_1 x_i$



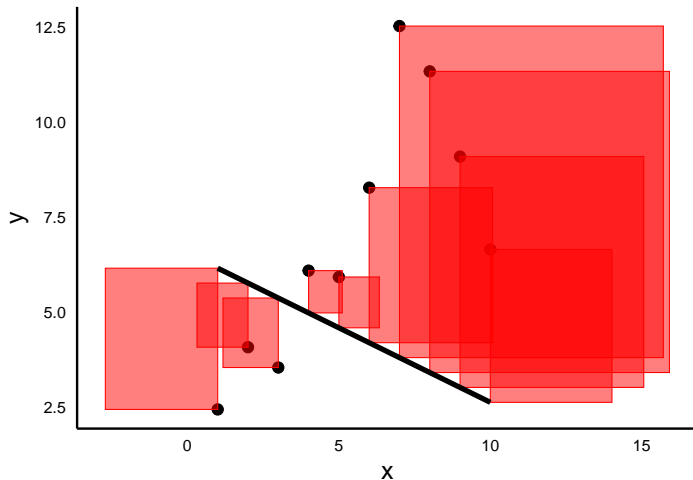
OLS: Intuition

Now consider the square of each residual



OLS: Intuition

Let's consider a different regression line: the squares are much larger!



OLS: Intuition

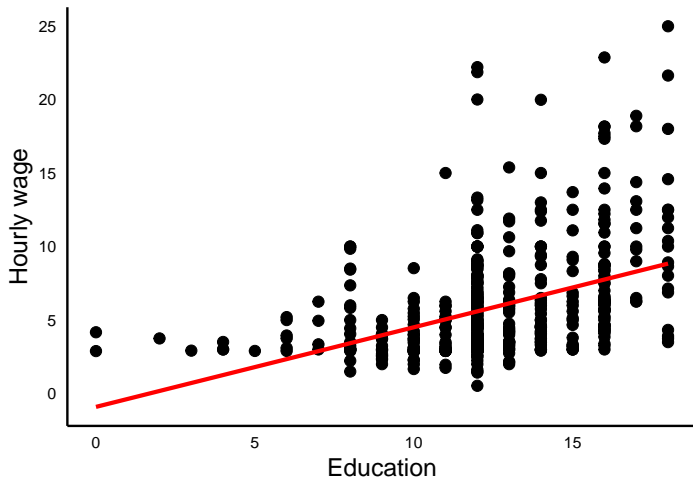
OLS **minimizes the average size of these squares**

It **minimizes the sum of squared residuals (SSR)**

The result is the **best-fitting line** that describes the relationship between x and y in the sample

OLS: Data Example

Let's look at the relationship between education and wages with data from the U.S.



Regression Output: Interpretation

Table 1: Effect of Education on Wages

<i>Dependent variable:</i>	
	wage
educ	0.541*** (0.053)
Constant	−0.905 (0.685)
Observations	526
R ²	0.165
Adjusted R ²	0.163
Residual Std. Error	3.378 (df = 524)
F Statistic	103.363*** (df = 1; 524)
Note:	* p<0.1; ** p<0.05; *** p<0.01

A 1-year increase in education is associated with a 0.54 USD increase in hourly wages

OLS: The Math

The Ordinary Least Squares (OLS) estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are derived through the minimization problem

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg \min_{\widehat{b}_0, \widehat{b}_1} \sum_{i=1}^n [(y_i - \widehat{b}_0 - \widehat{b}_1 x_i)^2] \quad (2)$$

The sample means $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ are the sample analogs of the population means $E(y_i)$ and $E(x_i)$

OLS: The Math

The **residuals of the regression** are defined as $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$.

When we “run an OLS regression”, we **minimize the sum of squared residuals (SSR)**, $\sum_{i=1}^n \hat{u}_i^2$ and obtain values for $\hat{\beta}_0$ and $\hat{\beta}_1$

Solving the minimization problem (2) yields the **estimators**

$$\begin{aligned}\hat{\beta}_1 &= \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\widehat{Cov}(y_i, x_i)}{\widehat{V}(x_i)} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}\tag{3}$$

The Sampling Distribution of the OLS Estimator

To **draw inference about the population**, we need to know the **sampling distribution of the OLS estimator**

What we want to know:

- ▶ When will $\hat{\beta}_1$ be **unbiased**?
- ▶ What is its **variance**?

To answer these questions, we need to make some **assumptions about the sample and population**

1. Population model is linear in parameters
2. Sample is randomly drawn from the population
3. Variation in x
4. **Zero conditional mean assumption (ZCM)**

OLS Assumptions

The four **OLS assumptions must be fulfilled** for the OLS estimator to be **unbiased and consistent**

Unbiasedness: $E(\hat{\beta}_1) = \beta_1$

- ▶ across many random samples, the estimator gets it right on average

Consistency: $\hat{\beta}_1 \xrightarrow{P} \beta_1$ as $n \rightarrow \infty$

- ▶ if the sample size increases, the estimator converges to the true value
- ▶ this is a consequence of the Law of Large Numbers (LLN)
- ▶ As n gets larger, the sample becomes more representative of the population

The Zero Conditional Mean (ZCM) Assumption

The **ZCM assumption is the most important assumption** in this module

- ▶ It is **not testable with the data** at hand
- ▶ It rarely holds in practice (except in randomised experiments)
- ▶ Causal inference techniques exploit scenarios where ZCM holds approximately

Other names for the ZCM assumption:

- ▶ **Conditional independence assumption (CIA)**
- ▶ **Exogeneity assumption**

The Zero Conditional Mean (ZCM) Assumption

Consider the **population model**

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

The **ZCM assumption** states that the **conditional mean of the error term is zero**

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

What does this mean?

- ▶ The error term is **not systematically related** (speak: uncorrelated) with x
- ▶ At any level of x , the average value of u is zero

ZCM Assumption: Example

Does **higher education (causally) increase earnings?**

$$wage_i = \beta_0 + \beta_1 education_i + u_i$$

What is the error term u_i here?

- ▶ Any determinant of a person's wage that is **not education**
- ▶ E.g., innate ability, motivation, personality, etc.

ZCM Assumption: Example

Say u_i includes ability. According to the ZCM assumption, the following must hold:

$$E(\text{ability} \mid \text{education} = 8) = E(\text{ability} \mid \text{education} = 12) = E(\text{ability} \mid \text{education} = 16)$$

So it **must hold that**:

- ▶ the average ability of people with 8 years of education is the same as
- ▶ the average ability of people with 12 years of education and
- ▶ the average ability of people with 16 years of education

This is hardly plausible \Rightarrow **ZCM assumption is violated**

What if ZCM is Violated?

The OLS estimator of β_1 is **biased and inconsistent**

Bias: $E(\widehat{\beta}_1) \neq \beta_1$

- ▶ The expected value of the OLS estimator is not equal to the true value of β_1
- ▶ Across many samples, the estimates are systematically too big or too small

Inconsistency: $\widehat{\beta}_1$ does not converge to β_1 as $n \rightarrow \infty$

- ▶ Even if the sample size is very large, the OLS estimator does not converge to the true value of β_1

How to think about Violations of ZCM

The variable x is **typically a choice**

- ▶ People choose how much education to get, how often they go to the gym, how much they save, who they want to date, etc
- ▶ Firms choose how much to invest, how many workers to hire, how much to pollute, etc.
- ▶ Governments choose how much to spend on education, how much to tax, etc.

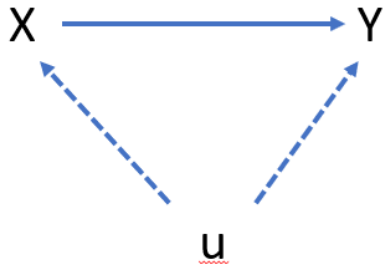
The **choice** of x is typically **influenced by other factors** u

- ▶ Individual factors: ability, motivation, preferences, etc.
- ▶ Firm factors: technology, market conditions, etc.
- ▶ Government factors: ideology, political pressure, etc.

Problem: u **affects** y not just through x but also **through other paths** or directly

How to think about Violations of ZCM

The error term u includes **one or more confounders**



Here u includes a confounder y directly and through x

Example: x is education, y is earnings, u is ability

Omitted Variable Bias

Suppose the **true model** is $y = \beta_0 + \beta_1 x + \beta_2 s_1 + e$

However, we **estimate the model** $y = \tilde{\beta}_0 + \tilde{\beta}_1 x + u$

It can be shown that the **OLS estimator is biased**

$$\tilde{\beta}_1 = \beta_1 + \underbrace{\beta_2 \frac{\text{Cov}(x, s_1)}{\text{Var}(x)}}_{OVB}$$

So when does ZCM hold?

ZCM holds if x is **as good as randomly assigned** to individuals

- ▶ This is the case if x is assigned in a **randomised experiment**
- ▶ Or if x is assigned in a **quasi-experiment** that mimics random assignment
- ▶ Or if we can **control for all confounders** in the analysis

We should **always assume that ZCM is violated**. Researchers need to **think hard about confounders and how to eliminate them**.

Controlling for Confounders: Multivariate Regression

We can include **confounders in the regression model** to control for them

$$y = \beta_0 + \beta_1 x + \mathbf{S}\boldsymbol{\gamma} + u$$

Here, \mathbf{S} is a vector of covariates $\mathbf{S} = (s'_1, s'_2, \dots, s'_k)$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_k)$ is a vector of coefficients,¹ i.e.

$$\mathbf{S}\boldsymbol{\gamma} = \gamma_1 s_1 + \gamma_2 s_2 + \dots + \gamma_k s_k$$

We are **only interested in β_1 , the causal effect of x on y**

- ▶ The other coefficients $\gamma_1, \gamma_2, \dots, \gamma_k$ are not of interest (nuisance parameters)
- ▶ We include the covariates \mathbf{S} to control for confounders

¹Note: each element of \mathbf{S} is in itself an $(1 \times n)$ vector. E.g. $s_1 = (s_{11}, s_{12}, \dots, s_{1n})$ so \mathbf{S} is actually an $(n \times k)$ matrix

Interpretation of β_1 in Multivariate Regression

β_1 now has a **ceteris paribus interpretation**

- ▶ **Holding all other variables \mathbf{S} constant**, a one unit increase in x leads to a β_1 unit increase in y

The **inclusion of \mathbf{S}** allows for a **like-with-like comparison**

- ▶ We **compare units with the same values** of \mathbf{S} but different values of x
- ▶ But the like-with like comparison is only valid if \mathbf{S} contains all confounders

Conditional Mean Independence Assumption

The **Conditional Mean Independence Assumption (CMIA)** is a “light” version of the ZCM assumption.

$$E(u \mid x, \mathbf{S}) = E(u \mid \mathbf{S}) = 0$$

In plain English: as long as **\mathbf{S} is included, the error term u is uncorrelated with x**

- ▶ x is exogenous conditional on **\mathbf{S}**
- ▶ x is as good as random conditional on **\mathbf{S}**

Controlled Regression in Practice

Summary: What you need to understand for this module

Logic of linear regression

- ▶ Why we use linear regression
- ▶ Why and how we use OLS to estimate the parameters of the linear regression model
- ▶ How to interpret the OLS estimator $\hat{\beta}_1$

Limitations of linear regression for causal inference

- ▶ The ZCM assumption is violated in most applications, leading to OVB
- ▶ How control variables can be used to control for confounders

Appendix

Regression with R

```
# Required packages (install if necessary)  
library(tidyverse)  
library(wooldridge)  
library(stargazer)
```

Regression with R

This code shows how to estimate and present regressions with R

```
data('wage1') # load the data
df <- wage1
reg1 <- lm(wage ~ educ, data = df) # estimate simple regression
reg2 <- lm(wage ~ educ + exper, data = df) # estimate multivariate regression
stargazer(reg1, reg2, type = "text") # print regression results
```

Regression outputs with Stargazer and R Markdown

You can generate nice regression tables with the `stargazer` package and R Markdown/Quarto. Here is a code chunk that gives you a nicely formatted latex table.

```
```{r, results='asis', echo=FALSE}  
stargazer(reg1, reg2,
 header=FALSE,
 type='latex',
 title="Effect of Education on Wages")
```
```

Regression with R

We can also generate a scatter with a regression line

```
ggplot(df, aes(x = educ, y = wage)) + # generate scatter plot
  geom_point() + # add points
  geom_smooth(method = "lm", se = FALSE) + # add regression line
  theme_minimal()
```

Group Work Questions I

What is the difference between unbiasedness and consistency of an estimator?

What is the sampling distribution of an estimator? Provide an intuitive explanation

How is the sampling variance of an estimator related to the sample size?

Group Work Questions II

Suppose you have a variable z that is omitted from the population model

$$y = \beta_0 + \beta_1 x + \beta_2 z + u$$

The correct population parameter, β_1 , is positive.

- ▶ a) Suppose you know that $\text{cov}(x, z) > 0$ and $\beta_2 < 0$. How will the OLS estimator $\hat{\beta}_1$ be biased when z is omitted? Will it be over- or underestimated?
- ▶ b) Draw a DAG with z as a confounder and indicate the two factors that make up the omitted variable bias

Group Work Questions III

On the following slide, you see a regression output with two columns. In data on school districts in California, we want to study whether larger class sizes lead to lower test scores. The dependent variable is the average test score in a school district; the regressor of interest is the average class size in a school district. In Column (2), we control for the percentage of students in a school district who are not English native speakers (a proxy for poverty).

- ▶ a) Interpret the slope coefficient on class size in Column (1)
- ▶ b) Given the results in Column (2), what can we say about the correlation between class size and the share of non-native English speakers in a school district?

Group Work Questions III cont'd

Table 2:

| | <i>Dependent variable:</i> | |
|--|----------------------------|-----------------------|
| | testscr | |
| | (1) | (2) |
| str | -2.280***
(0.480) | -1.101***
(0.380) |
| el_pct | | -0.650***
(0.039) |
| Constant | 698.933***
(9.467) | 686.032***
(7.411) |
| Observations | 420 | 420 |
| R ² | 0.051 | 0.426 |
| Adjusted R ² | 0.049 | 0.424 |
| <i>Note:</i> *p<0.1; **p<0.05; ***p<0.01 | | |



benjamin.elsner@ucd.ie



www.benjaminelsner.com



Sign up for office hours



YouTube Channel



@ben_elsner



LinkedIn

Contact

Prof. Benjamin Elsner

University College Dublin

School of Economics

Newman Building, Office G206

benjamin.elsner@ucd.ie

Office hours: book on Calendly