SciencesPo
DEPARTMENT OF ECONOMICS

ScPoEconometrics: Advanced

Intro and Recap 1

Nikiforos Zampetakis based on Florian Oswald's slides
SciencesPo Paris
2024-01-29

Welcome to *ScPoEconometrics: Advanced*!

Today

1. Who Am I
2. This Course
3. Recap 1 of topics from intro course

Next time

- Quiz 1 (before next time)
- Recap 2

Who Am I

- I'm a PhD student in the Dept of Economics at SciencesPo Paris.
- I work on Empirical Industrial Organization and Antitrust Economics:

  1. How producers' cartels work?
  2. How the bargaining between producers and retailers affect the contracts they sign?
  3. Does product varieties lead to higher markups?
  4. How firms' productivity is affected by industrial policies?

- I combine theoretical models, *data and econometrics* to try answer these questions.
- For the empirical part of my work I mostly use `R`.

This Course

Prerequisites

- This course is the *follow-up* to Introduction to Econometrics with R which we teach to 2nd years.
- You are supposed to be familiar with all the econometrics material from the slides of that course and/or chapters 1-9 in our textbook.
- We also assume you have basic `R` working knowledge at the level of the intro course! ()
  - basic `data.frame` manipulation with `dplyr`
  - simple linear models with `lm`
  - basic plotting with `ggplot2`
  - Quiz 1 will try and test for that, so be on top of this chapter

This Course

Grading

1. There will be ***four quizzes*** on Moodle roughly every two weeks => 40%
2. There will be ***two take home exams / case studies*** => 60%
3. There will be ***no*** final exam.

Course Materials

1. Book chapter 10 onwards
2. The Slides
3. The interactive shiny apps
4. Quizzes on Moodle

Syllabus

1. Intro, Recap 1 (*Quiz 1*)
2. Recap 2 (*Quiz 2*)
3. Tools: `Rmarkdown` and `data.table`
4. Instrumental Variables 1 (*Quiz 3*)
5. Instrumental Variables 2 (*Midterm exam*)
6. Panel Data 1
7. Panel Data 2 (*quiz 4*)
8. Discrete Outcomes
9. Intro to Machine Learning 1
10. Intro to Machine Learning 2
11. Recap / Buffer (*Final Project*)
12. Recap / Buffer (*Final Project*)

Recap 1

Let's get cracking!

Population *vs.* sample

Models and notation

We write our (simple) population model

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

and our sample-based estimated regression model as

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

An estimated regression model produces estimates for each observation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

which gives us the *best-fit* line through our dataset.

(A lot of this set slides - in particular: pictures! - have been taken from Ed Rubin's outstanding material. Thanks Ed 🙏)

Task 1: Run Simple OLS (4 minutes)

1. Load data here. in `dta` format. (Hint: use `haven::read_dta("filename")` to read this format.)

2. Obtain common summary statistics for the variables `classize`, `avgmath` and `avgverb`. Hint: use the `skimr` package.

3. Estimate the linear model

$$\text{avgmath}_i = \beta_0 + \text{classize}_i x_i + u_i$$

Task 1: Solution

1. Load the data

```
grades = haven::read_dta(file ="https://www.dropbox.com/s/wwp2cs9f0dubmhr/grade5.dta?dl=1")
```

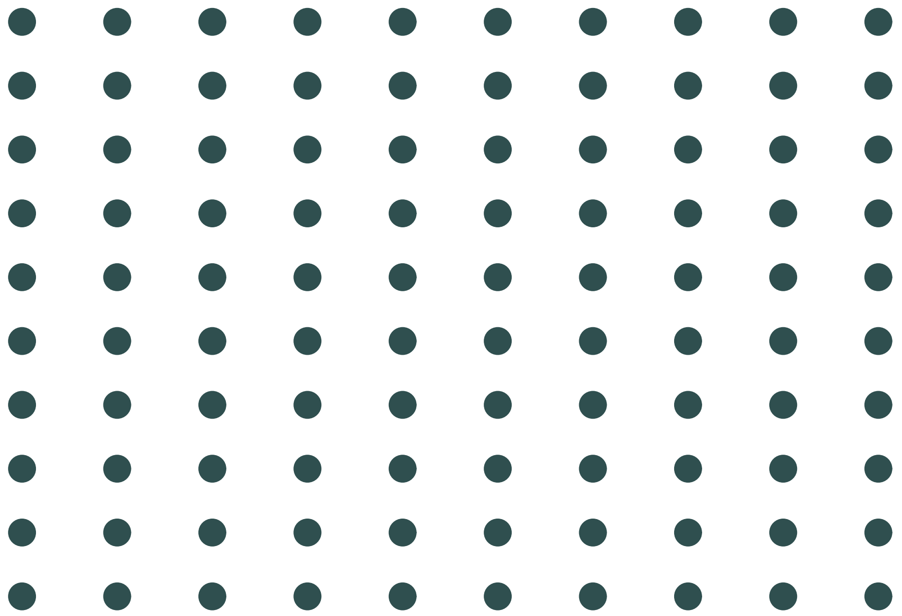
1. Describe the dataset:

```
library(dplyr)
grades %>%
select(classize,avgmath,avgverb) %>%
skimr::skim()
```

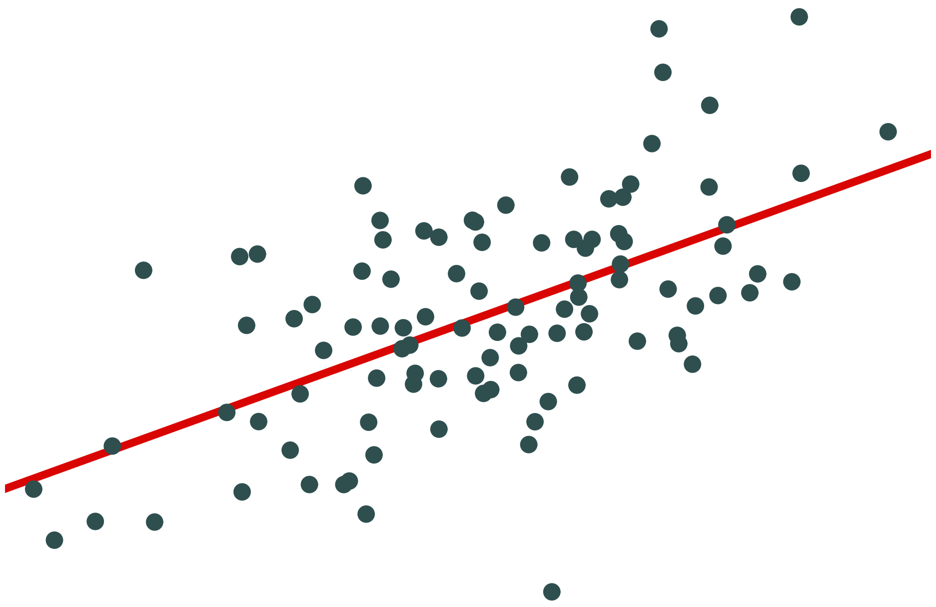
2. Run OLS to estimate the relationship between class size and student achievement?

```
summary(lm(formula = avgmath ~ classize, data = grades))
```

Question: Why do we care about *population vs. sample*?



Population

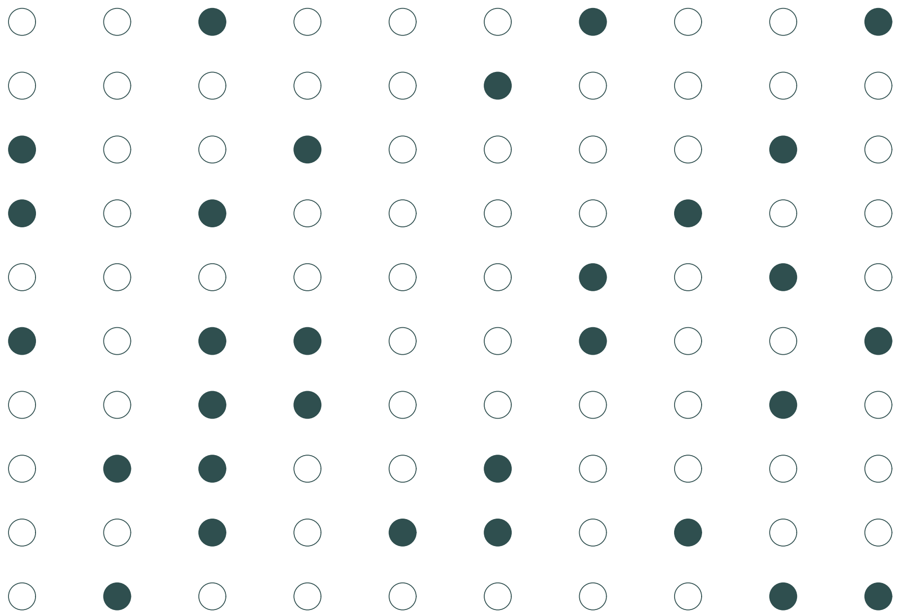


Population relationship

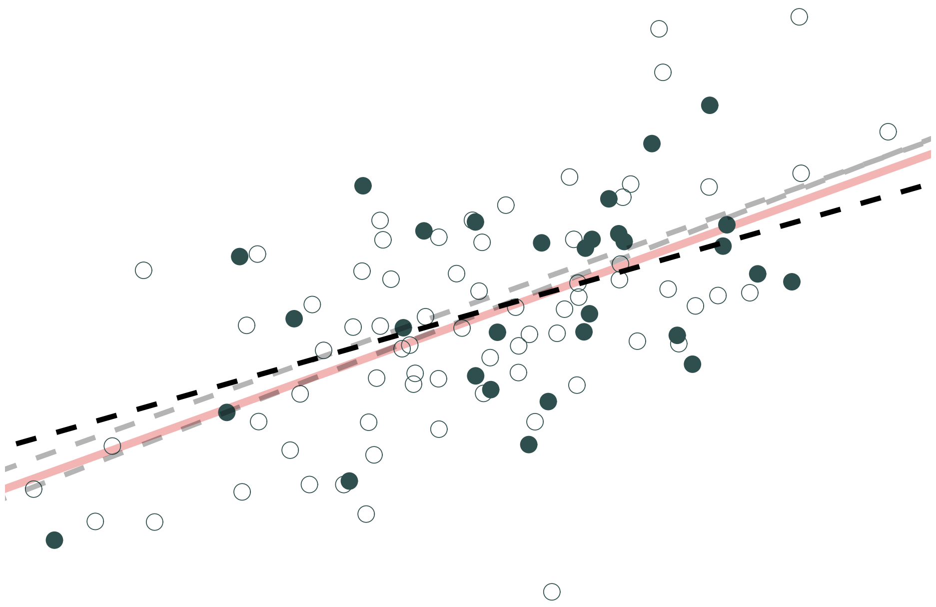
$$y_i = 2.53 + 0.57 x_i + u_i$$

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

Question: Why do we care about *population vs. sample*?



Sample 3: 30 random individuals

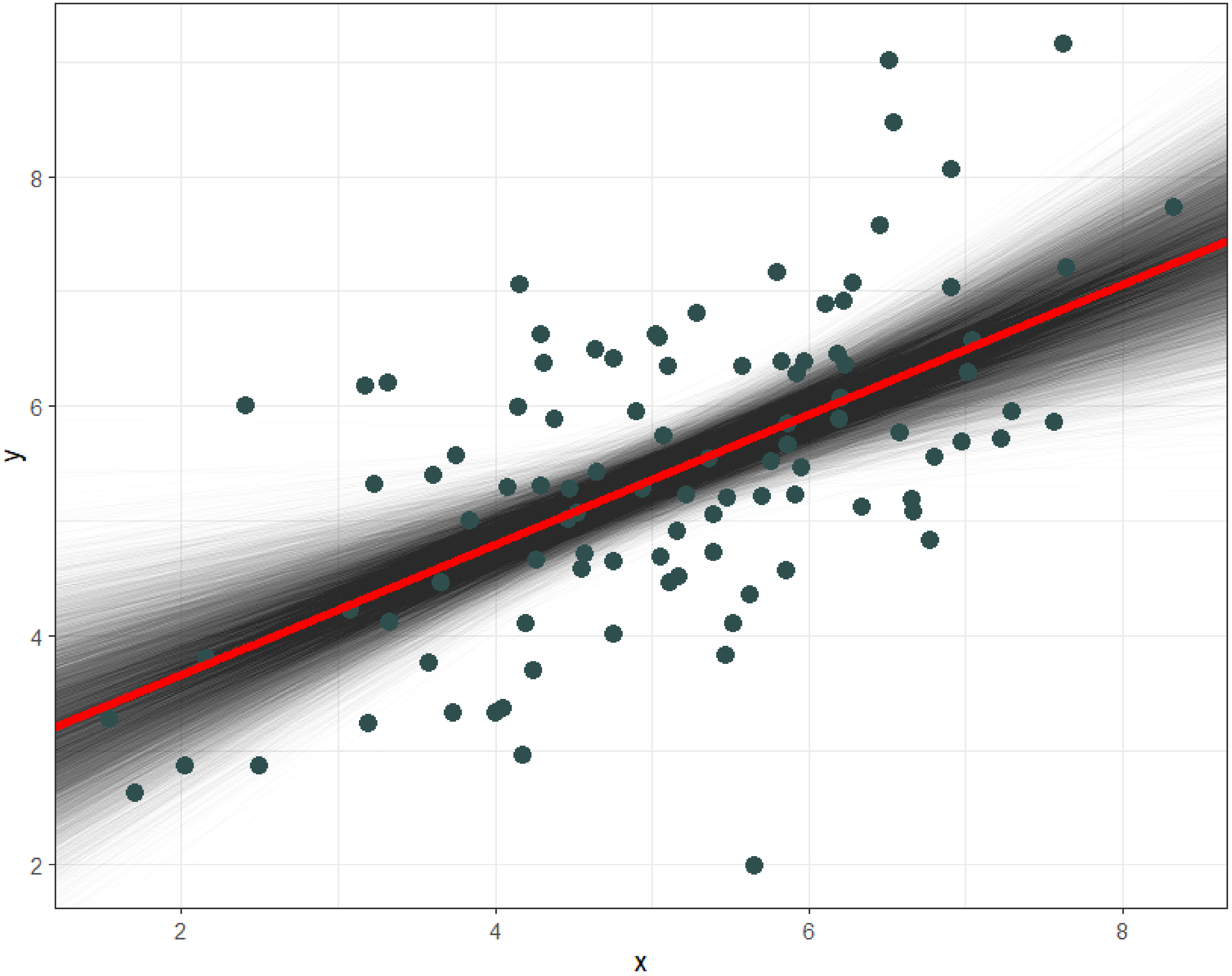


Population relationship

$y_i = 2.53 + 0.57 x_i + u_i$

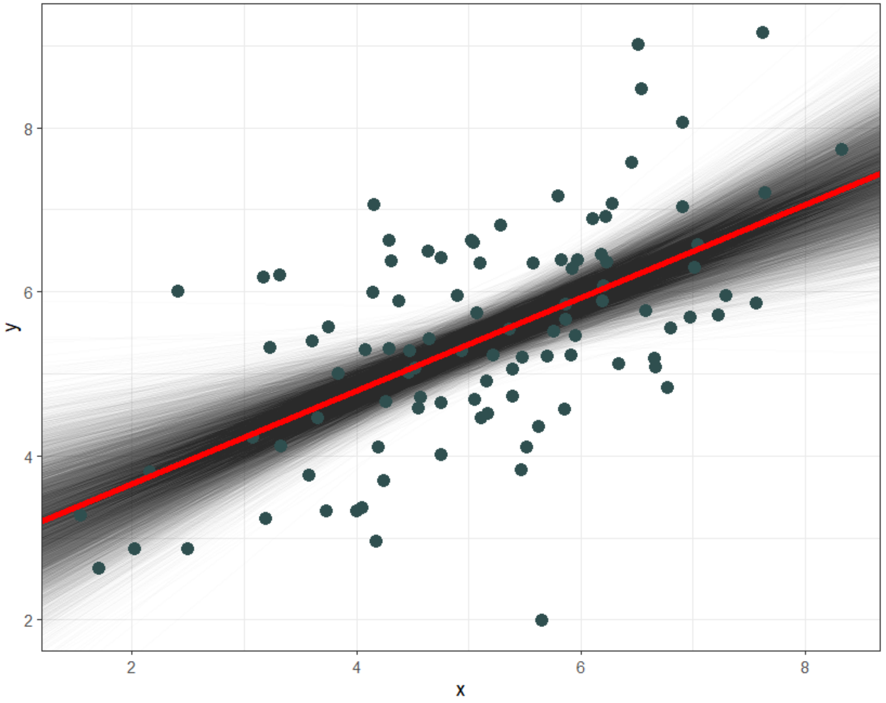
Sample relationship

$\hat{y}_i = 3.21 + 0.45 x_i$


8
6
4
2
y
2
4
6
8
x

Population *vs.* sample

Question: Why do we care about *population vs. sample*?



- On **average**, our regression lines match the population line very nicely.
- However, **individual lines** (samples) can really miss the mark.
- Differences between individual samples and the population lead to **uncertainty** for the econometrician.

Population *vs.* sample

Question: Why do we care about *population vs. sample*?

Answer: Uncertainty matters.

- Every random sample of data is different.
- Our (OLS) estimators are computed from those samples of data.
- If there is sampling variation, there is variation in our estimates.

- OLS inference depends on certain assumptions.
- If violated, our estimates will be biased or imprecise.
- Or both.

Linear regression

The estimator

We can estimate a regression line in `R` (`lm(y ~ x, my_data)`) and `stata` (`reg y x`). But where do these estimates come from?

A few slides back:

> $$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
>
> which gives us the *best-fit* line through our dataset.

But what do we mean by "best-fit line"?

Being the "best"

Question: What do we mean by *best-fit line*?

Answers:

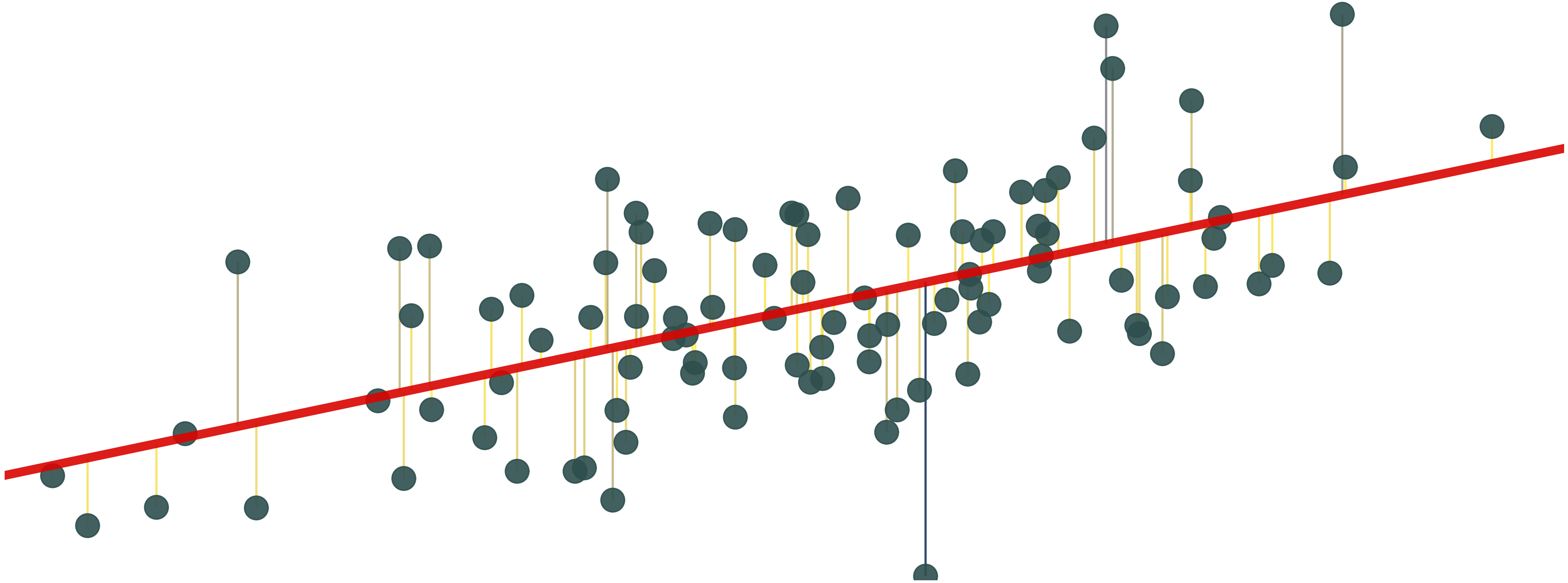
- In general (econometrics), *best-fit line* means the line that minimizes the sum of squared errors (SSE):

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 \quad \text{where} \quad e_i = y_i - \hat{y}_i$$

- Ordinary **least squares** (**OLS**) minimizes the sum of the squared errors.
- Based upon a set of (mostly palatable) assumptions, OLS
  - Is unbiased (and consistent)
  - Is the *best* (minimum variance) linear unbiased estimator (BLUE)

OLS *vs.* other lines/estimators

The OLS estimate is the combination of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize SSE.



```r
ScPoApps::launchApp("simple_reg")
```

OLS

Formally

In simple linear regression, the OLS estimator comes from choosing the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared errors (SSE), *i.e.*,

$$\min_{\hat{\beta}_0,\,\hat{\beta}_1} \text{SSE}$$

but we already know $\text{SSE} = \sum_i e_i^2$. Now use the definitions of e_i and $\hat{y}$.

$$\begin{aligned} e_i^2 &= \left(y_i - \hat{y}_i\right)^2 = \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \\ &= y_i^2 - 2 y_i \hat{\beta}_0 - 2 y_i \hat{\beta}_1 x_i + \hat{\beta}_0^2 + 2 \hat{\beta}_0 \hat{\beta}_1 x_i + \hat{\beta}_1^2 x_i^2 \end{aligned}$$

Recall: Minimizing a multivariate function requires (**1**) first derivatives equal zero (the *1st-order conditions*) and (**2**) second-order conditions (concavity).

OLS

Interactively

```r
ScPoApps::launchApp("SSR_cone")
```

OLS

Interactively

We skipped the maths.

We now have the OLS estimators for the slope

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

and the intercept

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Remember that *those* two formulae are amongst the very few ones from the intro cours that you should know by heart!

We now turn to the assumptions and (implied) properties of OLS.

OLS: Assumptions and properties

Question: What properties might we care about for an estimator?

Tangent: Let's review statistical properties first.

OLS: Assumptions and properties

Refresher: Density functions

Recall that we use **probability density functions** (PDFs) to describe the probability a **continuous random variable** takes on a range of values. (The total area = 1.)

These PDFs characterize probability distributions, and the most common/famous/popular distributions get names (*e.g.*, normal, *t*, Gamma).

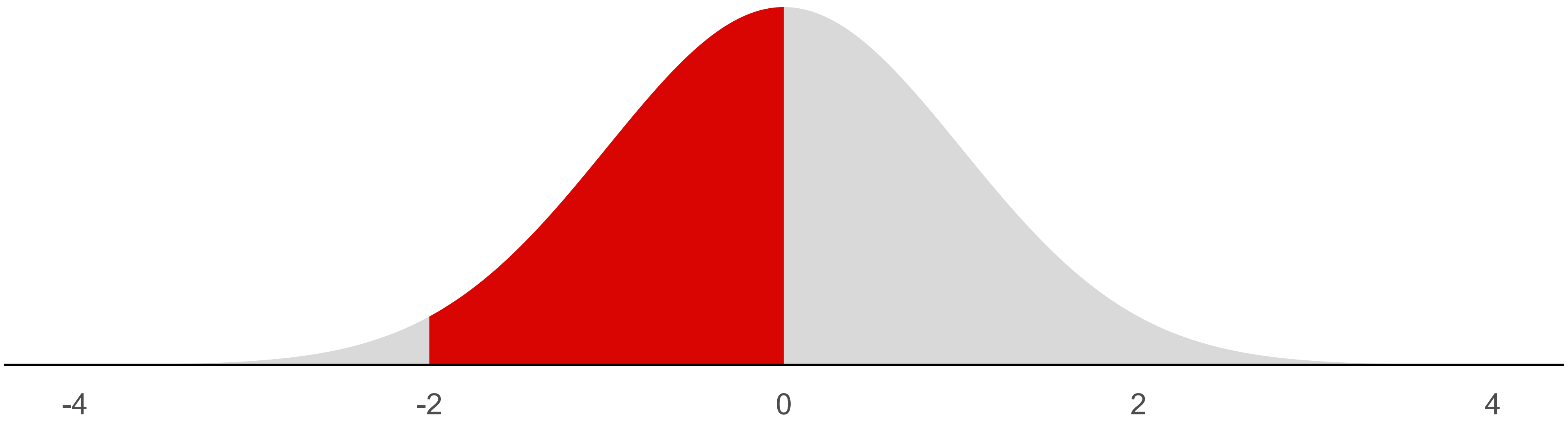
Here is the definition of a *PDF* f_X for a *continuous* RV X:

$$\mathrm{Pr}[a \leq X \leq b] \equiv \int_a^b f_X(x) dx$$

OLS: Assumptions and properties

Refresher: Density functions

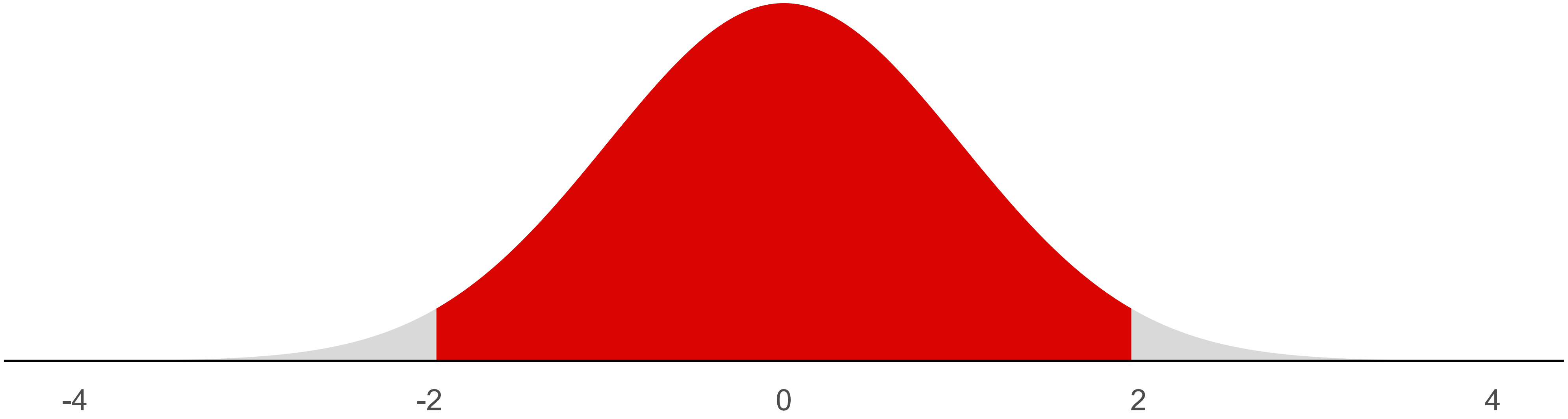
The probability a standard normal random variable takes on a value between -2 and 0: $\mathrm{P}(-2 \leq X \leq 0) = 0.48$



OLS: Assumptions and properties

Refresher: Density functions

The probability a standard normal random variable takes on a value between -1.96 and 1.96:
$\mathrm{P}(-1.96 \leq X \leq 1.96) = 0.95$



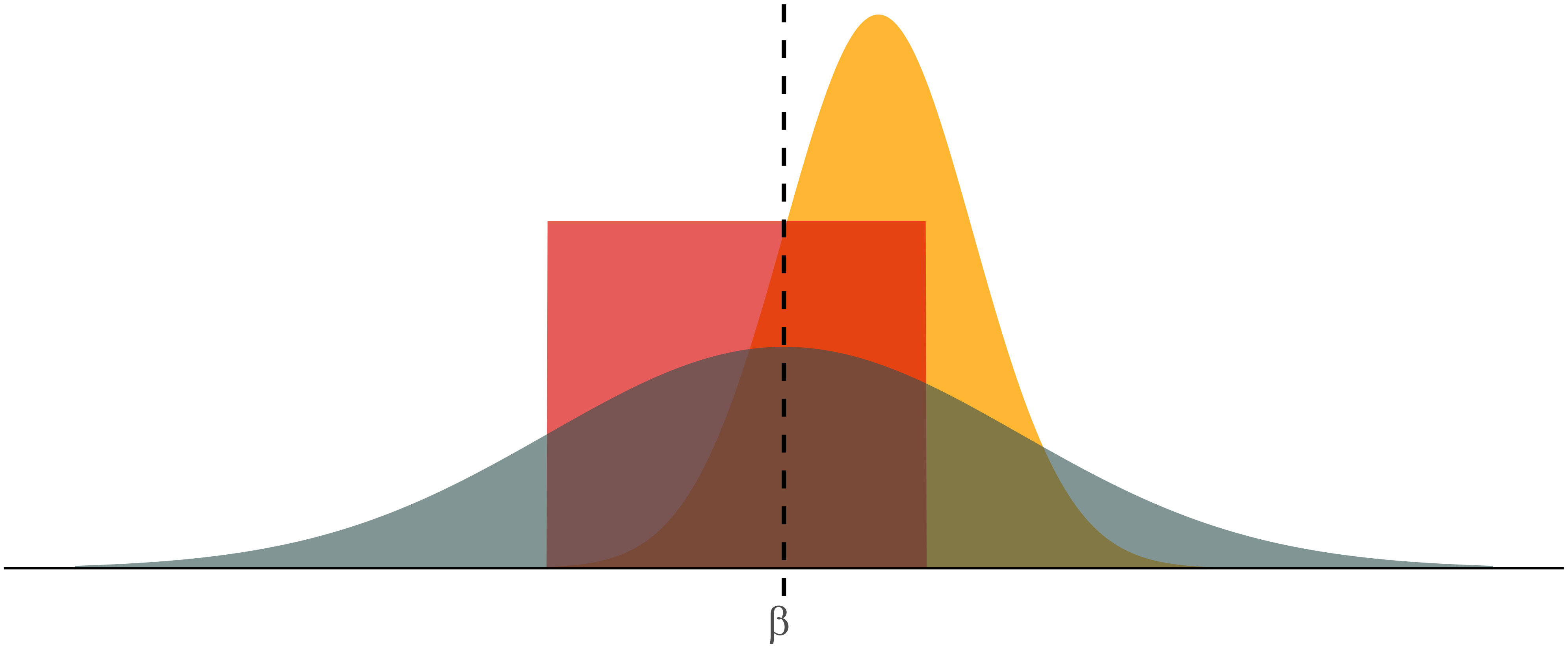
OLS: Assumptions and properties

Refresher: Density functions

The probability a standard normal random variable takes on a value beyond 2: $\mathrm{P}(X > 2) = 0.023$

OLS: Assumptions and properties

Imagine we are trying to estimate an unknown parameter β, and we know the distributions of three competing estimators. Which one would we want? How would we decide?



OLS: Assumptions and properties

Question: What properties might we care about for an estimator?

Answer one: Bias.

On average (after *many* samples), does the estimator tend toward the correct value?

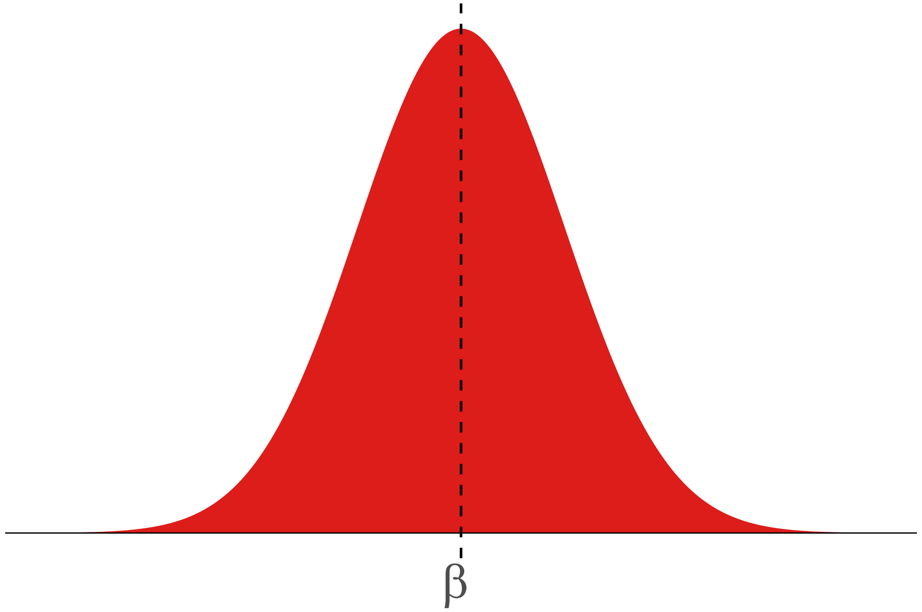
More formally: Does the mean of estimator's distribution equal the parameter it estimates?

$$\mathop{\text{Bias}}_\beta \left(\hat{\beta} \right) = \boldsymbol{E}\left[\hat{\beta} \right] - \beta$$

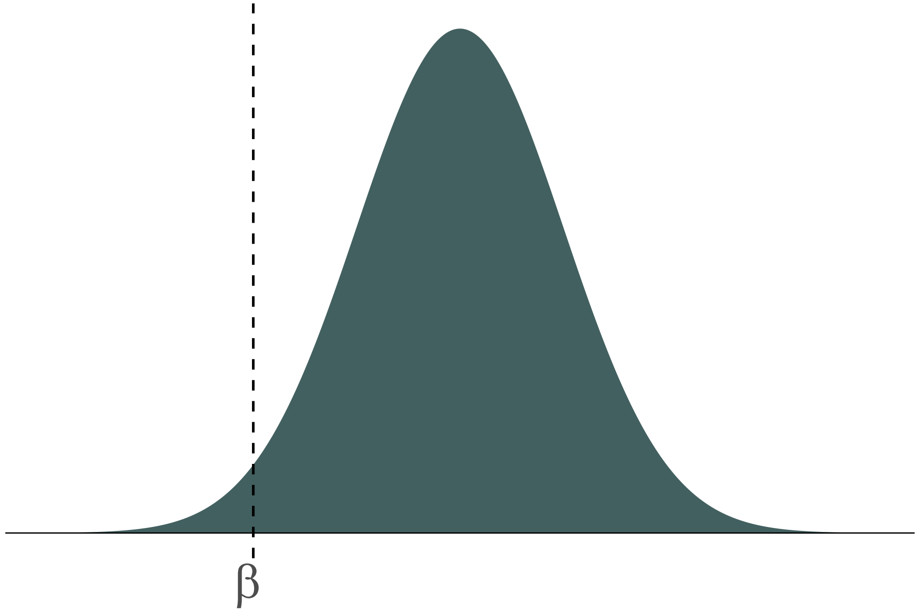
OLS: Assumptions and properties

Answer one: Bias.

Unbiased estimator: $\boldsymbol{E}\left[\hat{\beta}\right] = \beta$

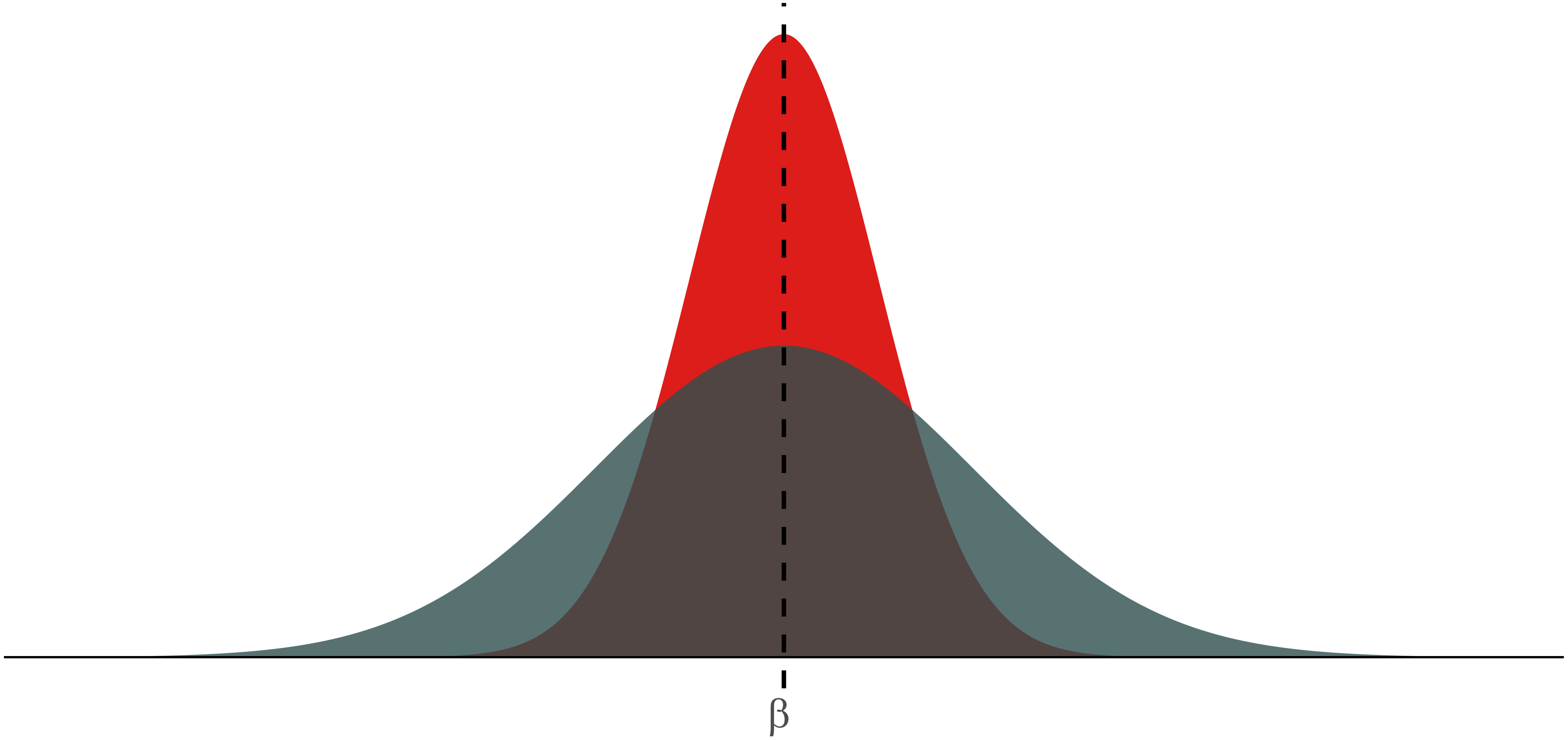


Biased estimator: $\boldsymbol{E}\left[\hat{\beta}\right] \neq \beta$



OLS: Assumptions and properties

Answer two: Variance.



OLS: Assumptions and properties

Answer one: Bias.

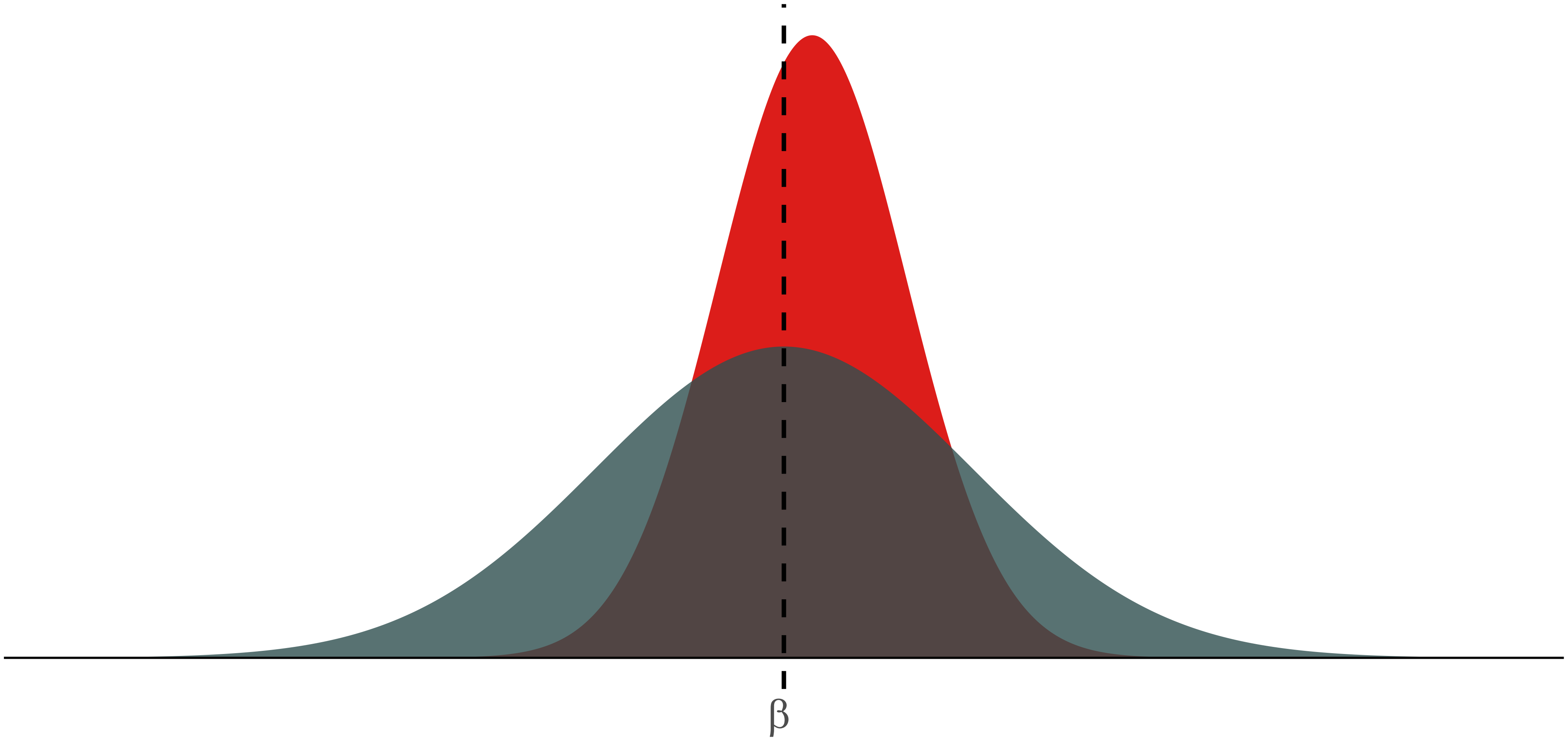
Answer two: Variance.

Subtlety: The bias-variance tradeoff.

Should we be willing to take a bit of bias to reduce the variance?

In econometrics, we generally stick with unbiased (or consistent) estimators. But other disciplines (especially computer science) think a bit more about this tradeoff.

The bias-variance tradeoff.



OLS: Assumptions and properties

Properties

As you might have guessed by now,

- OLS is **unbiased**.
- OLS has the **minimum variance** of all unbiased linear estimators.

OLS: Assumptions and properties

Properties

But... these (very nice) properties depend upon a set of assumptions:

1. The population relationship is linear in parameters with an additive disturbance.

2. Our X variable is **exogenous**, *i.e.*, $\boldsymbol{E}[u|X] = 0$.

3. The X variable has variation. And if there are multiple explanatory variables, they are not perfectly collinear.

4. The population disturbances u_i are independently and identically distributed as normal random variables with mean zero $\left(\boldsymbol{E}[u] = 0\right)$ and variance σ^2 (*i.e.*, $\boldsymbol{E}\left[u^2\right] = \sigma^2$). Independently distributed and mean zero jointly imply $\boldsymbol{E}\left[u_i u_j\right] = 0$ for any $i \neq j$.

OLS: Assumptions and properties

Assumptions

Different assumptions guarantee different properties:

- Assumptions (1), (2), and (3) make OLS unbiased.
- Assumption (4) gives us an unbiased estimator for the variance of our OLS estimator.

We will discuss solutions to **violations of these assumptions**. See also our discussion in the book

- Non-linear relationships in our parameters/disturbances (or misspecification).
- Disturbances that are not identically distributed and/or not independent.
- Violations of exogeneity (especially omitted-variable bias).

OLS: Assumptions and properties

Conditional expectation

For many applications, our most important assumption is **exogeneity**, *i.e.*,

$$E[u|X] = 0$$

but what does it actually mean?

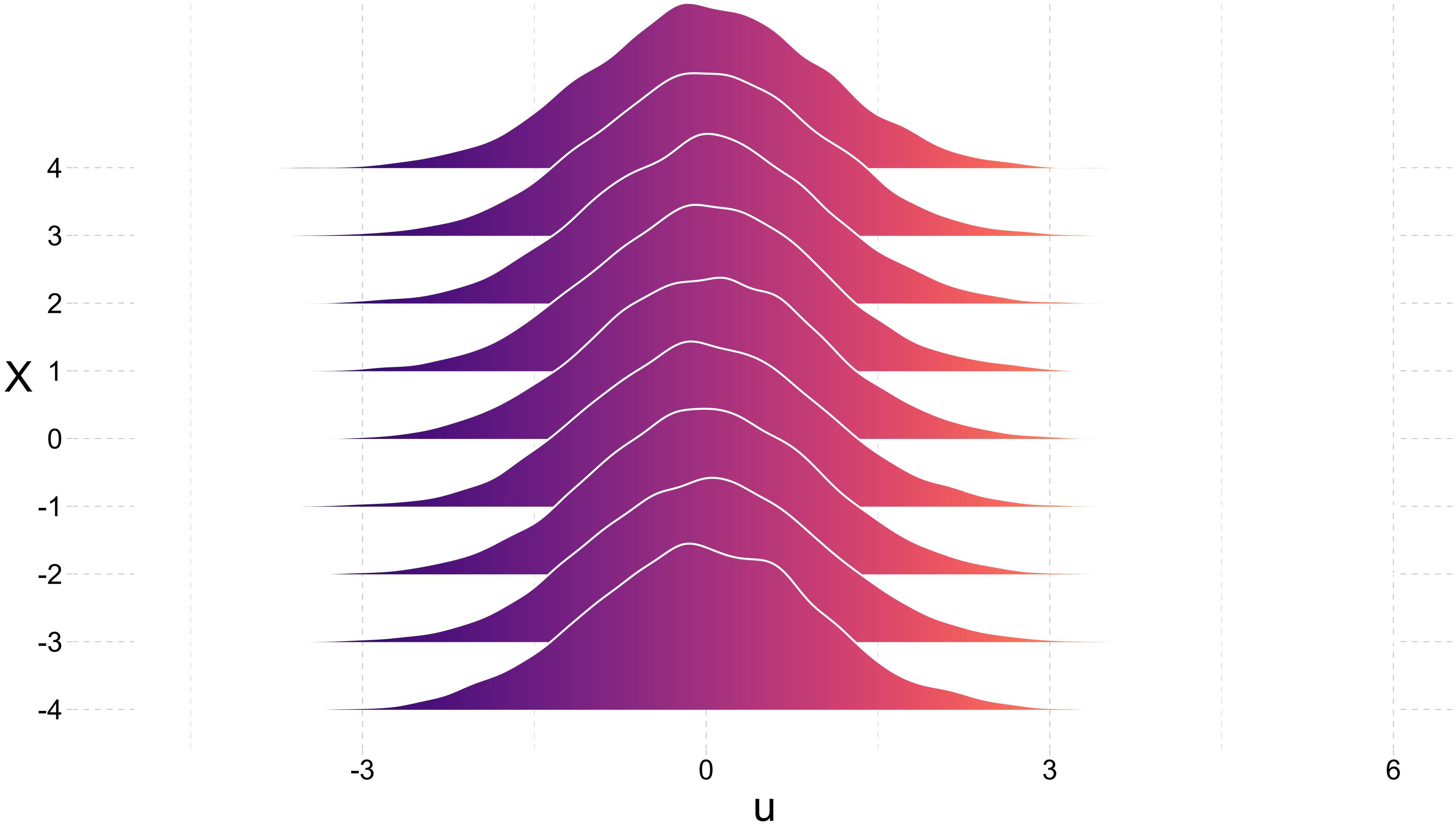
One way to think about this definition:

> For *any* value of X, the mean of the residuals must be zero.

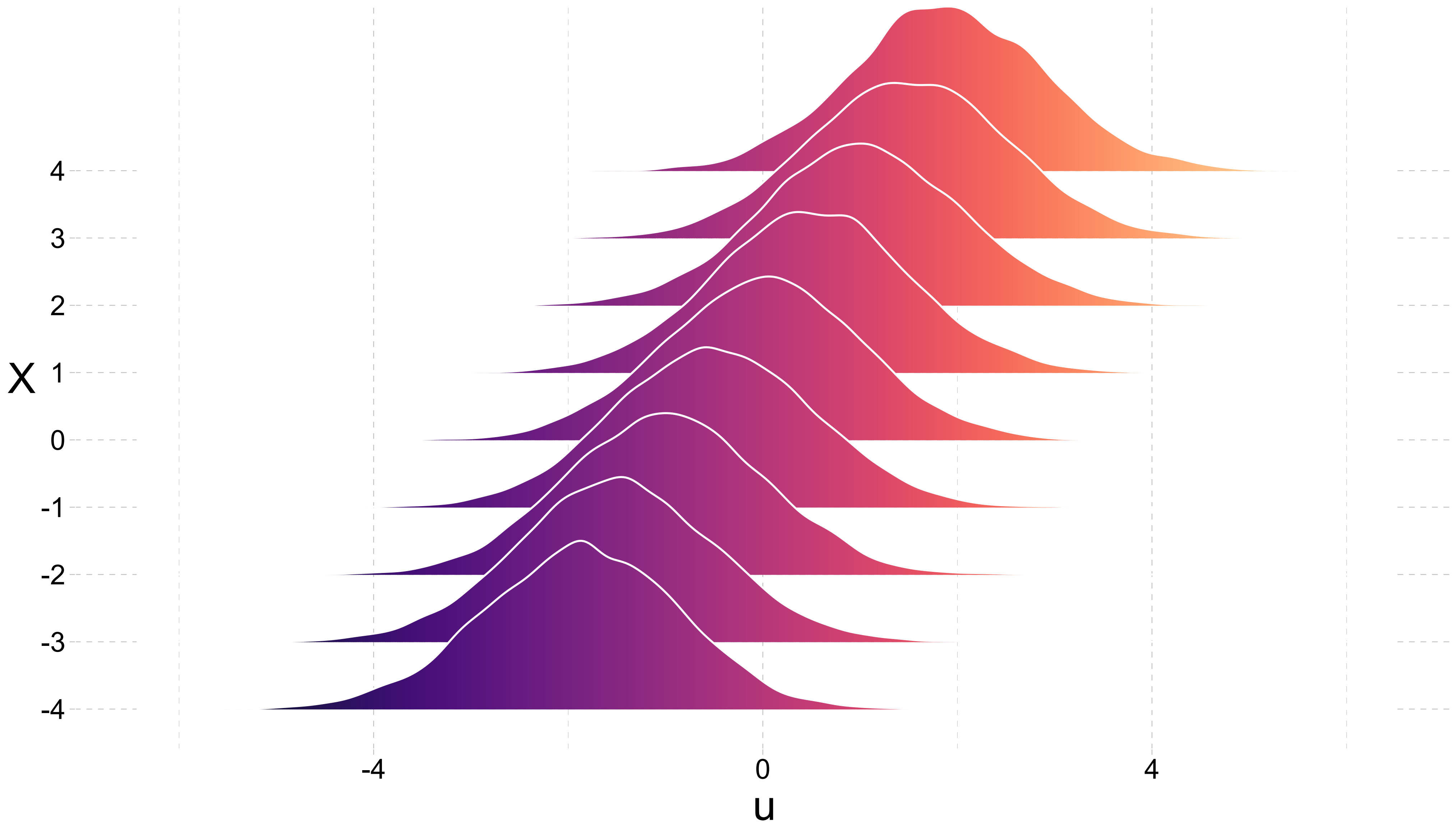
- *E.g.*, $E[u|X=1]=0$ *and* $E[u|X=100]=0$
- *E.g.*, $E[u|X_2=\text{Female}]=0$ *and* $E[u|X_2=\text{Male}]=0$
- Notice: $E[u|X]=0$ is more restrictive than $E[u]=0$

Graphically…

Valid exogeneity, *i.e.*, $E[u|X] = 0$



Invalid exogeneity, *i.e.*, $E[u|X] \neq 0$



SciencesPo
DEPARTMENT OF ECONOMICS

Thanks!

This is the final slide

you can add your email, twitter, github, etc. info here

Here is an example:

	nikiforos.zampetakis@sciencespo.fr
	slides
	@ScPoEcon
	@ScPoEcon