When A/B testing isn’t an option

An introduction to quasi-experimental methods

Inga Jańczuk (OLX)

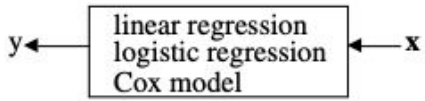
PyConDE & PyData Berlin 2023

Statistical Modeling: Two Cultures

“There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown.”

L. Breiman (StatSci, 2001)

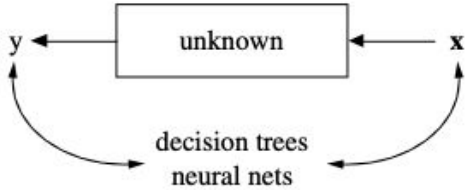
Data modeling culture



1. Theory driven.
2. Allows to identify quantities of interest, e.g. treatment effects.
3. Usually not very flexible.

Typical application: **causal inference**

Algorithmic modeling culture

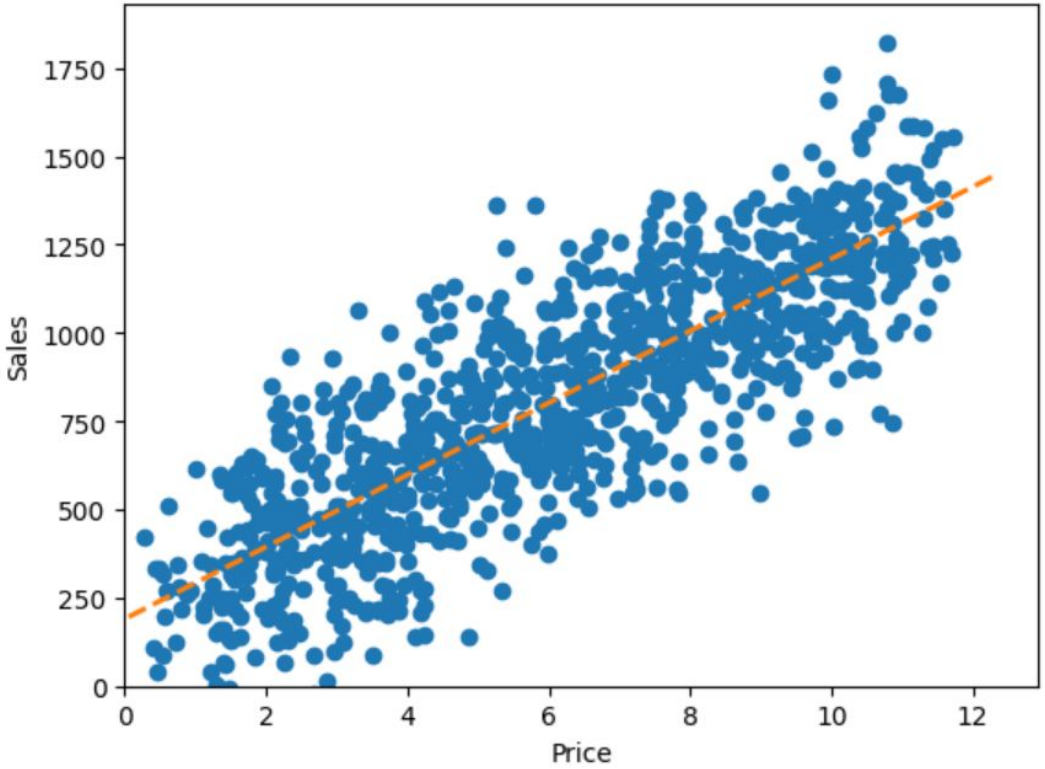


1. Theory agnostic, data driven.
2. Focuses on prediction power.
3. Ignores non-predictive considerations like causality, equilibrium, feedback effects.

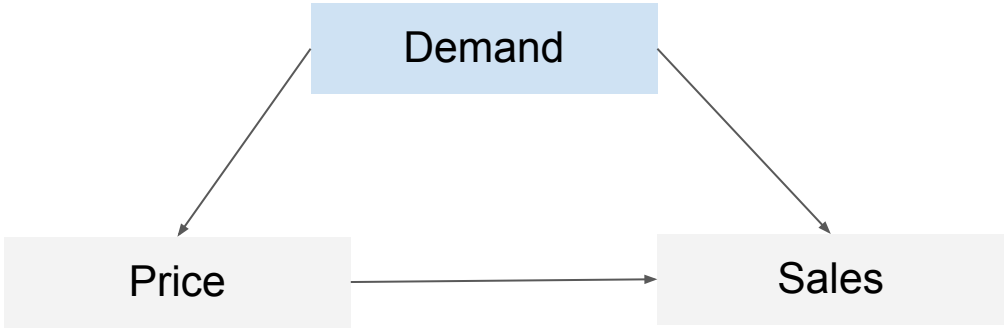
Typical application: **prediction problems**

Diagrams from Breiman, 2001

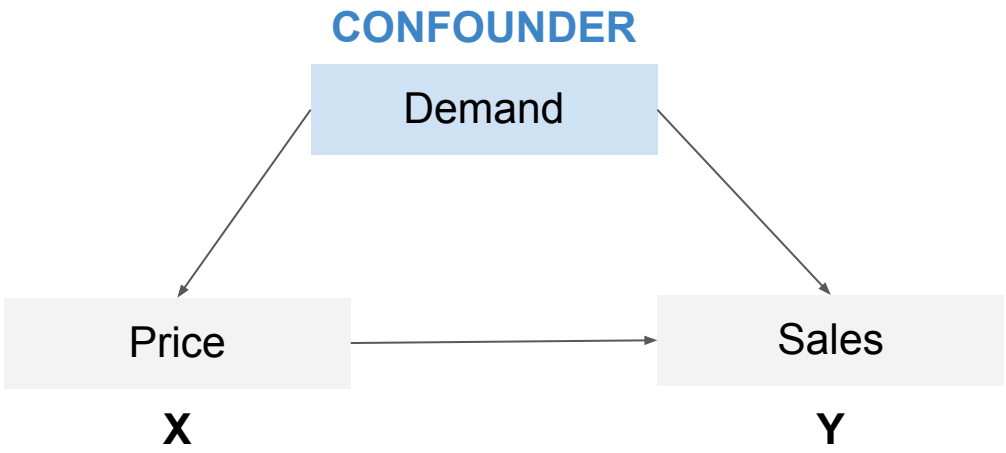
Prices and Quantities



Why the model is wrong?



Why the model is wrong?



Price optimization requires counterfactual prediction:

What if I hold everything fixed and then change the price?

Solution A: **a structural model**
Solution B: **a randomized trial**

Experimentation is the gold standard for measuring the effect of an action.

A/B Testing

also known as *completely randomized design*

An experimentation technique in which subjects are **randomly** assigned to **control** and **treatment** groups.

The random assignment process performs an important function. It removes confounding factors.

This so-called potential outcome framework allows to estimate the *average treatment effect* (ATE):

$$ATE = E[y|d = 1] - E[y|d = 0]$$

where *y* is the response variable and *d* is indicator of treatment.

However…

Experiments can be costly, long-lasting, unethical, or illegal.

In other cases, the underlying assumptions for identification cannot be met due to

imperfect randomization

or dependency between treated subjects.

Quasi-experimental setup

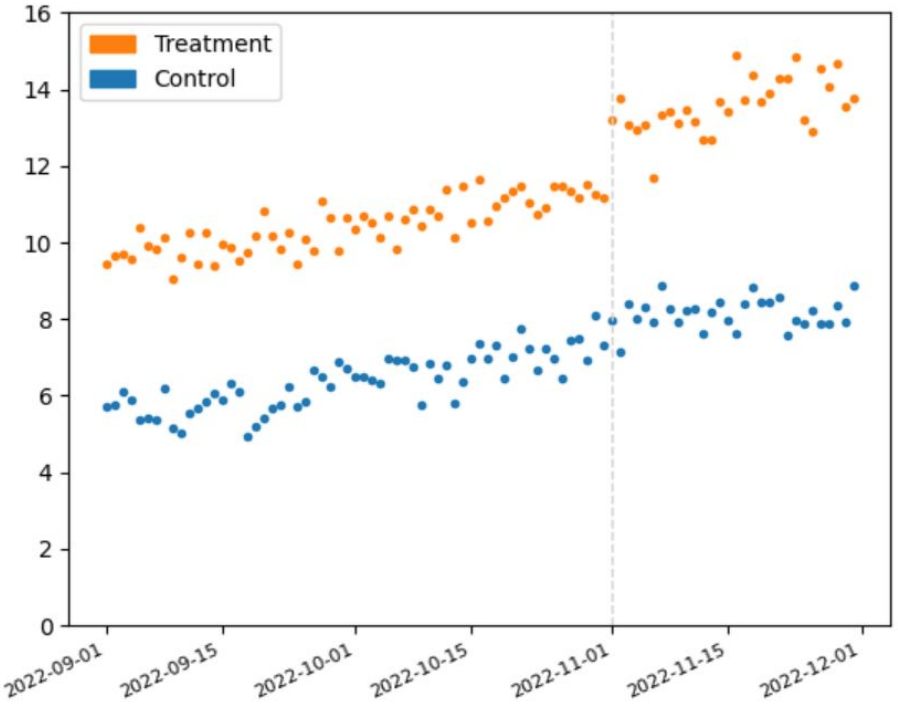
also known as *near-experimental setup*

Quasi-experiments are used to inspect past events that resemble laboratory experiments. They allow to discover treatment effects when randomized experiments are not feasible (e.g., ethical reasons, they are too costly).

Treatment effects in near-experimental setups can be uncovered under a small set of **additional assumptions** which define **identification strategy**.

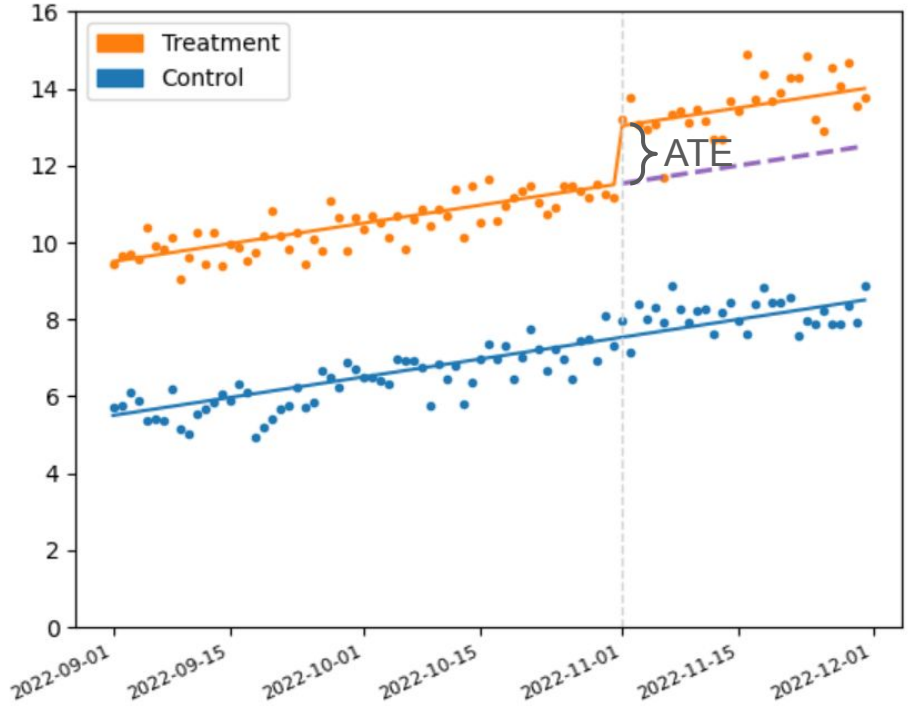
Difference-in-Differences

also known as *parallel trends design*



Difference-in-Differences

also known as *parallel trends design*



Comparing the set of units where the event happened (treatment group) in relation to units where the event did not happen (control group).

Assumptions:
The two groups share the same trend before the treatment happens ➡
if the event never happens, the differences between treatment and control groups should stay the same overtime.

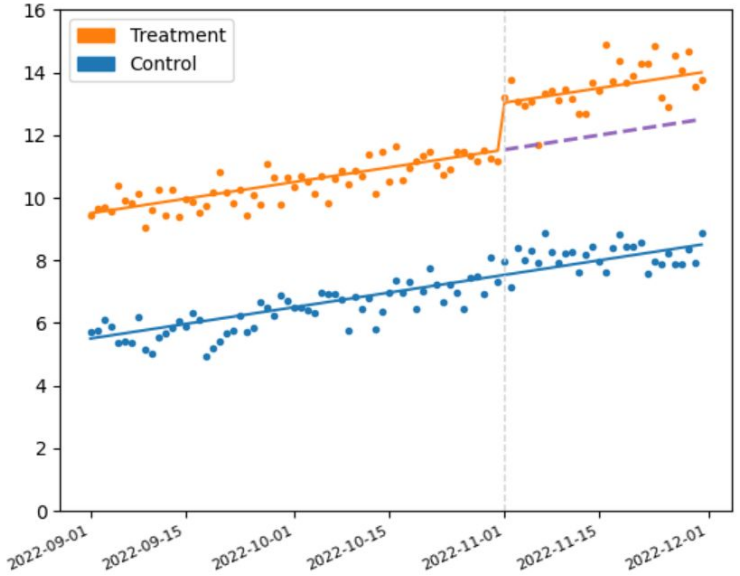
The diff-in-diff regression is given by

$$E[y_{it}] = \alpha + \beta_d * d_i + \beta_t * t + \gamma d_i * t$$

where
$d \in \{0, 1\}$ encodes the treated group of a unit i, with
d = 0 for the control group,
d = 1 for the treatment group,
$t \in \{0, 1\}$ denotes time, with
t = 0 for periods before the treatment,
t = 1 for periods after the treatment.

The coefficient of interest is γ - the coefficient on the *interaction* between d_i and t.

Difference-in-Differences

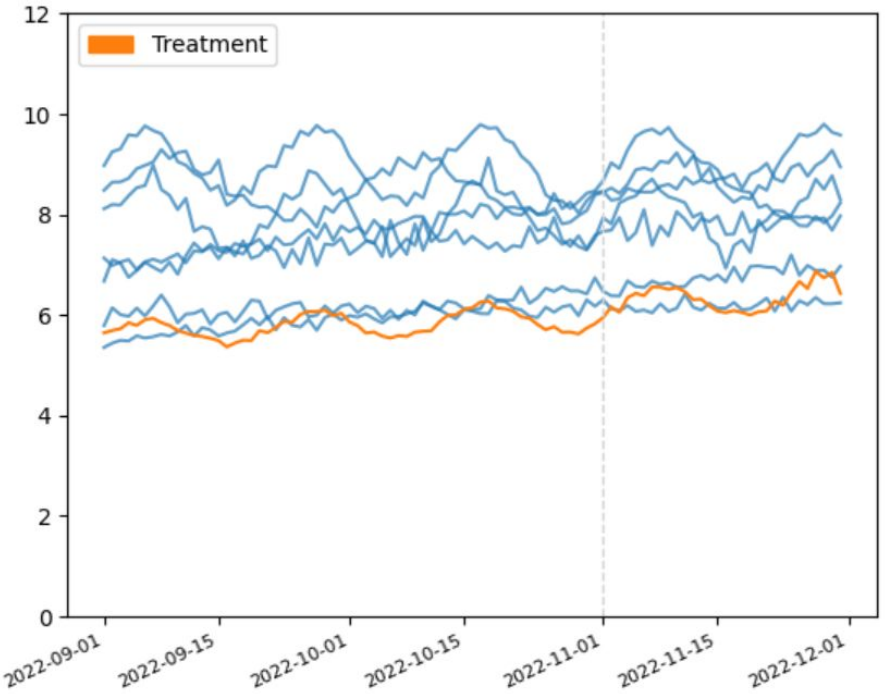


	date	group	value	t	d
34	2022-10-05	control	7.426192	0	0
88	2022-11-28	control	8.967618	1	0
163	2022-11-12	treatment	12.862654	1	1
168	2022-11-17	treatment	13.621904	1	1
32	2022-10-03	control	6.370706	0	0
20	2022-09-21	control	6.356972	0	0
169	2022-11-18	treatment	12.451310	1	1

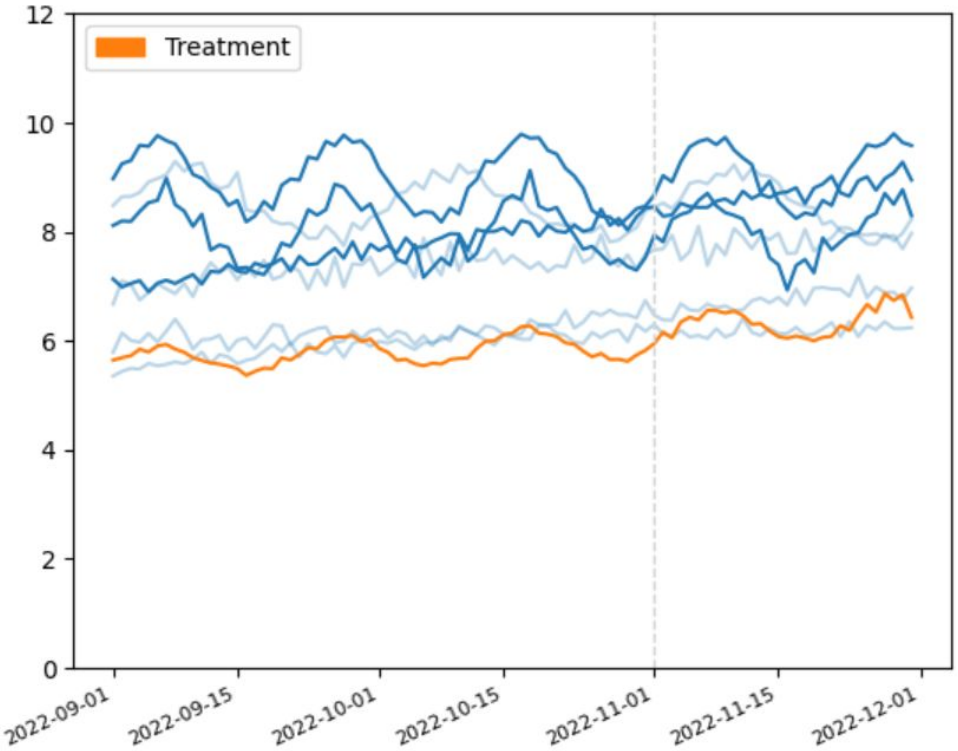
	coef	std err	t	P>|t|
Intercept	6.4420	0.081	79.718	0.000
d	3.8957	0.114	34.089	0.000
t	1.5567	0.141	11.061	0.000
d:t	1.6121	0.199	8.100	0.000

Synthetic control

also known as the *most important development in program evaluation in the last decade (Athey and Imbens, 2016)*



Synthetic control

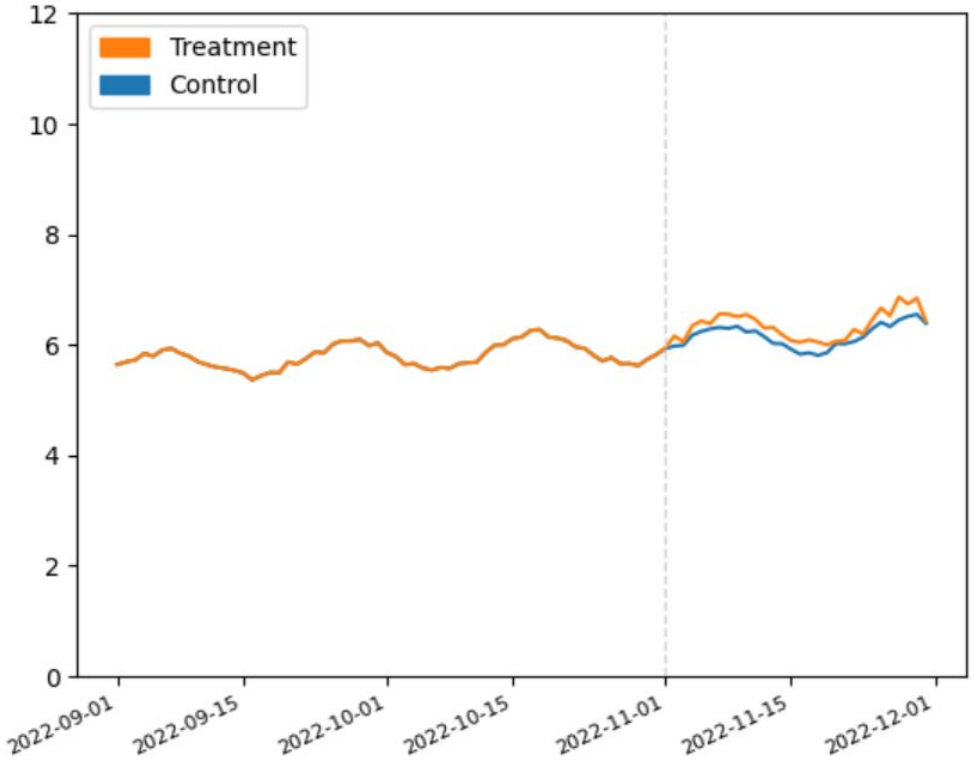


It involves construction of a weighted combination of groups used as controls, to which the treatment group is compared.

We want to create a forged control that as closely as possible resembles the treated group *before* the treatment.

But… how to build the synthetic control group?

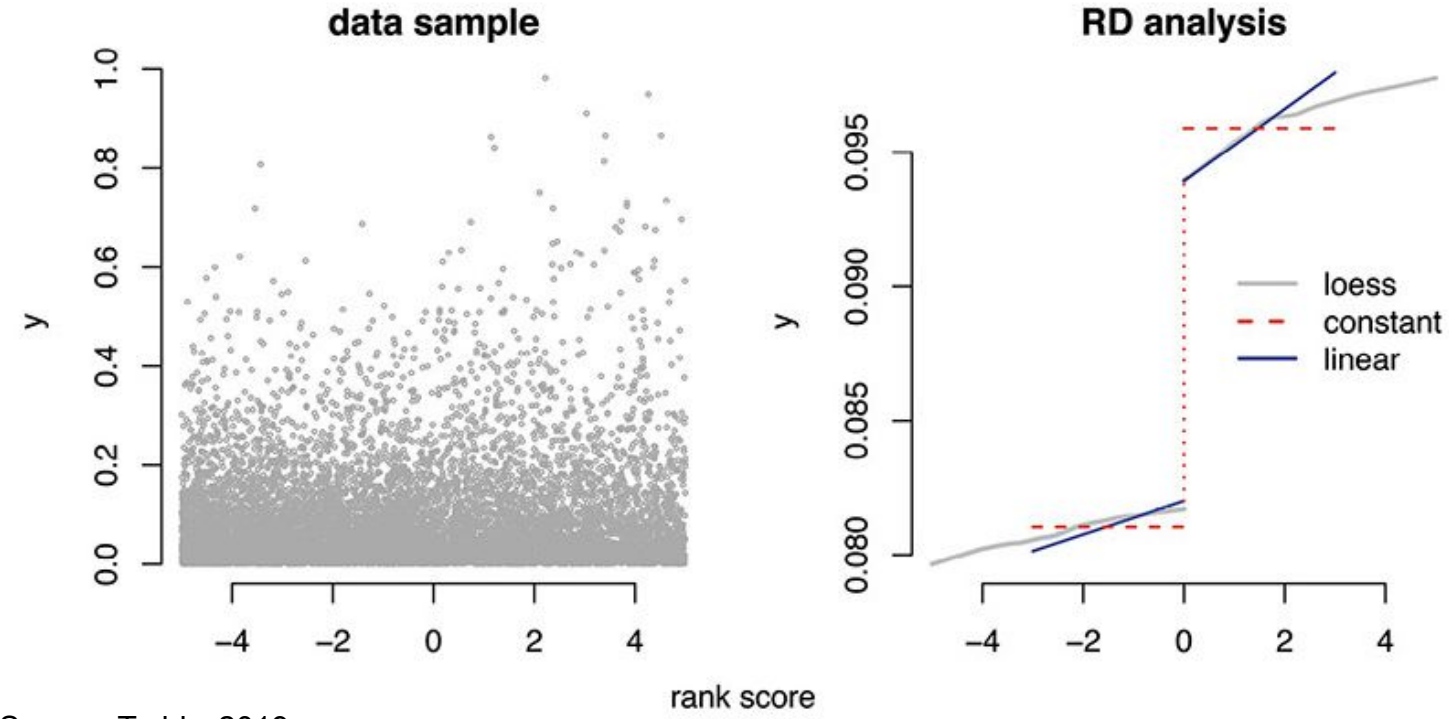
Synthetic control



The synthetic control group follows closely the control group in the pre-treatment periods.

In the post-treatment periods, the difference between the synthetic control group and the treatment group is equal to the treatment effect.

Regression discontinuity



Source: Taddy, 2019

Regression discontinuity

1. The allocation of treatment is dictated by a **running variable**.
2. Individuals near the threshold, regardless of their position on either side, can be considered equivalent for the purpose of estimating causal effects.
3. Continuity assumption - if the threshold were slightly changed, individuals switching treatment groups would behave similarly to those near them in their new treatment group.
4. The estimated treatment effects are identified only near the threshold - they are **local ATEs**.

Takeaways

1. Prediction and causality are not necessarily the same. Correlation does not imply causation.
2. Counterfactuals (rather than predictions) are critical for policy makers.
3. A/B testing is a gold standard for identifying causal effects.
4. When A/B testing cannot be used, alternative quasi-experimental setups can be useful.
5. Quasi-experiments allow to make use of the past events that resemble laboratory experiments.

Thank you