Module 6: More on Identification and Other Issues

Econometrics II

Max Heinze (mheinze@wu.ac.at)
Department of Economics, WU Vienna

Sannah Tijani (stijani@wu.ac.at)
Department of Economics, WU Vienna

January 15, 2026

Panel Data

Quasi-Experiments

More Methods

Introduction to Panel Data

We already know **cross-sectional data** well. Cross-sectional data covers **many individuals** at **one point in time**:

$$x_i$$

In Econometrics III / Applied Econometrics, we will learn about **time series data**. Time series data covers **one individual** at **many different points in time**:

$$x_t$$

Today, we will talk briefly about **panel data** because it is very useful for answering causal questions. In a panel, we follow **many individuals** over **many time periods**:

$$x_{it}$$

Panel data is especially useful because it allows us to **control for some unobserved effects** without actually observing them.

Examples

This is an example of a **panel dataset**. You can see that we have data on **two individuals** for **two points in time**.

Individual	Date	Income	Age	Education
A	2020	1200	20	medium
A	2021	1300	21	medium
B	2020	1800	24	medium
B	2021	2600	25	high

Panel data is not as uncommon as you might imagine. Examples for panel data include:

- Many **surveys**, e.g. the EU-SILC (Statistics on Income and Living Conditions) or the HFCS (Household Finance and Consumption Survey).
- Most **remotely sensed data**, e.g. data on temperature or precipitation that is measured by satellites.
- Data that **companies** have on their user base. Google follows me through time, and so does it with you.

Why Panel Data?

Panel data and models have some useful **advantages**, such as:

- **More data** is good.
- We can **follow** relationships **over time**.
- We can consider **unobserved** individual or time-specific **effects**.

However, there also some potential **issues**, including:

- **panel mortality** (individuals drop out),
- **panel effects** (impacts of repeated data collection), and
- decreasing marginal returns of observations.

Pooled Cross-Sections

The **simplest model** can be obtained by stacking cross-sectional models like this:

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_T \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_T \end{pmatrix} + \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_T \end{pmatrix}.$$

Alternatively, we can write down the model for **a single cross-sectional unit** like this:

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_{it}.$$

This is what we call **pooled cross-sections**. **Coefficients** are assumed constant across time and individuals.

In this setting, (pooled) **OLS** is **consistent** as long as the assumption $\mathrm{E}(u_{it} \mid x_{it}) = 0$ holds.

Fixed Effects

In many cases, this assumption is problematic. Let's start discussing this by splitting up the error term in two components, **time-invariant heterogeneity between individuals** μ_i and a **time-varying error** ε_{it}:

$$u_{it} = \mu_i + \varepsilon_{it}.$$

There are now two **problems** with the assumption that $\mathrm{E}(u_{it} \mid x_{it}) = 0$, which is equivalent to

$$\mathrm{E}(u_{it} \mid x_{i1}, x_{i2}, \ldots, x_{iT}) = 0 \text{ for } i = 1, 2, \ldots, N.$$

- The **unobserved individual characteristics** μ_i are likely correlated with the treatment (and outcome).
- Within individuals, the **time-varying error** ε_{it} at a certain t will likely depend on the value of the error at $t - 1$.

The Fixed Effects Estimator (Within Estimator)

We can circumvent this problem by considering **individual fixed effects** explicitly:

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \mu_i + \varepsilon_{it}.$$

This is equivalent to estimating the model with **dummy variables for each individual**. These fixed effects capture unobserved individual heterogeneity.

- The resulting estimator is called the **Within Estimator** (because we are comparing observations "within" an individual) or alternatively, the **Fixed Effects Estimator**.
- If we additionally use **time fixed effects**, we call the estimator **Two-Way Fixed Effects** (**TWFE**) Estimator.

There are **two things** that these estimators **cannot do**:

- They cannot capture heterogeneity that is **both time-varying and individual-specific**.
- Even if there is no such heterogeneity, and thus all unobserved factors are captured, this does not remedy **endogeneity from other sources** such as reverse causality.

Difference-in-Differences

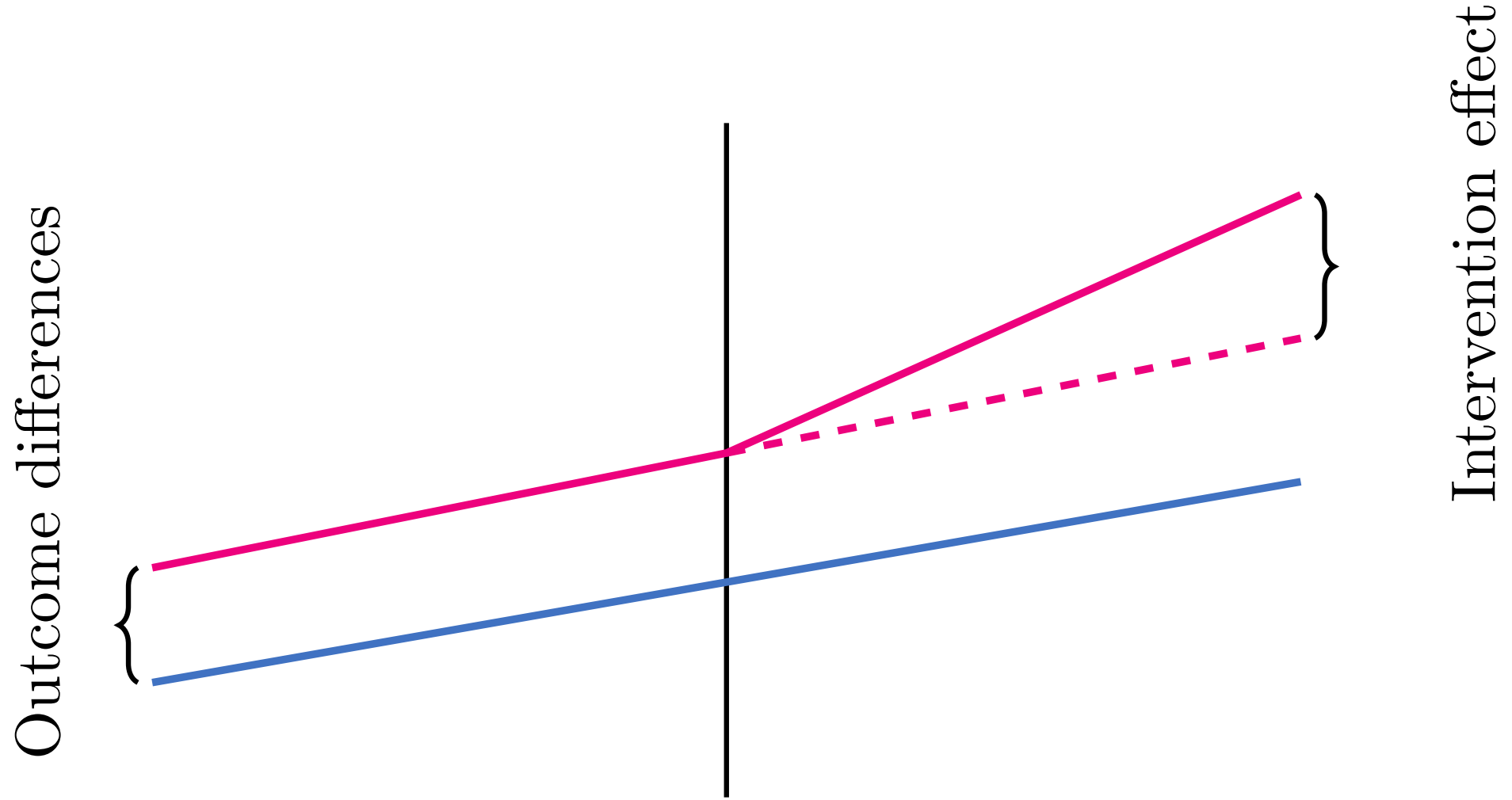
If we have panel data, we can use a **difference-in-differences** (**DiD**) approach. We divide our data in four and estimate:

$$y_{it} = \alpha + x_{\text{after}}\phi + x_{\text{treated}}\theta + x_{\text{interacted}}\delta + \dots$$

	Before	**After**	**Difference**
Control	α	$\alpha + \phi$	ϕ
Treatment	$\alpha + \theta$	$\alpha + \theta + \phi + \delta$	$\phi + \delta$
Difference	θ	$\theta + \delta$	δ

By **comparing** and evaluating the **difference** between the two **before-after differences** (one for the treatment group, and one for the control group), we can directly obtain the treatment effect, $\hat{\delta}$.

Diff-in-Diff Illustration



Panel Data

Quasi-Experiments

More Methods

Natural Experiments

A **natural experiment** is a study where an **experimental setting** is **induced by nature** or other factors outside our control.

- It is an **observational study** with properties of randomised experiments.
- This provides a good basis for **causal inference**, and
- does not suffer from **potential issues** of a conducting an experiment, such as cost, ethics, feasibility, etc.
- For a natural experiment, we need something to **happen exogenously** and **create variation in treatment**.

U.S. Representative Alexander Pirnie of New York drawing the first capsule in the Vietnam war draft lottery.

Cholera

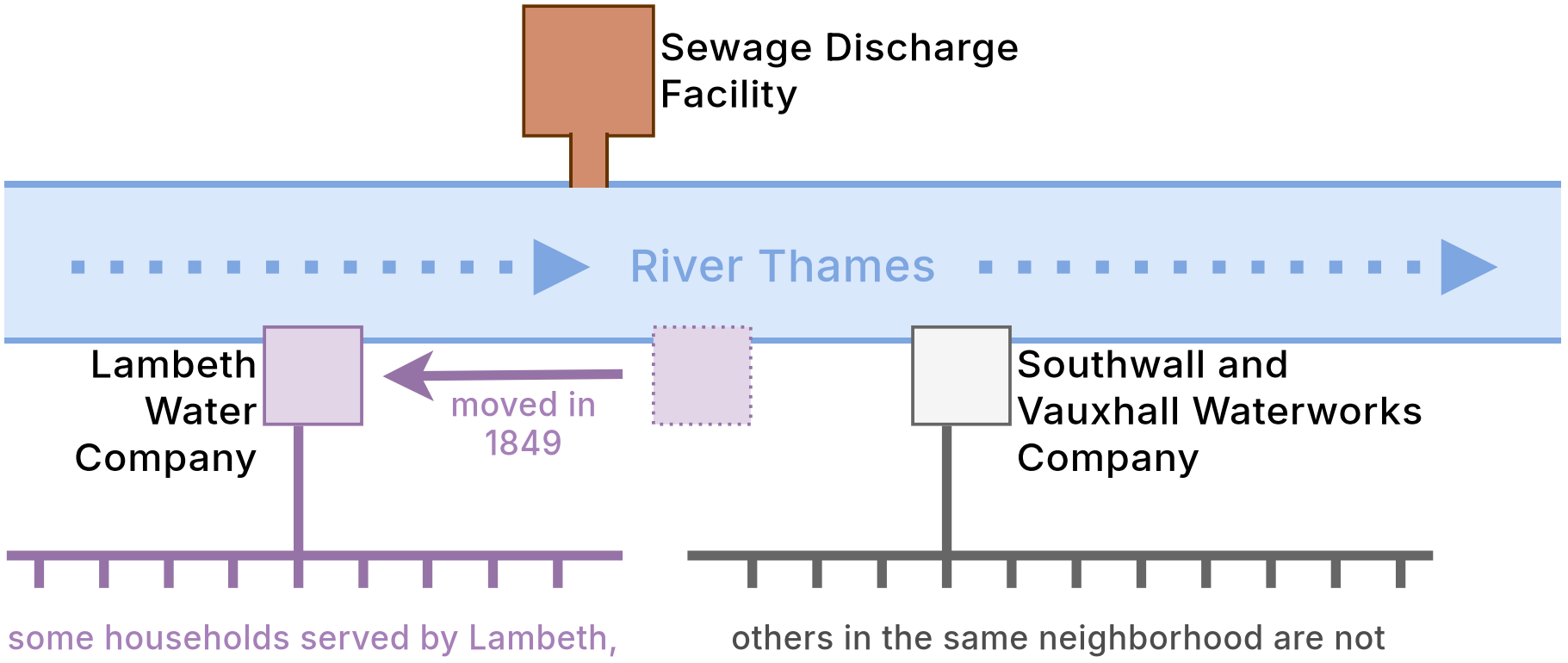
In the 1800s, London (as well as many other places) was repeatedly hit by waves of a **cholera epidemic**.

- The predominant theory at the time was that the disease was spread by small **inanimate particles** that floated through the air (which is obviously incorrect).
- **John Snow** was a physician working in London at the time, and he suspected instead that cholera was caused by **microscopic living organisms** that entered the body through water and food, multiplied in the body, and then exited the body through urine and feces.
- This would imply that **clean water supply** was a way to slow the spread of the disease.
- Unfortunately, he was only able to collect **anecdotal evidence**, which did not allow him to make a causal claim.

Of course, running **an experiment** is infeasible in this context. It would requre randomizing households, and allocating clean water to only a subset of them. This was both logistically infeasible and ethically questionable.

A Natural Experiment

In 1849, the following happened:



One water company **moved its pipes** further upstream, to a location that incidentally was upstream of the **main sewage discharge facility**. Suddenly, **households in the same neighborhoods** had access to **different qualities of water**.

A Natural Experiment

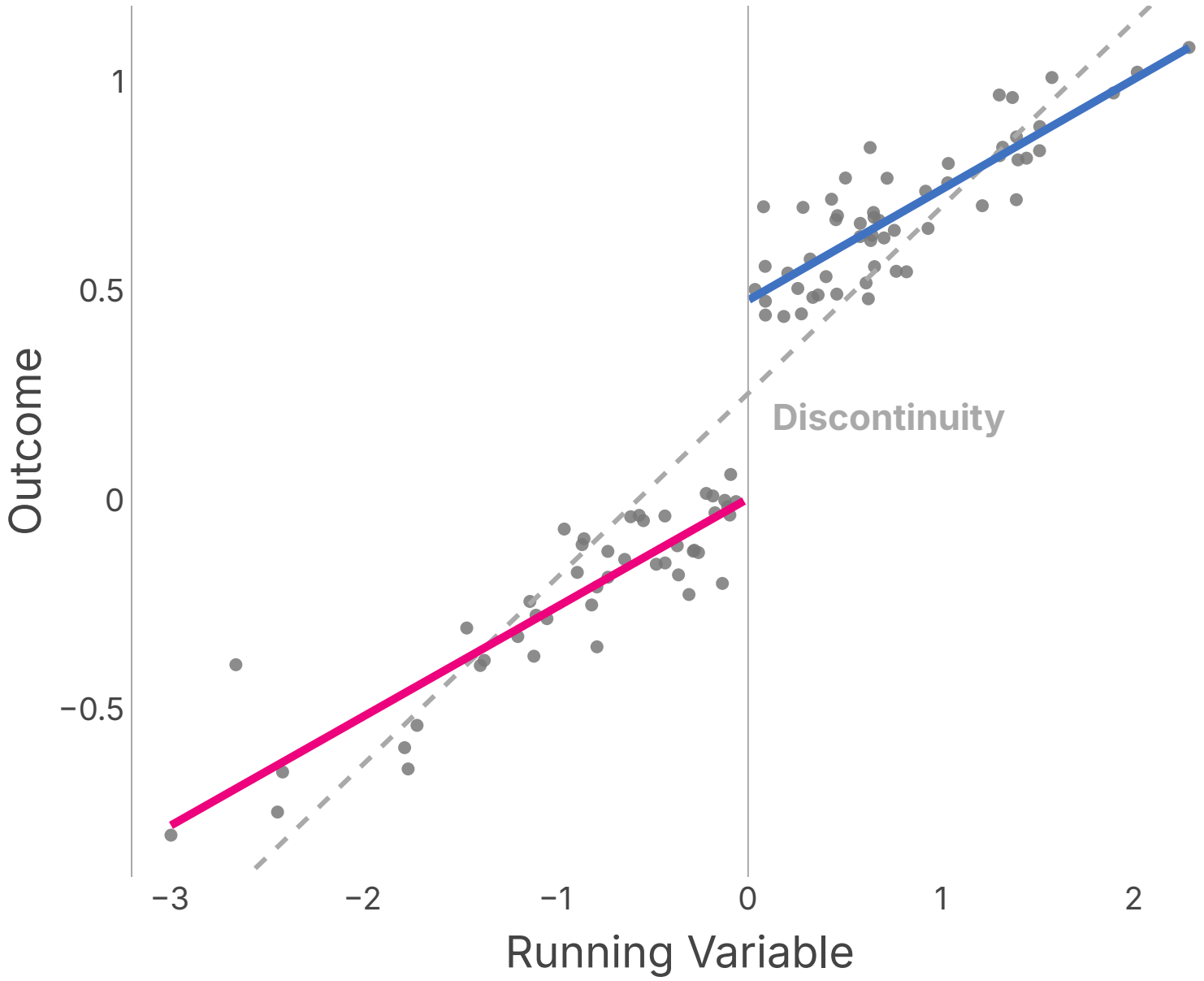
There were a few other factors that made this situation a **natural experiment**:

- Water companies were not serving disjoint geographical areas. Their networks intersected and often houses in the same street were chaotically served by pipes from different companies.
- John Snow collected extensive additional data and compared characteristics of treatment and control households to confirm their comparability.
- Most crucially, the change in water supply happened for other reasons and thus induced **exogenous variation**.

Photograph by Hisgett (2015).

In the end, John Snow collected very convincing evidence for his theory and went on to identify a certain contaminated water pump. The theory, however, was deemed politically unpleasant and was thus not accepted until long after Snow's death.

Regression Discontinuity



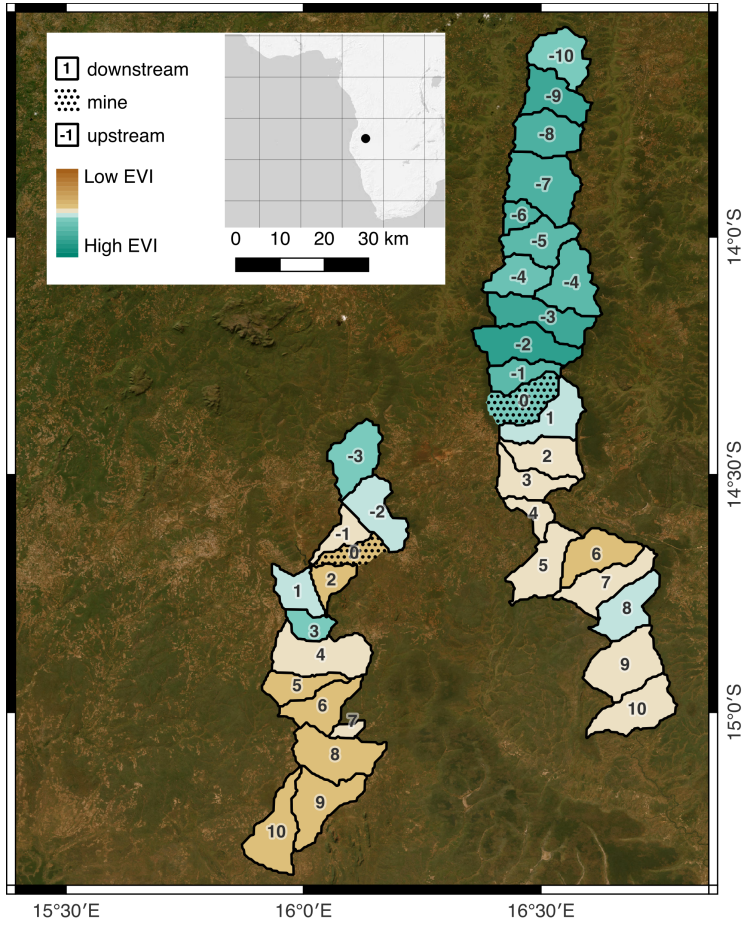
A **Regression Discontinuity Design** (**RDD**) is another type of **quasi-experimental design**.

We make use of **a sharp cutoff** in some **runnning variable** and compare values immediately below and immediately above the cutoff.

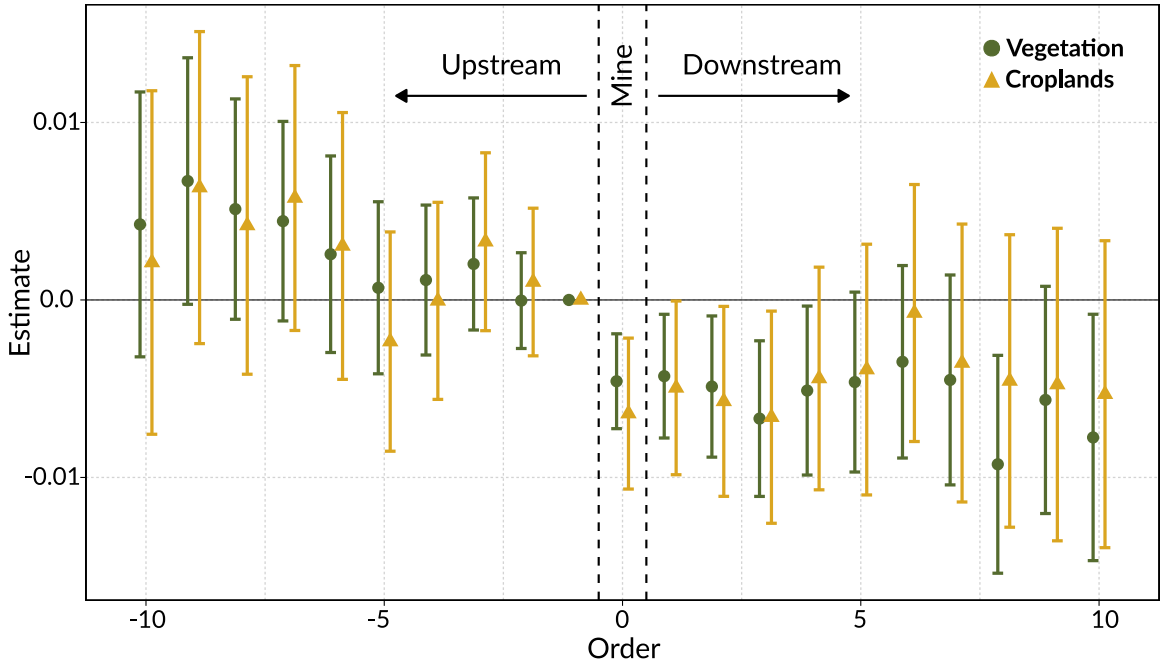
The **size** of the **discontinuity in outcomes** gives us the **local treatment effect**.

RDD Applications

RDDs are commonly used where there is some kind of artificial **cutoff**, e.g. test scores exceeding a minimal threshold for admission to a program. But they are not limited to that.



We (Vashold et al., 2026) made use of a **discontinuity in space**: Mines pollute water flows, but only in one direction. We found that vegetation is less healthy downstream of a mine.



Requirements for an RDD

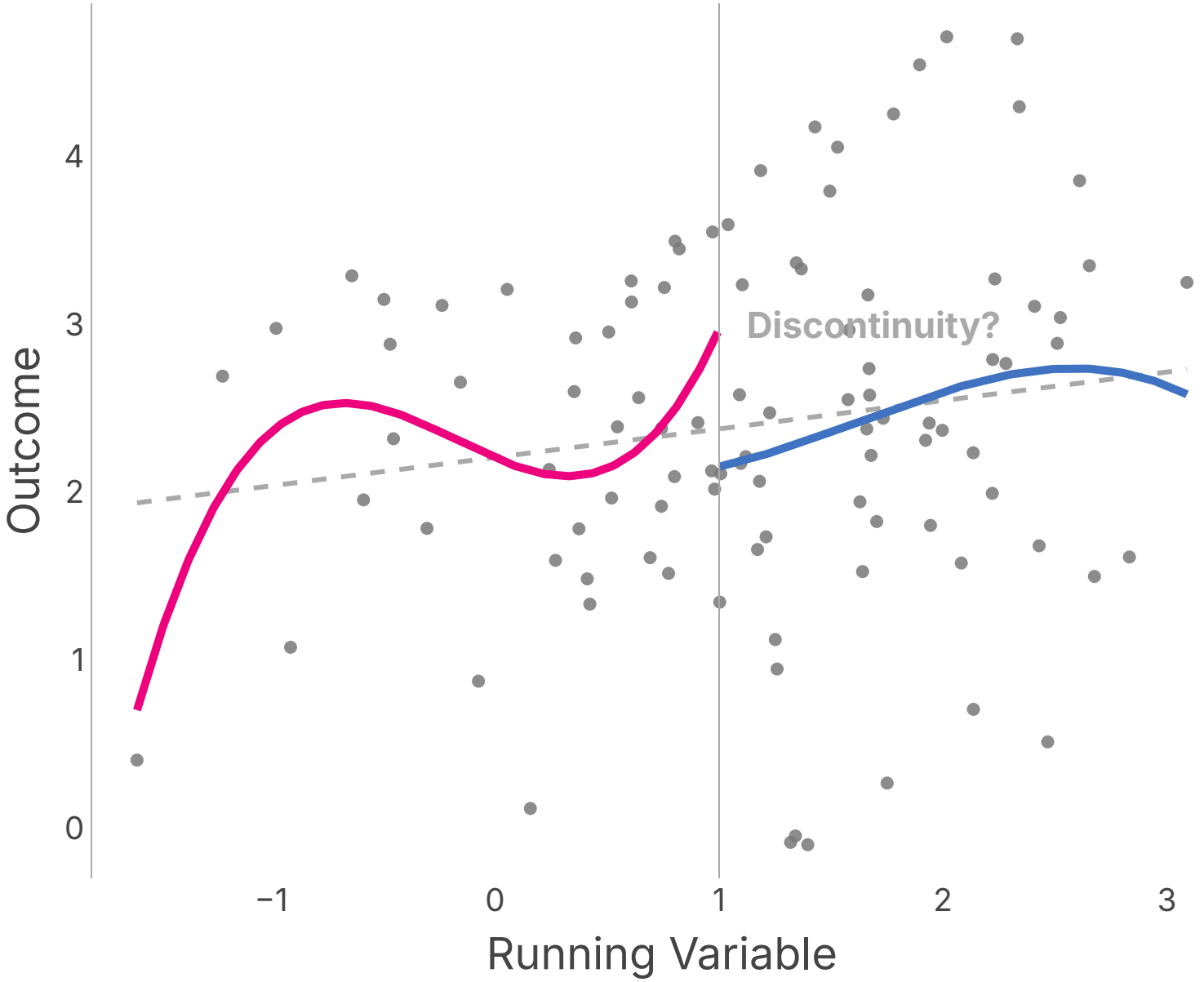
For an **ideal RDD**, we need a few things:

- All **other relevant variables** should be continuous at the cutoff, meaning that **they** do not jump.
- There needs to be **randomness** in the assignment around the cutoff. People just below and just above the threshold should be otherwise comparable.
- We also need to model the **functional form** (i.e., of the relationship between the running variable and the outcome) correctly.

In practice, these requirements are **hard to check**.

- Effects are often **contaminated** by other factors, as cutoffs often trigger multiple things simultaneously.
- We (obviously) never truly know the **functional form**.
- Treatment assignment can sometimes be **manipulated**. Think of us giving you a half-point you don't deserve in the exam to make you get the better grade.

Look, I've Found a Discontinuity



A common problem is "fabricating" a discontinuity by **overfitting** the data to both sides of the cutoff.

In the example on the left, there is obviously no discontinuity – yet we can fit something that makes one appear.

Panel Data

Quasi-Experiments

More Methods

Matching

Recall the fundamental problem of causal inference: We **cannot observe the counterfactual** to our treatment. What we can do is find specific treated observations that are very similar to other untreated observations. We call this procedure **matching**. It works like this:

- We start by dividing the dataset in **treated** and **control** units.
- Next, we find the ones whose **characteristics match best**.
- Then, we **discard unmatched observations** without creating selection bias.
- Finally, we **perform** our **analysis** with the matched dataset.

This procedure allows us to create a **sample with balanced confounders**, emulating the balance induced by completely randomized or blocked experiments.

How to Match

Propensity score matching uses the **propensity of being treated** for each observation.

- We estimate this propensity, assign a **propensity score**, and match observations with similar propensity scores. (Note that we reduce all information to one dimension.)
- The method can sometimes induce imbalance rather than remedy it.

Distance matching uses some measure of **distance** between observations.

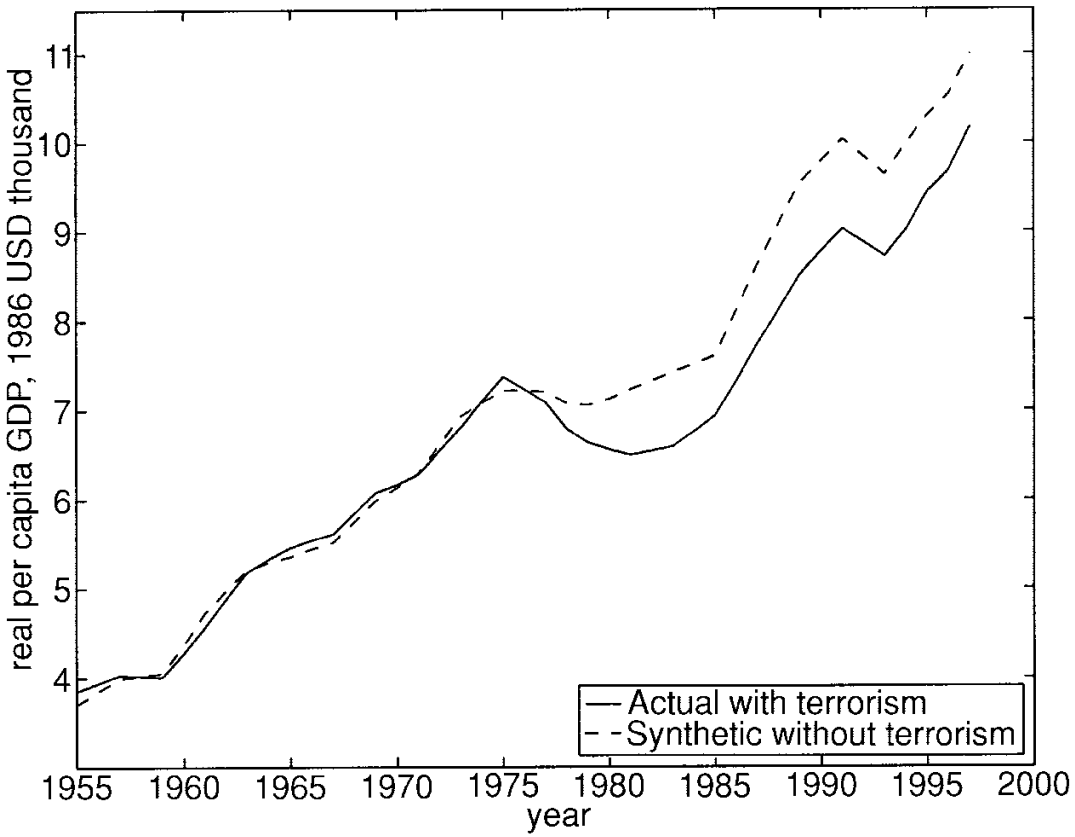
- Explaining what "distance" means in this context would require discussing Euclidean geometry.
- Intuitively, you can think distances between points in a scatterplot of covariates.

Coarsened exact matching sorts variables into different **bins**.

- First, we **coarsen** covariates, e.g. separating age values into bins.
- Then, we group observations that are in the same set of bins, and discard all sets of bins that only have treated or control observations.

Synthetic Controls

The basic idea of the **synthetic control estimator** is that we have a treated unit, and multiple untreated units that do not match the characteristics, or trajectory, of the treated unit. So, what we do is to compute a **weighted average** of untreated units that matches the treated unit (using a data-driven approach).



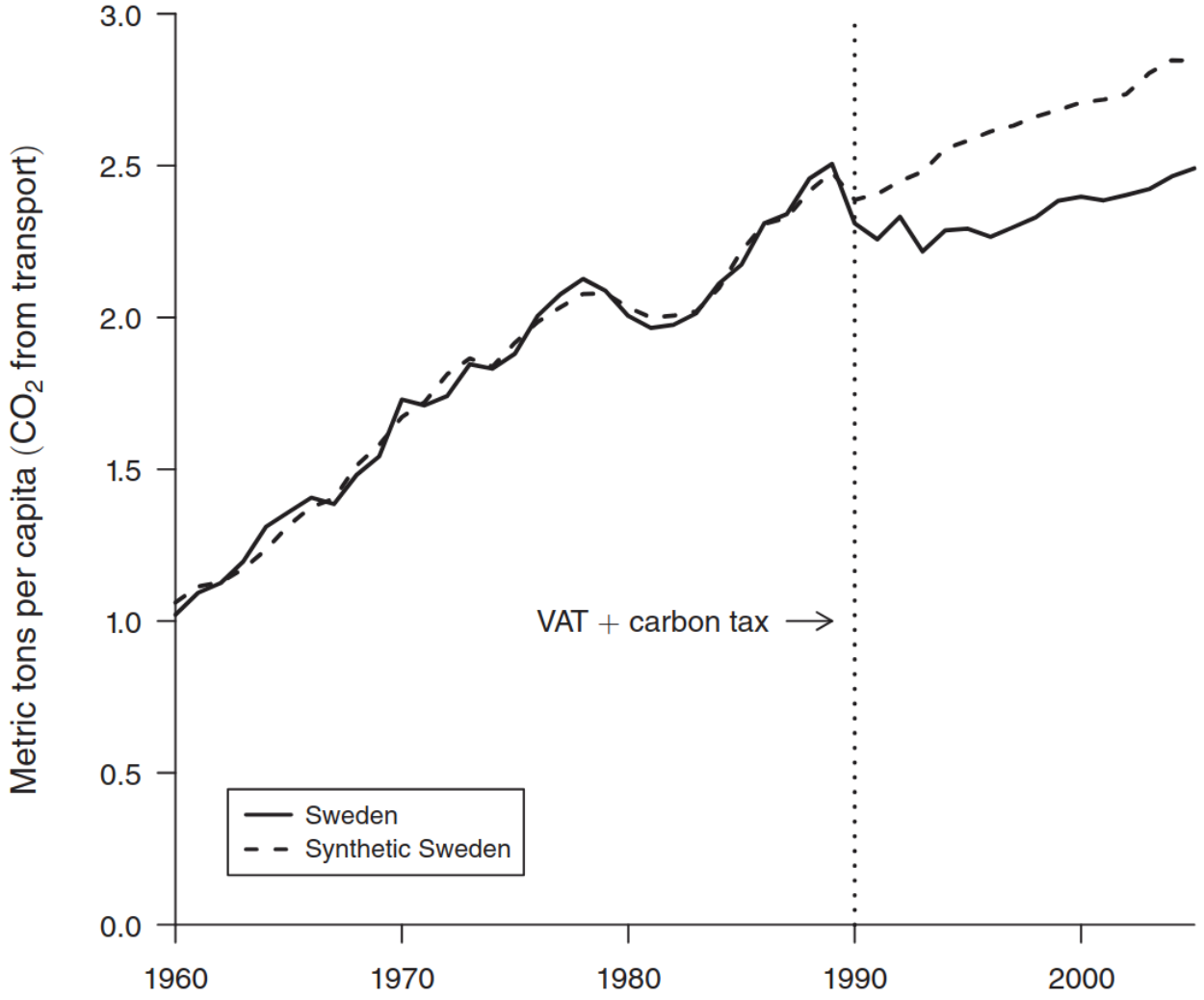
In the first study to use this design, Abadie & Gardeazabal (2003) were researching the economic cost of the terrorist activity in the Basque Country in the 1970s.

They construct the synthetic control from other Spanish regions. The mix they end up with is 85.1% Catalonia, 14.9% Madrid, and 0% of all other regions.

A Second Synthetic Control Example

Let us look at one more example. Andersson (2019) investigated the effects of a carbon tax and a fuel-specific value added tax on CO_2 emissions from Sweden's transport sector.

The synthetic control is constructed from other OECD countries. The final mix is 38.4% Denmark, 19.5% Belgium, 17.7% New Zealand, 9% Greece, 8.8% U.S., 6.1% Switzerland, and 0.1% each of Australia, Iceland, and Poland.



References

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review*, *93*(1), 113–132. https://doi.org/10.1257/000282803321455188

Andersson, J. J. (2019). Carbon taxes and CO2 emissions: Sweden as a case study. *American Economic Journal: Economic Policy*, *11*(4), 1–30. https://doi.org/10.1257/pol.20170144

Cunningham, S. (2021). *Causal inference*. Yale University Press. https://doi.org/10.12987/9780300255881

Hisgett, T. (2015). *Dr john snow*. Flickr. https://flickr.com/photos/37804979@N00/24023399742

Vashold, L., Pirich, G., Heinze, M., & Kuschnig, N. (2026). Downstream impacts of mines on agriculture in africa. *Journal of Development Economics*, *179*, 103671. https://doi.org/10.1016/j.jdeveco.2025.103671