Econometrics I

Lecture 1: Introduction

Matías Cabello
Chair of Economic Growth and Development
MLU Halle-Wittenberg

14 October 2025

Where and when

- Tuesday: 10:15 - 11:45, weekly, Hörsaal A [Mel]
- Thursday: 14:15 - 15:45, weekly, Hörsaal XX [Mel]
- Thursday: 16:15 - 17:45, weekly, Hörsaal B [Mel]
- Monday: 16:15 - 17:45, weekly, Großer Hörsaal [WiWi]

No lectures (I must attend a conference):

- Monday October 20th
- Tuesday October 21st

There is no sharp distinction between lectures and tutorials: we will have a mixture of both, spending roughly half of the time on tutorial-like exercises.

Contact & interaction

- **Via email** (or StudIP): matias.cabello@wiwi.uni-halle.de

- **Personal meetings possible**:
  - By appointment
  - Universitätsring 3, Chair of Growth and Development (Prof. Grieben)

- However: I will offer 30-45 minutes **Q&A every Monday and Thursday** (during the “tutorial” meetings)

Evaluation

- Written examination on

  December 1, 2025 (1st attempt)

  or

  February 23, 2006 (2nd attempt)

- Important:
  - Register for module & exam (two registrations!)
  - The exact dates and locations can be changed by the Examinations Office.

Literature

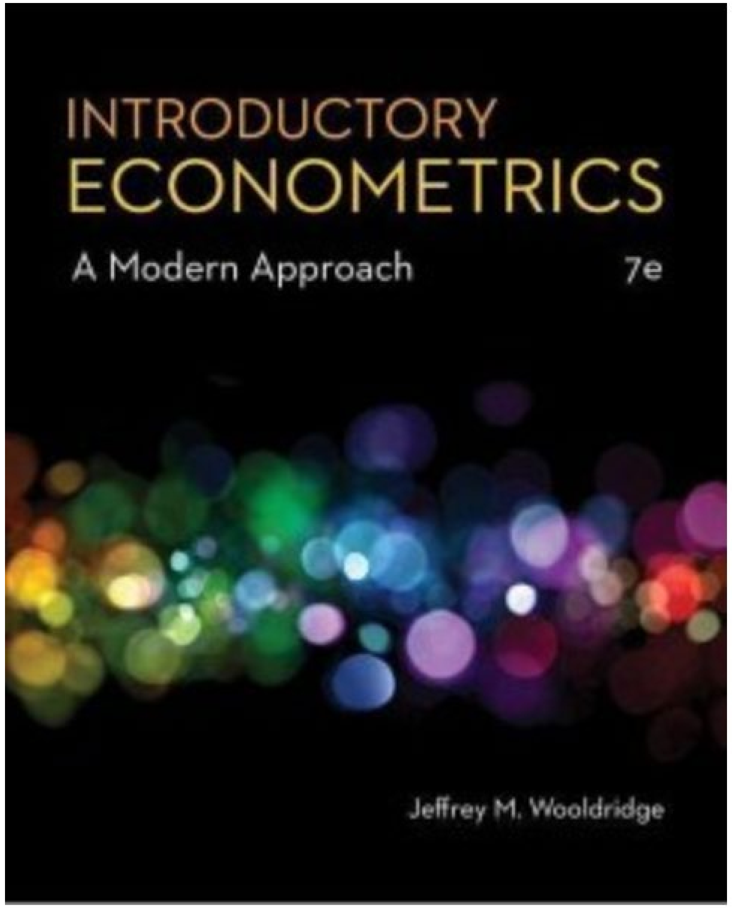
Main textbook

Introductory Econometrics: A Modern Approach,

by Jeffrey Wooldridge,

7th ed. (or any other), Cengage.

Available in university library.

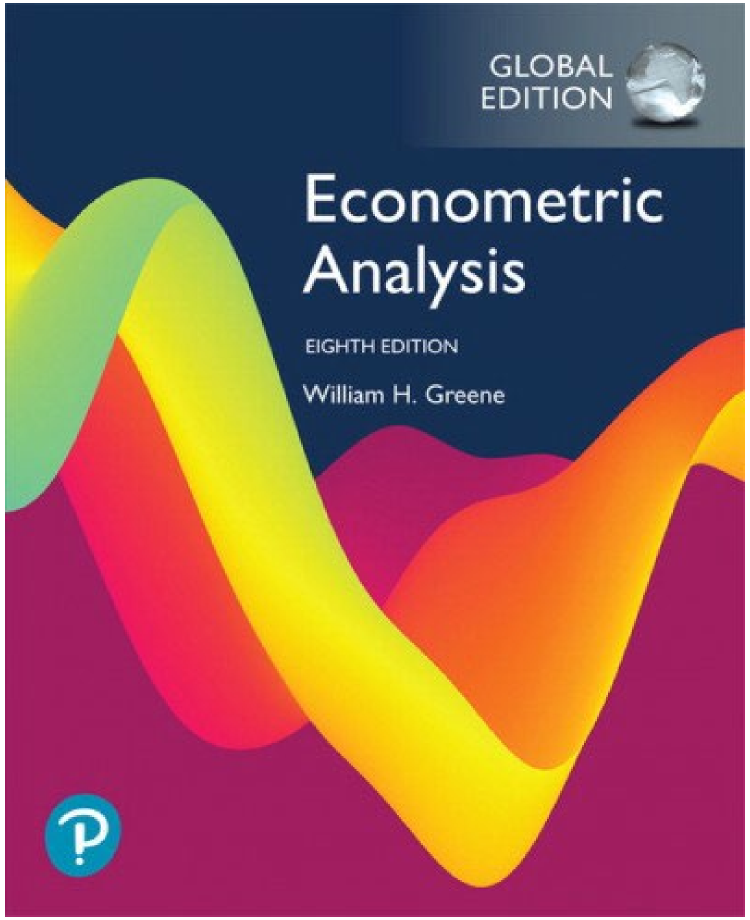


Literature

Secondary textbook

Econometric Analysis,
by William Greene,
8th ed. (or any other), Pearson.

Useful for matrix notation.
Available in university library.



Literature

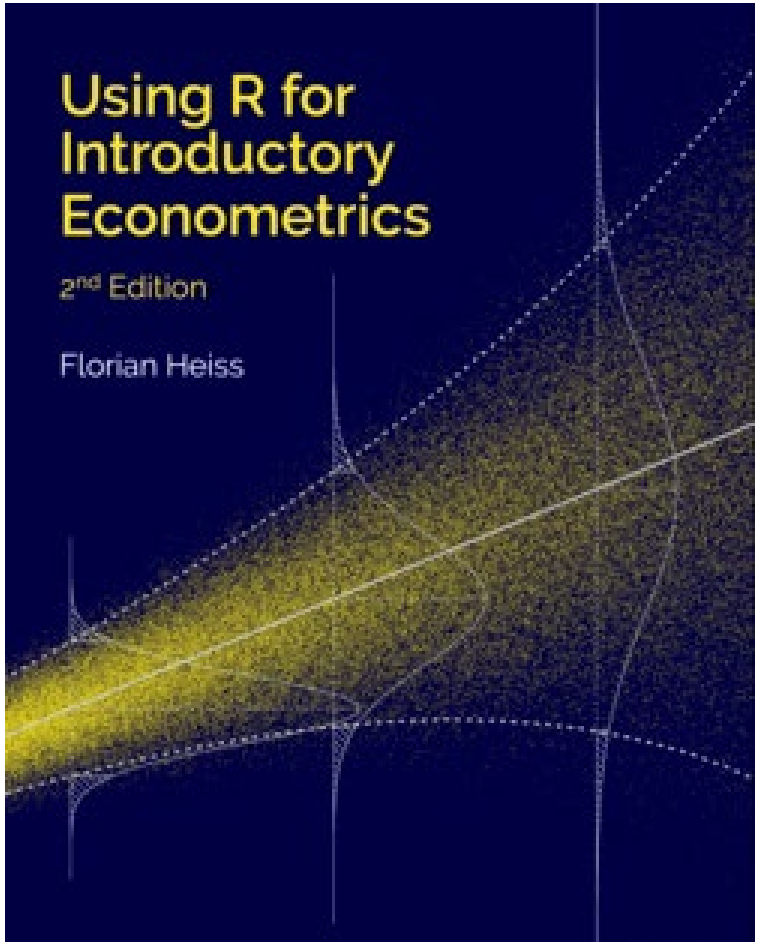
Applied textbook

Using *R* for Introductory Econometrics, 2nd ed.,

by Florian Heiss.

Can be downloaded as PDF at
https://www.urfie.net/

(Whoever is interested can learn Python and/or Julia at the same time, check out the books available at: https://www.urfie.net/)



Software: *R*

We will use the free econometric software *R* (http://www.r-project.org/)

and the user interface *Rstudio* (https://rstudio.com/products/rstudio/download/).

Please install them in your personal computer.

What is Econometrics?

What is Econometrics?

- Econometrics is <u>not</u> "economic measurement."

- ***Econometrics*** ≈ **data**-based analysis
  - Data → Graphs, tables, statistical relationships
  - "Making the data speak"
  - Not necessarily "economic" data (e.g., health, psychology, crime, etc.)

- ***Econometrics*** ≈ statistical **prediction**
  - If variable x changes, how will y <u>react</u>?
  - E.g.: How should sales change if I raise the price? Who will likely get elected if unemployment falls?

What is Econometrics?

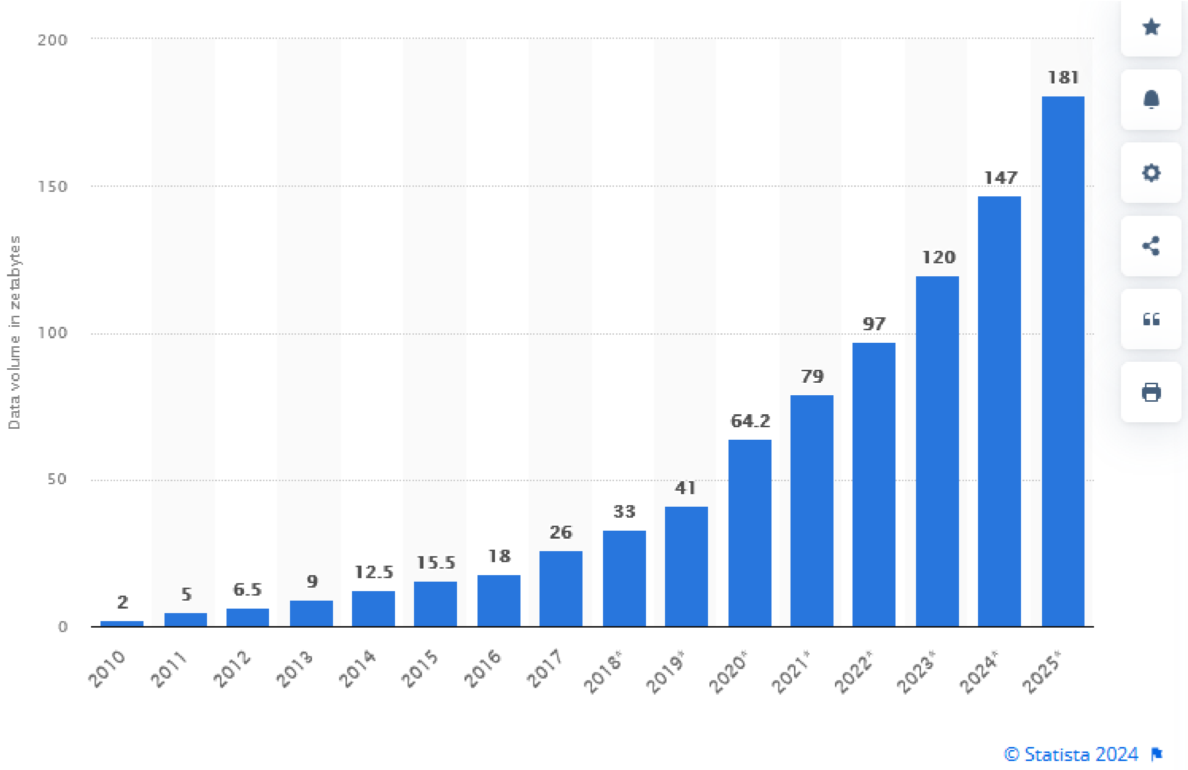
- ***Econometrics*** ≈ **causal** assesment
  - If variable x changes, will y *really* react? By how much?
  - Does better education really lead to higher wages?
  - Does minimal wages really create unemployment?
  - Did measures against Covid really reduce mortality?

- ***Econometrics*** ≈ **regression** analysis

outcome $y = f($**inputs** $x_1, x_2, \ldots)$

- ***Econometrics*** ≈ the most important subject of your studies?

Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

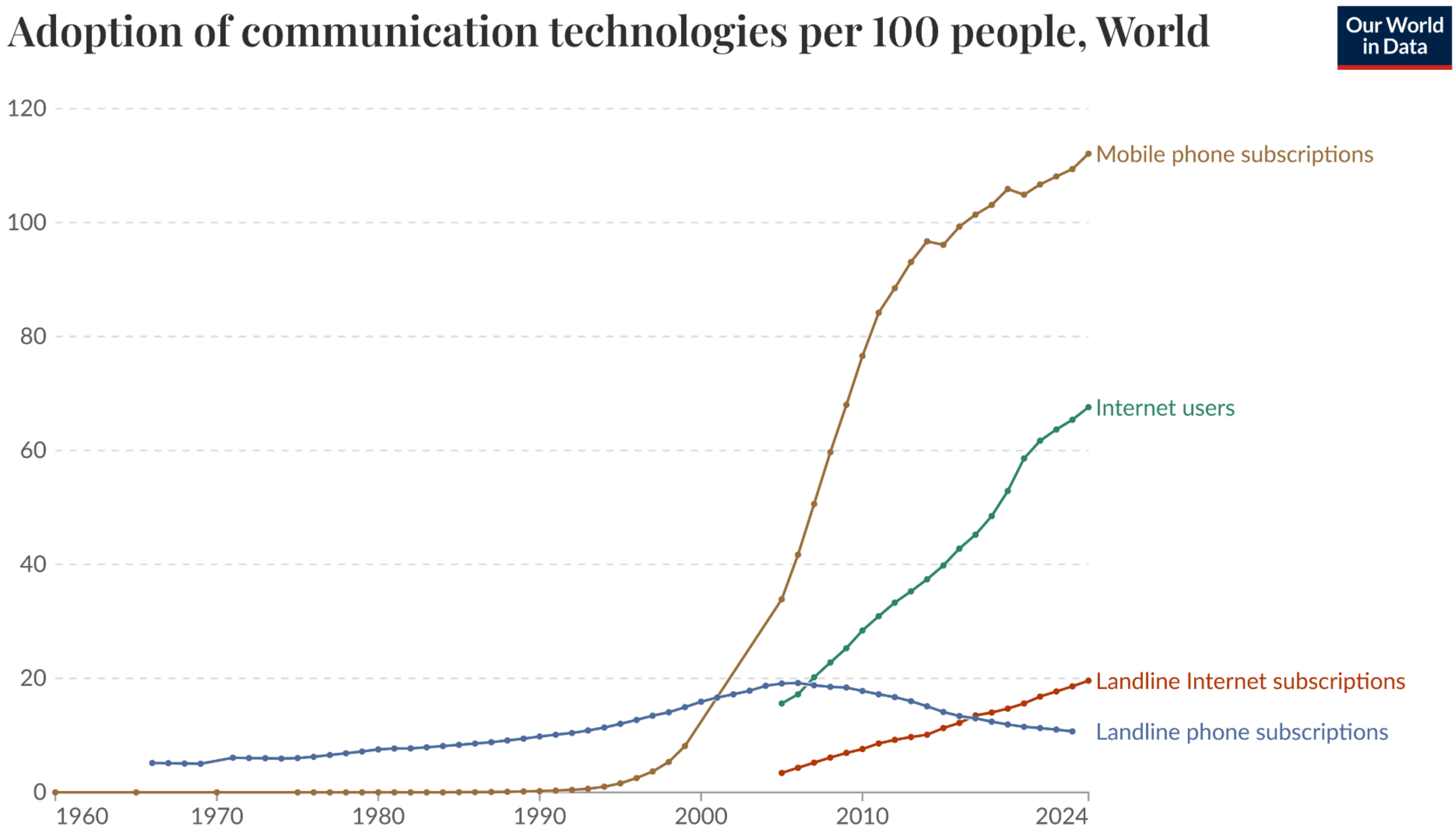


181 Zettabytes (1 followed by 21 zeros) amount of data created, captured, copied, and consumed in 2025 (3x more than 2020).

→ **Data analysis skills increasingly important!**

* This slide is based on one by Amelie Wuppermann

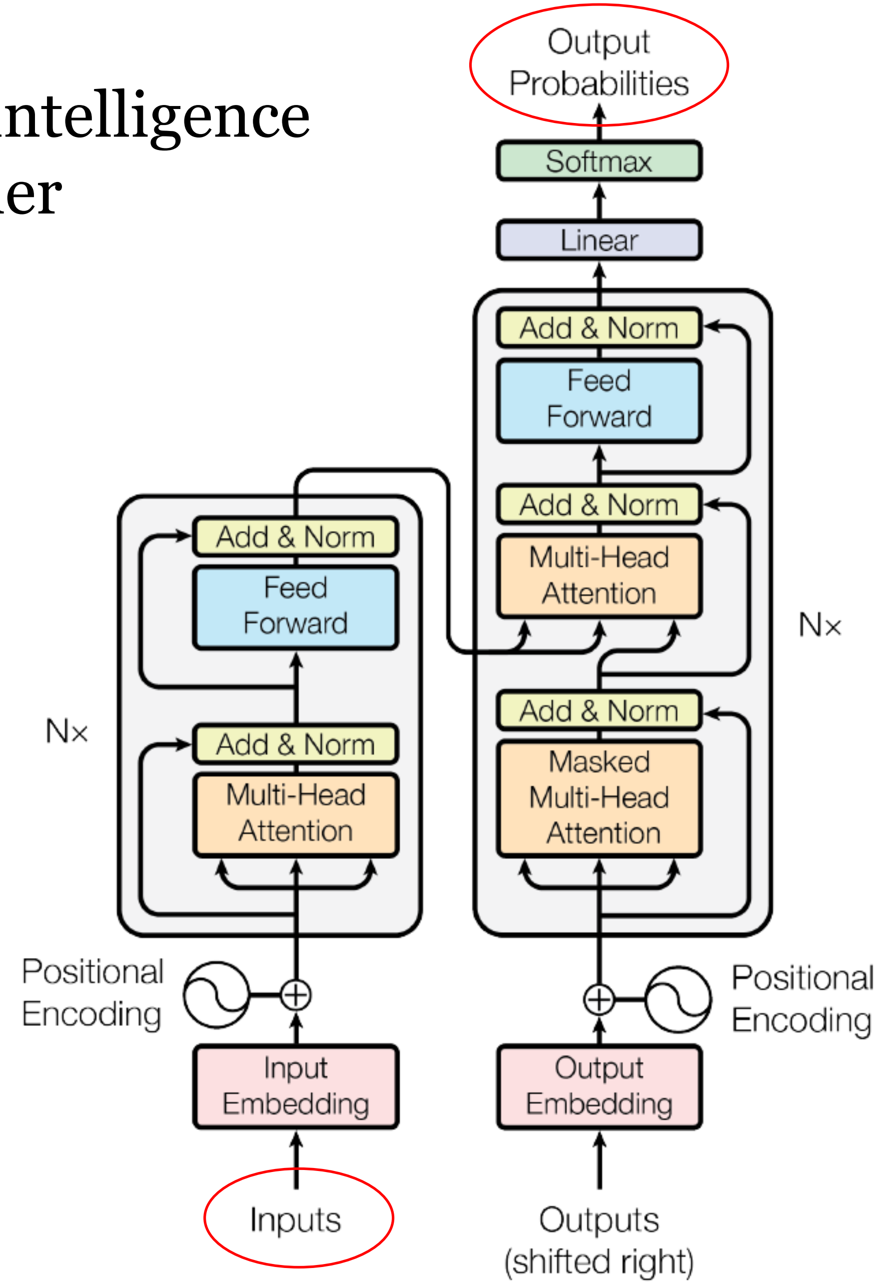
Adoption of communication technologies per 100 people, World



Data source: International Telecommunication Union (ITU), via World Bank (2025); World Telecommunication/ICT Indicators Database - International Telecommunication Union (ITU), via World Bank (2025)

Note: Landline Internet subscriptions are defined as a fixed access to the public Internet with a download speed of at least 256 kbit/s. Internet users are people who have accessed the Internet from any location in the last three months.

Artificial intelligence transformer



At its core, generative AI (ChatGPT, etc.) is a statistical prediction technology that shares many commonalities with a standard regression.

Source: Vaswani, et al. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30.

About this course

Course objectives

You will learn:

1. The logic of **regression analysis**
   - How does it work
   - When does it fail
2. Basic applications (with hands-on computer work)

This background will not only prepare you for Econometrics II (2nd semester half), but **help you understand** much better **the world we live in today**.

Core contents

1. **The basics** of multivariate OLS regressions: Coefficients, standard errors, significance, predictions, properties, asymptotic behavior.
2. **Regression design** (what variables should be included in a regression?). Omitted variable biases, efficiency-bias tradeoff.
3. **Departures from the classical model**: Heteroscedasticity, endogeneity, common misspecifications.

Additional contents

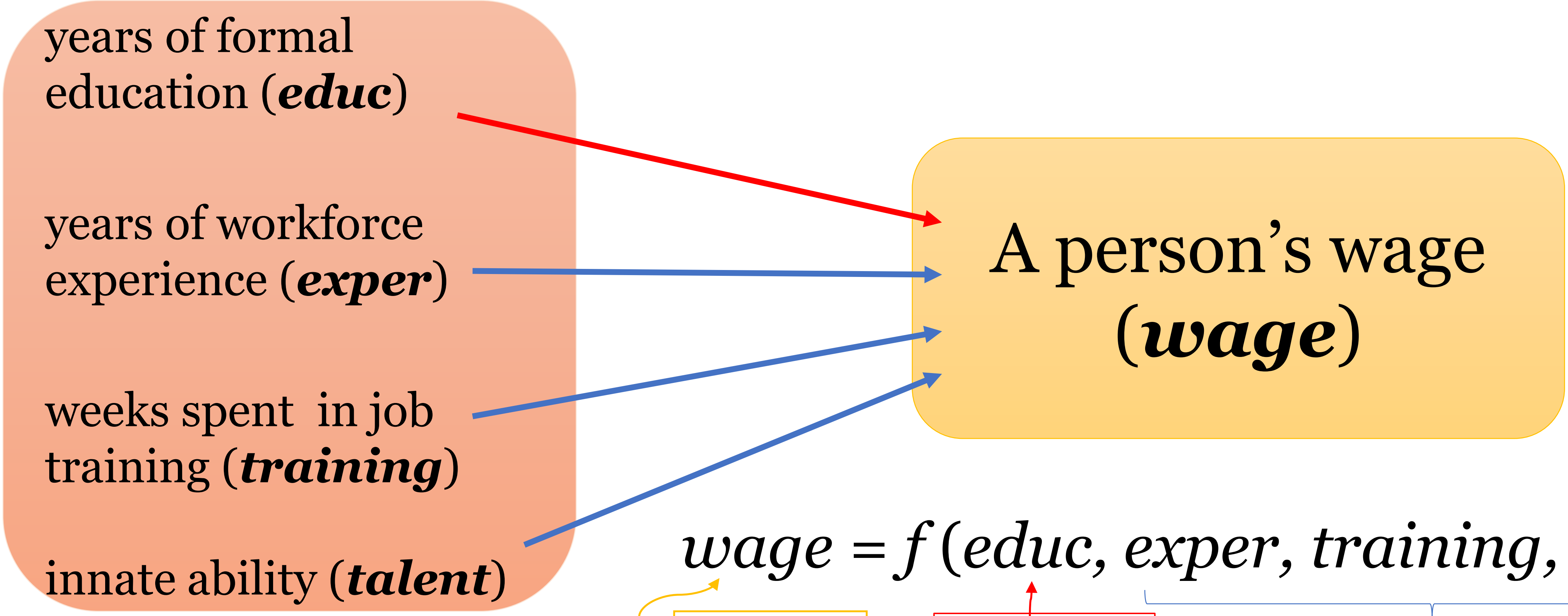
4. **Basic techniques**: Dummies, fixed effects, limited dependent variables, nonlinearities, perhaps: time series and panel data (Econometrics II content)
5. **Distinguishing correlation from** causality: Experiments and quasi-experiments, difference in difference, instrumental variables).
6. **Hands-on analysis**: handling databases, creating graphs, running regressions, and producing result tables.

What is a regression?

Wooldridge, Ch. 1

Regression example 1

Q: Does more education improve wages?



$$\underbrace{wage}_{\text{Outcome}} = f(\underbrace{educ}_{\text{Treatment}}, \underbrace{exper, training, talent}_{\text{Controls}})$$

Regression example 2

Q: Do harder laws reduce crime?

$x1$ = "wage" for an hour spent in criminal activity,

$x2$ = hourly wage in legal employment,

$x3$ = income other than from crime or employment,

$x4$ = probability of getting caught,

$x5$ = probability of being convicted if caught,

x_6 = expected sentence if convicted

A person's criminal activity ($\boldsymbol{y}$)

$$\underbrace{y}_{\text{outcome}} = f(\underbrace{x_1, x_2, x_3, \dots, x_6}_{\text{regressors}})$$

(incl. treatment x_6)

Regression example 3 (Women and the Plough)

Abstract: We test the hypothesis that traditional agricultural practices influenced the historical gender division of labor and the evolution of gender norms. We find that, consistent with existing hypotheses, the descendants of societies that traditionally practiced plough agriculture today have less equal gender norms, measured using reported gender-role attitudes and female participation in the workplace, politics, and entrepreneurial activities.

Alesina, Giuliano & Nunn. 2013. "**On the Origins of Gender Roles: Women and the Plough**." *QJE* 128 (2): 469–530. https://doi.org/10.1093/qje/qjt005.

Regression example 3 (Women and the Plough)

Contemporary controls
(income p.c., conflicts, communism, etc.)

Historical controls
(historical agriculture, family structure, etc.)

Historical plough use

Q: Are today's gender norms determined by historical plough agriculture?

A country's gender norms (e.g. female participation)

If *f*( ) is linear:

$$\underbrace{y}_{\text{Gender norm}} = \beta_0 + \underbrace{\color{red}{\beta_1}\color{black}{x_1}}_{\text{Plough history}} + \underbrace{\beta_2 x_2}_{\text{Income}} + \underbrace{\beta_3 x_3}_{\text{Conflicts}} + \underbrace{\beta_4 x_4}_{\text{Communism}} + \ldots + \underbrace{\epsilon}_{\text{Everything else}}$$

$\beta_1 < 0$: female participation negatively affected by historical plough use

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent variable: Female labor force participation in 2000							
Mean of dep. var.	51.35	51.55	51.35	51.48	51.26	52.09	51.48	52.13
Traditional plough use	-10.892***	-12.714***	-12.356***	-12.336***	-12.721***	-14.618***	-9.913***	-9.234**
	(3.848)	(3.255)	(2.993)	(3.019)	(3.364)	(3.482)	(3.160)	(4.301)
Historical controls:								
Practices intensive agriculture	yes							yes
Prop. of subsist. from herding	yes							yes
Prop. of subsist. from hunting	yes							yes
Absence of private property		yes						yes
Patrilocal marriages		yes						yes
Matrilocal marriages		yes						yes
Nuclear family structure		yes						yes
Extended family structure		yes						yes
Year ethnicity sampled			yes					yes
Contemporary controls:								
Years of civil conflicts (1816-2007)				yes				yes
Years of interstate conflicts (1816-2007)				yes				yes
Ruggedness				yes				yes
Communism indicator					yes			yes
Fraction of European descent					yes			yes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent variable: Female labor force participation in 2000							
Mean of dep. var.	51.35	51.55	51.35	51.48	51.26	52.09	51.48	52.13
Traditional plough use	-10.892***	-12.714***	-12.356***	-12.336***	-12.721***	-14.618***	-9.913***	-9.234**
	(3.848)	(3.255)	(2.993)	(3.019)	(3.364)	(3.482)	(3.160)	(4.301)
Historical controls:								
Practices intensive agriculture	yes							yes
Prop. of subsist. from herding	yes							yes
Prop. of subsist. from hunting	yes							yes
Absence of private property		yes						yes
Patrilocal marriages		yes						yes
Matrilocal marriages		yes						yes
Nuclear family structure		yes						yes
Extended family structure		yes						yes
Year ethnicity sampled			yes					yes
Contemporary controls:								
Years of civil conflicts (1816-2007)				yes				yes
Years of interstate conflicts (1816-2007)				yes				yes
Ruggedness				yes				yes
Communism indicator					yes			yes
Fraction of European descent					yes			yes
Oil production per capita						yes		yes
Agricultural share of GDP						yes		yes
Manufacturing share of GDP						yes		yes
Services share of GDP						yes		yes
Fraction of pop. Catholic							yes	yes
Fraction of pop. Protestant							yes	yes
Fraction of pop. Christian (other)							yes	yes
Fraction of pop. Muslim							yes	yes
Fraction of pop. Hindu							yes	yes
Baseline controls	yes	yes	yes	yes	yes	yes	yes	yes
Observations	165	163	165	163	153	154	163	142
R-squared	0.43	0.43	0.40	0.40	0.46	0.40	0.55	0.64

Regression example 4 (Brynjolfsson, Li & Raymond, 2025)

"**Generative AI at Work**."
QJE 140 (2): 889–942.
https://doi.org/10.1093/qje/qjae044.

Abstract: We study the staggered introduction of a generative AI–based conversational assistant using data from 5,172 customer-support agents. Access to AI assistance increases worker productivity, as measured by issues resolved per hour, by 15% on average, with substantial heterogeneity across workers.

We isolate the causal impact of access to AI recommendations using a standard difference-in-differences regression:

$$y_{it} = \delta_t + \alpha_i + \beta AI_{it} + \gamma X_{it} + \epsilon_{it}$$

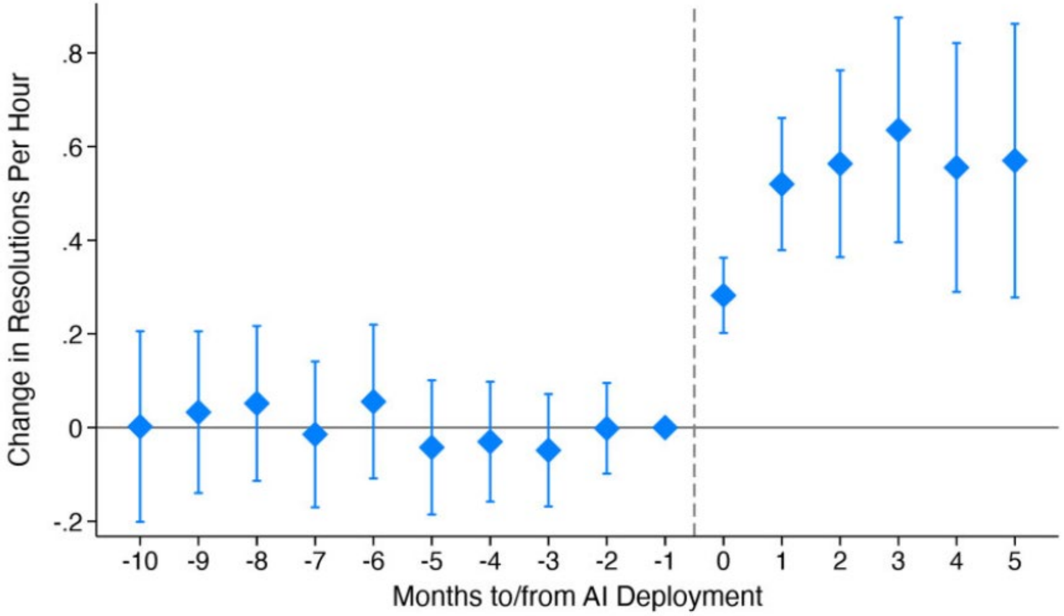
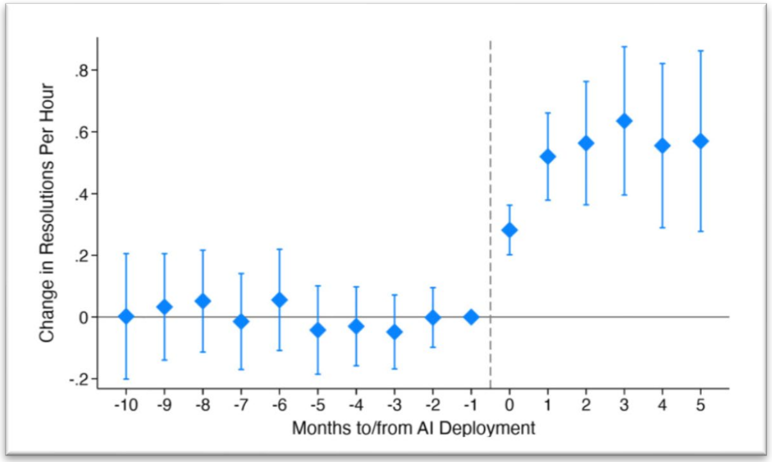


TABLE II

Main Effects: Productivity (Resolutions per Hour)

Variables	Resolutions/hour (1)	Resolutions/hour (2)	Resolutions/hour (3)
Post AI × Ever treated	0.469***	0.371***	0.301***
	(0.0325)	(0.0318)	(0.0329)
Ever treated	0.110**		
	(0.0440)		
Observations	13,192	12,295	12,295
R-squared	0.249	0.562	0.575
Year month FE	Yes	Yes	Yes
Location FE	Yes	Yes	Yes
Agent FE	—	Yes	Yes
Agent tenure FE	—	—	Yes
DV mean	2.123	2.176	2.176

Notes. This table presents the results of difference-in-difference regressions estimating the effect of AI model deployment on our main measure of productivity, resolutions per hour, the number of technical support problems resolved by an agent per hour (resolutions/hour). Post AI × Ever treated captures the impact of AI model deployment on resolutions per hour. Column (1) includes agent geographic location and year-by-month fixed effects. Columns (2) and (3) include agent-level fixed effects, and column (3), our preferred specification described by equation (1), also includes fixed effects that control for months of agent tenure. Observations for this regression are at the agent-month level and all standard errors are clustered at the agent level. Section IV.A describes the AI rollout procedure. Robust standard errors are in parentheses. * $p < .10$, ** $p < .05$, *** $p < .01$.



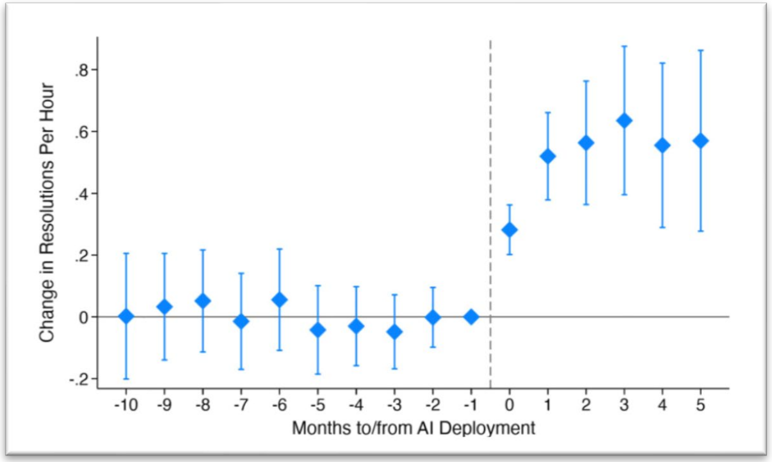
$$y_{it} = \delta_t + \alpha_i + \boxed{\beta AI_{it}} + \gamma X_{it} + \epsilon_{it}$$

TABLE II

MAIN EFFECTS: PRODUCTIVITY (RESOLUTIONS PER HOUR)

Variables	Resolutions/hour (1)	Resolutions/hour (2)	Resolutions/hour (3)
Post AI × Ever treated	0.469***	0.371***	0.301***
	(0.0325)	(0.0318)	(0.0329)
Ever treated	0.110**		
	(0.0440)		
Observations	13,192	12,295	12,295
R-squared	0.249	0.562	0.575
Year month FE	Yes	Yes	Yes
Location FE	Yes	Yes	Yes
Agent FE	—	Yes	Yes
Agent tenure FE	—	—	Yes
DV mean	2.123	2.176	2.176

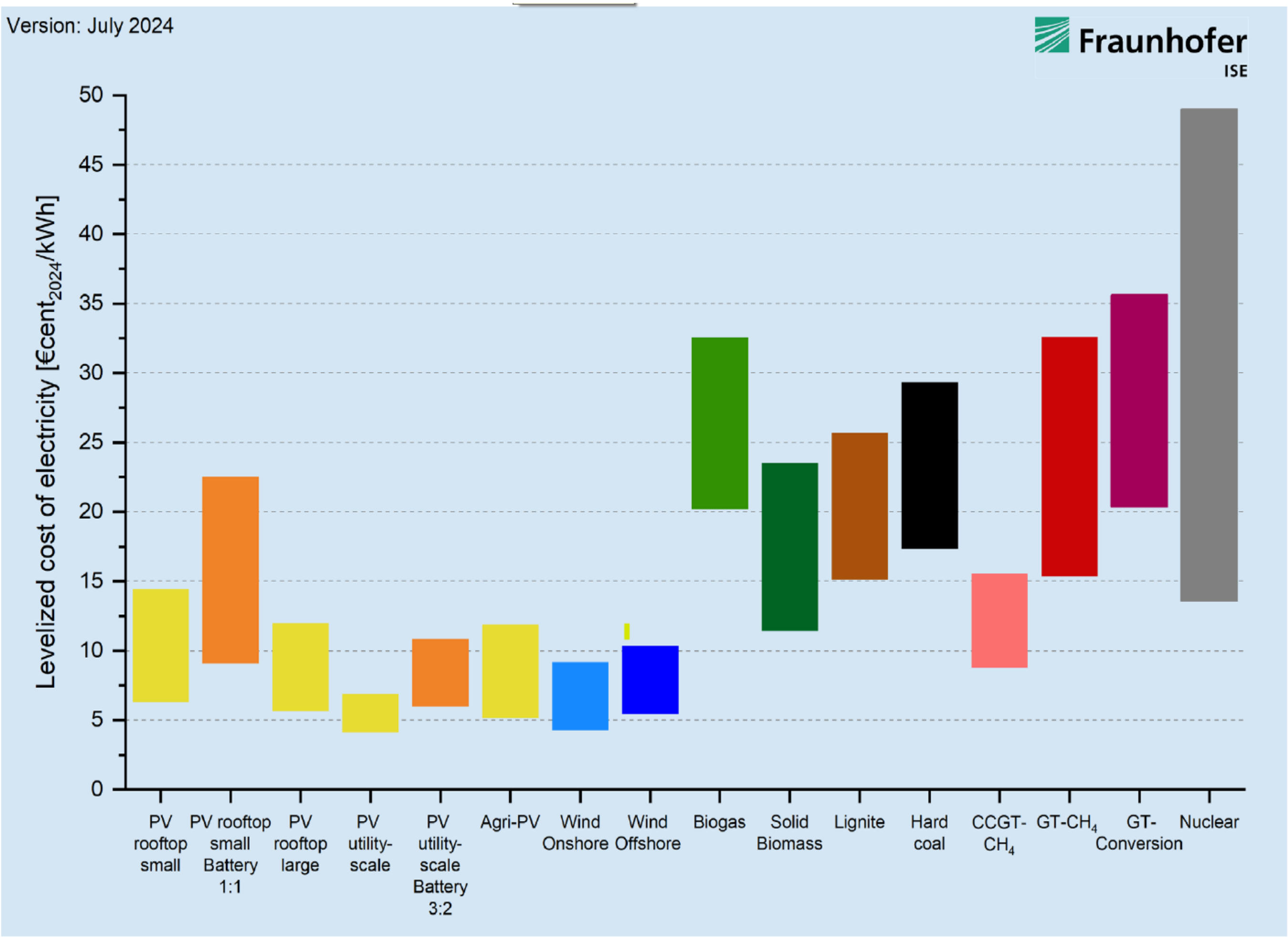
Notes. This table presents the results of difference-in-difference regressions estimating the effect of AI model deployment on our main measure of productivity, resolutions per hour, the number of technical support problems resolved by an agent per hour (resolutions/hour). Post AI × Ever treated captures the impact of AI model deployment on resolutions per hour. Column (1) includes agent geographic location and year-by-month fixed effects. Columns (2) and (3) include agent-level fixed effects, and column (3), our preferred specification described by equation (1), also includes fixed effects that control for months of agent tenure. Observations for this regression are at the agent-month level and all standard errors are clustered at the agent level. Section IV.A describes the AI rollout procedure. Robust standard errors are in parentheses. * $p < .10$, ** $p < .05$, *** $p < .01$.



$$y_{it} = \delta_t + \alpha_i + \beta AI_{it} + \gamma X_{it} + \epsilon_{it}$$

Regression example 4

How did solar energy become so cheap?



Regression example 4

Univariate regression model:

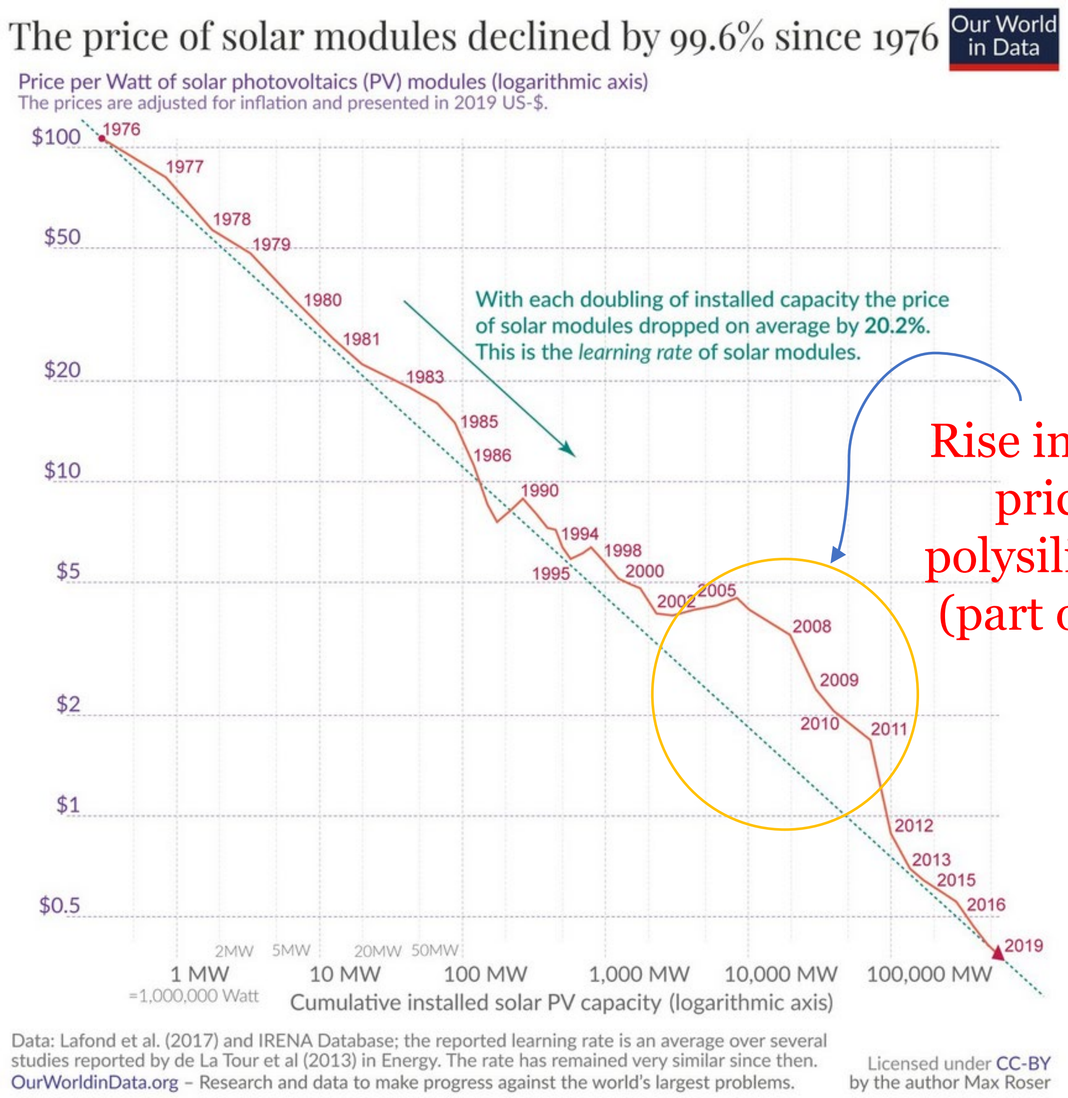
$$y_i = \beta_0 + \beta_1 x_i + u_i$$

where

y_i = log(Price) in year i

x_i = log(installed capacity) at year i

u_i = other unexplained factors



Rise in the price of polysilicon (part of u)

Estimating β with OLS

Wooldridge, Ch. 2; Greene, Ch. 3.

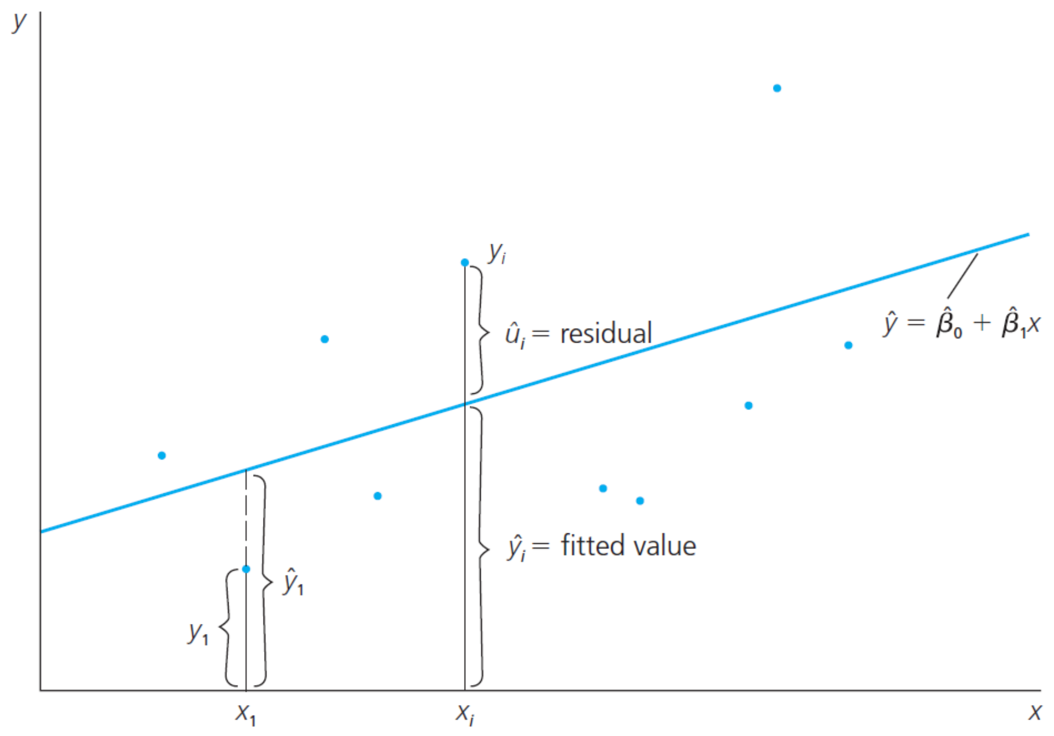
The basics: the linear univariate model

- Dataset with n observations $i = 1, \dots, n$.

- **Univariate** model:

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

where we have an intercept (β_0), just <u>one</u> explanatory variable (x) and everting else is captured by the unobserved error term (u).



The basics: the linear univariate model

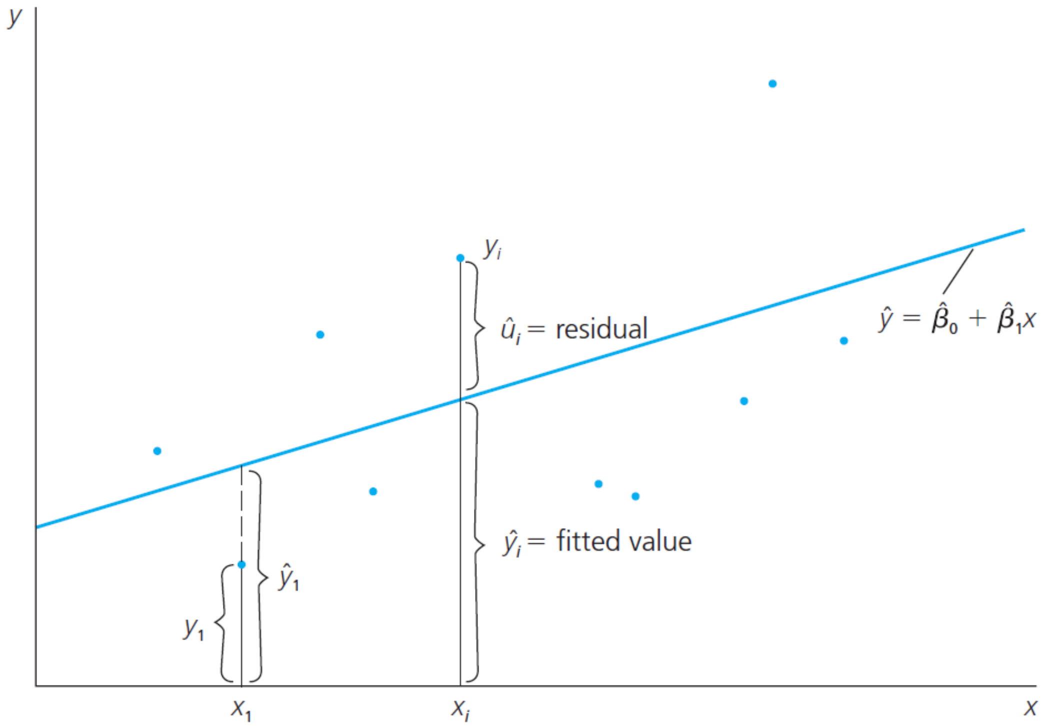
Important distinction:

- Theoretically-true “**population**”:

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

- Empirical **estimate** (or fit):

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$



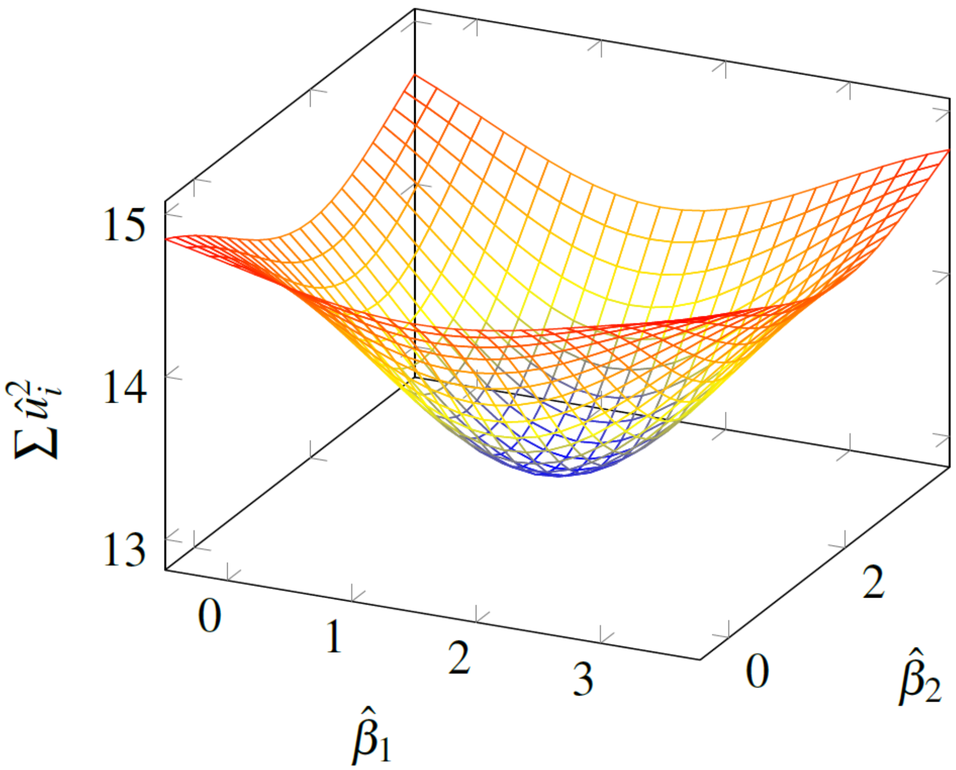
Note that the **residual** $\hat{u}_i = y_i - \hat{y}_i$ is also an estimate.

How do I estimate β?

Residuals $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i$ should be as small as possible.

We thus want to minimize the **sum of squared residuals**:

$$\mathrm{SSR}(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2$$



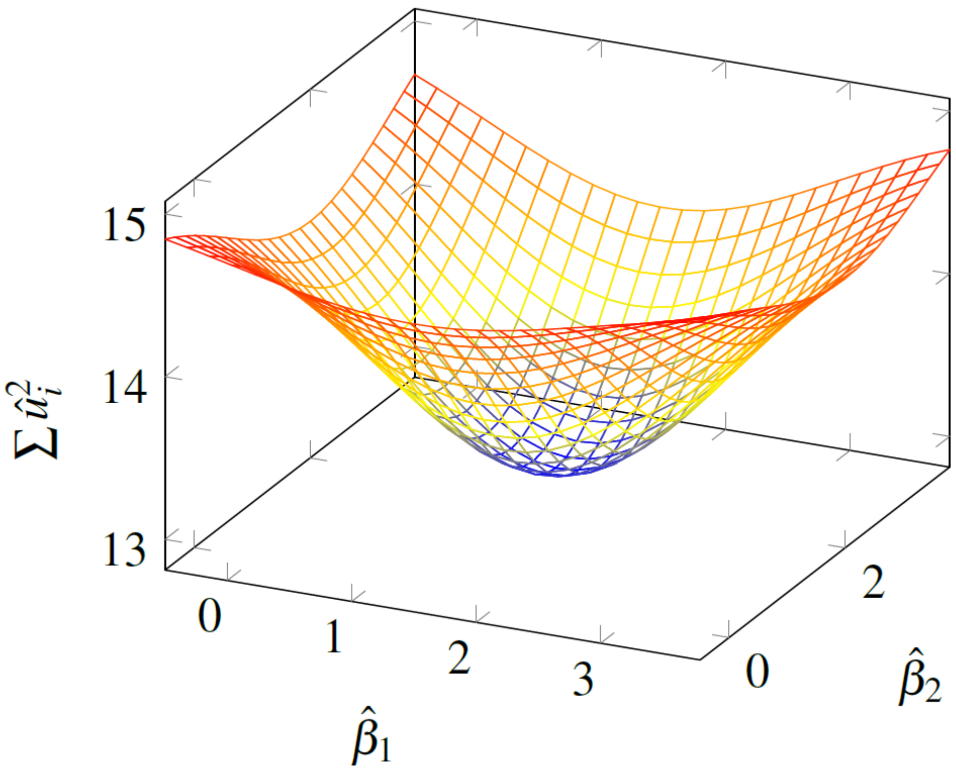
Hence the name: **Ordinary Least Squares (OLS)**

How do I estimate β?

First order conditions:

$$\frac{\partial \text{SSR}}{\partial \hat{\beta}_1} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0,$$

$$\frac{\partial \text{SSR}}{\partial \hat{\beta}_2} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0.$$



Combining:

$$\boxed{\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}.}$$

$$\boxed{\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\text{cov}(x, y)}{\text{var}(x)}.}$$

β in the multivariate case

If we have many regressors:

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2,i} + \ldots + \hat{\beta}_k x_{k,i} + \hat{u}_i \qquad \text{for all } i = 1, \ldots, n$$

$$\underset{(n\times1)}{y} = \hat{\beta}_1 \underset{(n\times1)}{1} + \hat{\beta}_2 \underset{(n\times1)}{x_2} + \hat{\beta}_3 \underset{(n\times1)}{x_3} + \ldots + \hat{\beta}_k \underset{(n\times1)}{x_k} + \underset{(n\times1)}{\hat{u}}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \hat{\beta}_1 + \begin{bmatrix} x_{2,1} \\ x_{2,2} \\ \vdots \\ x_{2,n} \end{bmatrix} \hat{\beta}_2 + \ldots + \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,n} \end{bmatrix} \hat{\beta}_k + \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix}$$

β in the multivariate case

If we have many regressors:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \hat{\beta}_1 + \begin{bmatrix} x_{2,1} \\ x_{2,2} \\ \vdots \\ x_{2,n} \end{bmatrix} \hat{\beta}_2 + \ldots + \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,n} \end{bmatrix} \hat{\beta}_k + \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & x_{2,1} & x_{3,1} & \cdots & x_{k,1} \\ 1 & x_{2,2} & x_{3,2} & \cdots & x_{k,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2,n} & x_{3,n} & \cdots & x_{k,n} \end{bmatrix}}_{X} \cdot \underbrace{\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}}_{\hat{\beta}} + \underbrace{\begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix}}_{\hat{u}}$$

β in the multivariate case

In compact notation

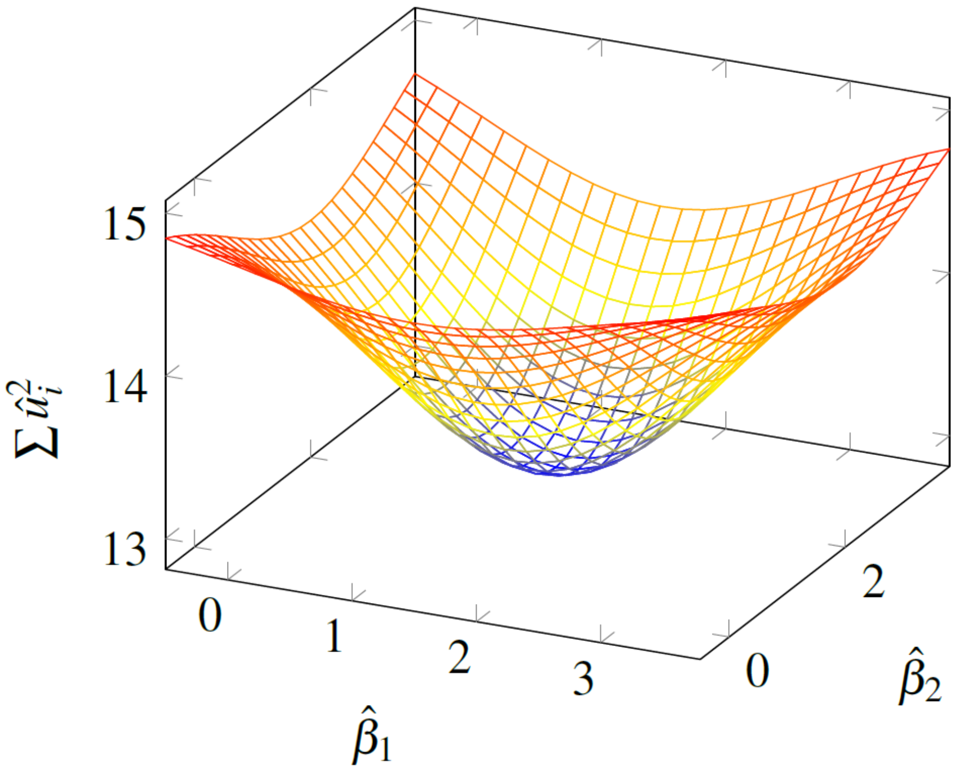
$$\underset{(n\times 1)}{y} = \underset{(n\times k)}{X}\underset{(k\times 1)}{\hat{\beta}} + \underset{(n\times 1)}{\hat{u}}$$

Again minimizing the SSR:

$$\sum_{i=1}^{n} \hat{u}_i^2 = \hat{u}'\hat{u} = (y - X\hat{\beta})'(y - X\hat{\beta})$$

This times yields

$$\boxed{\hat{\beta} = (X'X)^{-1}X'y}$$



- We will derive the OLS estimator $\hat{\beta}$ step by step soon and study its properties.
- Important for now: the intuition of where it comes from (the minimization of squared residuals)

- Next class (Thursday):
  - Optional reading: Wooldridge, Ch. 2.
  - We will use the free econometric software *R* (http://www.r-project.org/) and the user interface *Rstudio* (https://rstudio.com/products/rstudio/download/).
  - **Bring your laptops with both installed**.