Synthetic controls

Brief introduction and intuitions

Sebastian Daza

Outline

- Very short intro to **diff-in-diff** estimates
- Synthetic **control**
- Synthetic **diff-in-diff**
- Geo-experiments

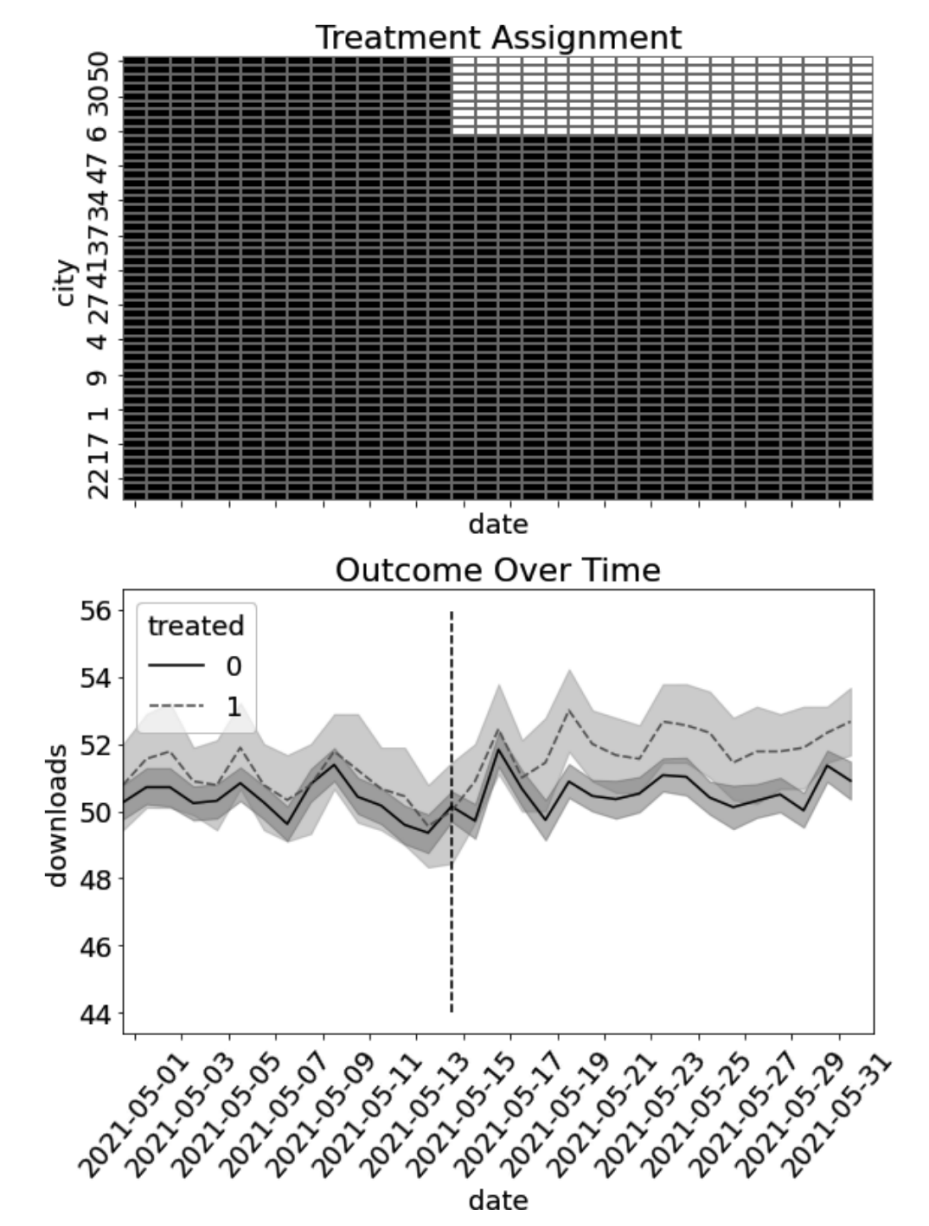
Diff-in-Diff (DID) 😎

- It's one of the oldest tools, first used in 1855 by **John Snow** in his analysis of the cause of **cholera**
- Goal: Identifying the causal effects when randomization is not possible
    - We have same units in **multiple times** so we can set what happens before and after a **treatment** takes place

Diff-in-Diff (DID): Example

Geo-experiment: Treat entire markets, such as a city or a state, while leaving others as control.

Diff-in-Diff (DID): Block design



Diff-in-Diff (DID) 😎

Let's assume we have theses data:

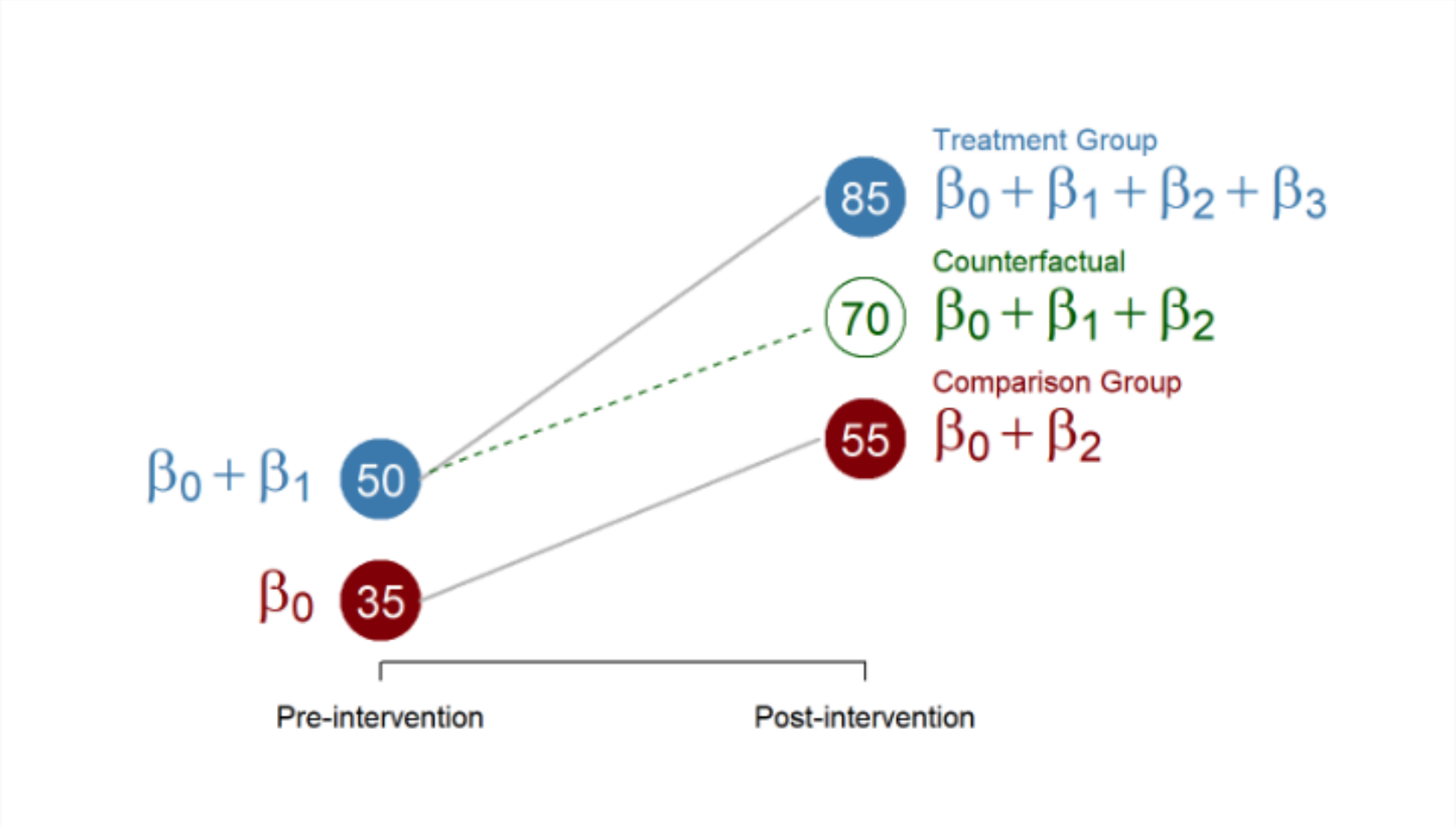
		downloads	date
treated	**post**		
0	0	50.335034	2021-05-01
	1	50.556878	2021-05-15
1	0	50.944444	2021-05-01
	1	51.858025	2021-05-15

- The DID estimate will be:

$$(51.85 - 50.94) - (50.55 - 50.33) = 0.69$$

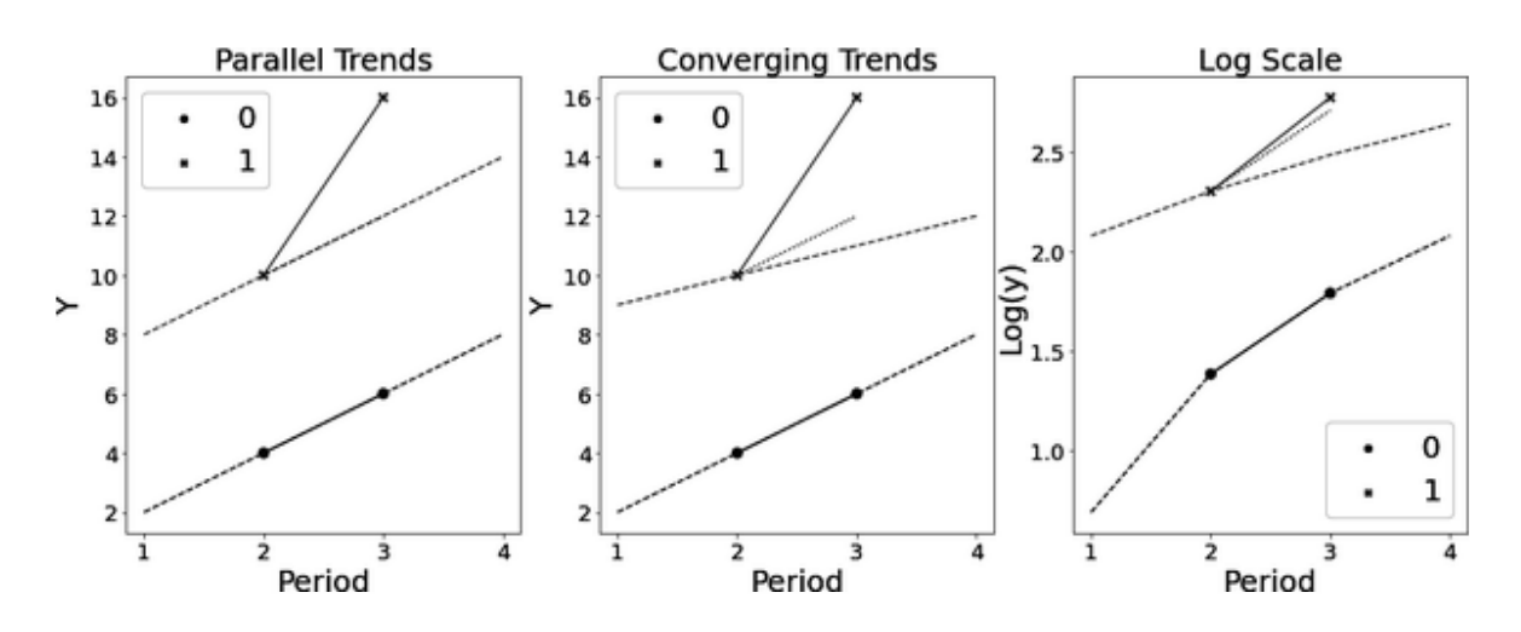
- DID measures the impact of a campaign on a city. It represents the Average Treatment Effect for the Treated (ATT), quantifying the campaign's effect post-launch
- With disaggregated data, we can use regression and adjusting for pre/post and city (fixed-effects)

Diff-in-Diff (DID) 😎



$$Y_{it} = \beta_0 + \beta_1 D_i + \beta_2 Post_t + \beta_3 D_i Post + e_{it}$$

DID assumptions 😬



- Parallel trends
- No anticipation assumption (SUTVA)
- Strict exogeneity
  - No time varying confounders
  - Assignment to treatment isn't based on future potential outcome trajectories
  - No carryover effect

Synthetic control (SC) 💊

- DID works great if you have a large number of units N compared to time periods T
- SC works fine with very few treatment units
  - Combine the control units to create a synthetic control that approximate the behavior of treated units in the absence of treatment
  - If that approximation is good and generalizes into the post-intervention period, we can estimate ATT
  - We can relax the parallel trend assumption

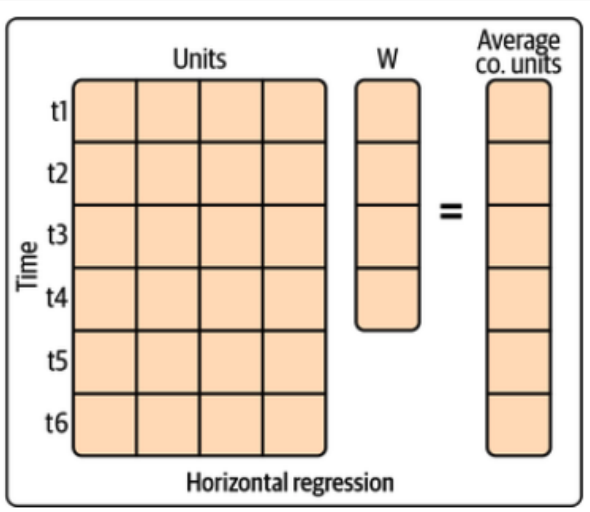
Synthetic control (SC) 💊

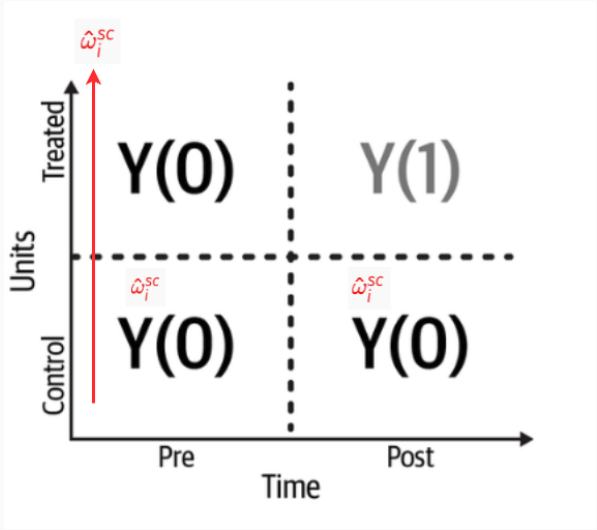
Goal: Combine the control units to approximate the average outcome for the treated units using only the pre-treatment period:

$$\hat{\omega}^{SC} = \arg\min_{\omega} ||\bar{y}_{pre,tr} - Y_{pre,co}\omega_{co}||^2$$

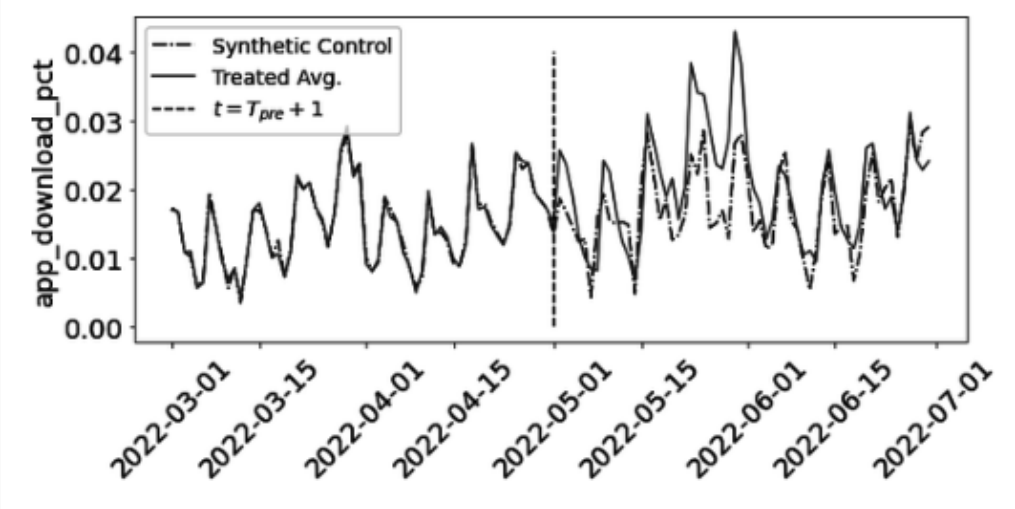
To make it simple, we can reframe the problem in linear regression terms:

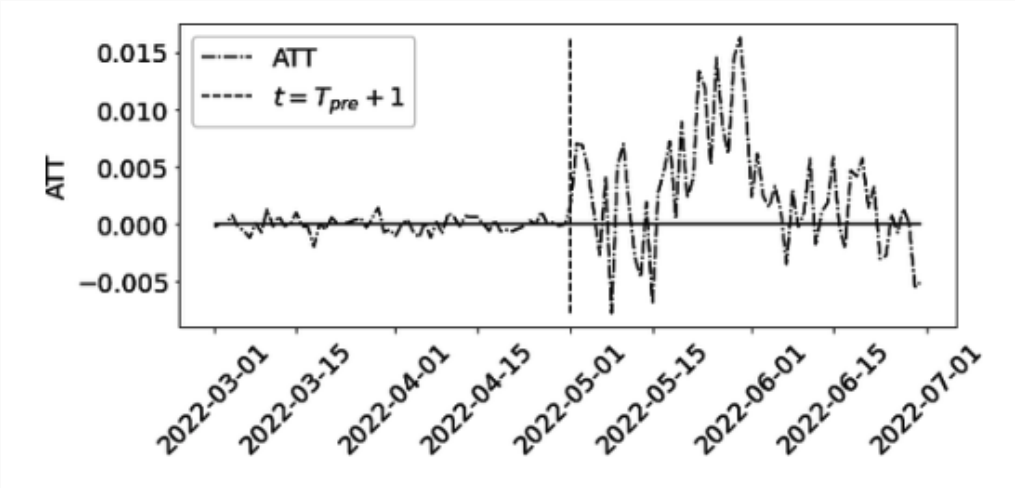
$$\beta^* = \arg\min_{\beta} ||Y_i - X_i'\beta||^2$$



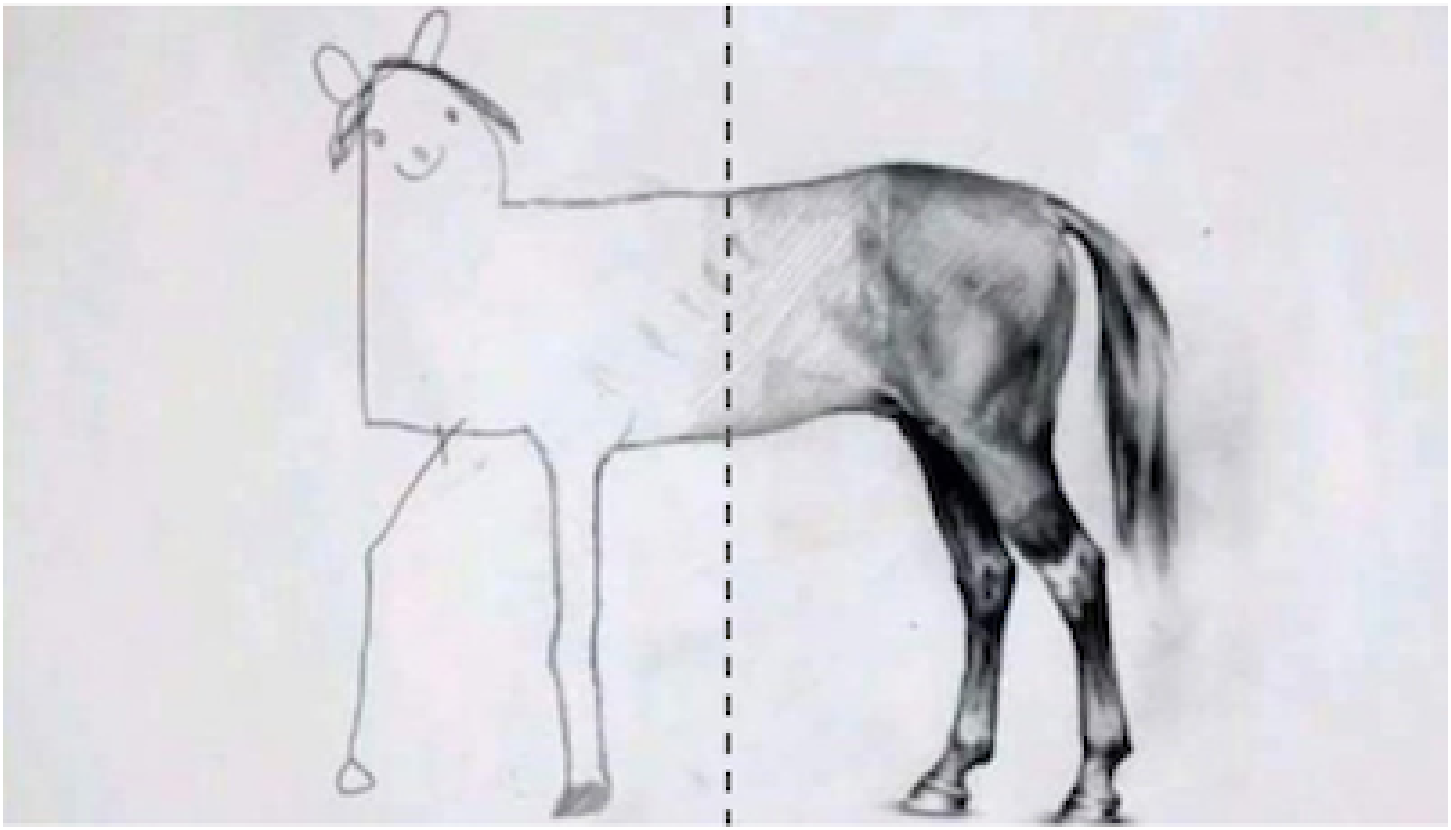


Synth Control: Regression approach





Synth Control: Regression approach

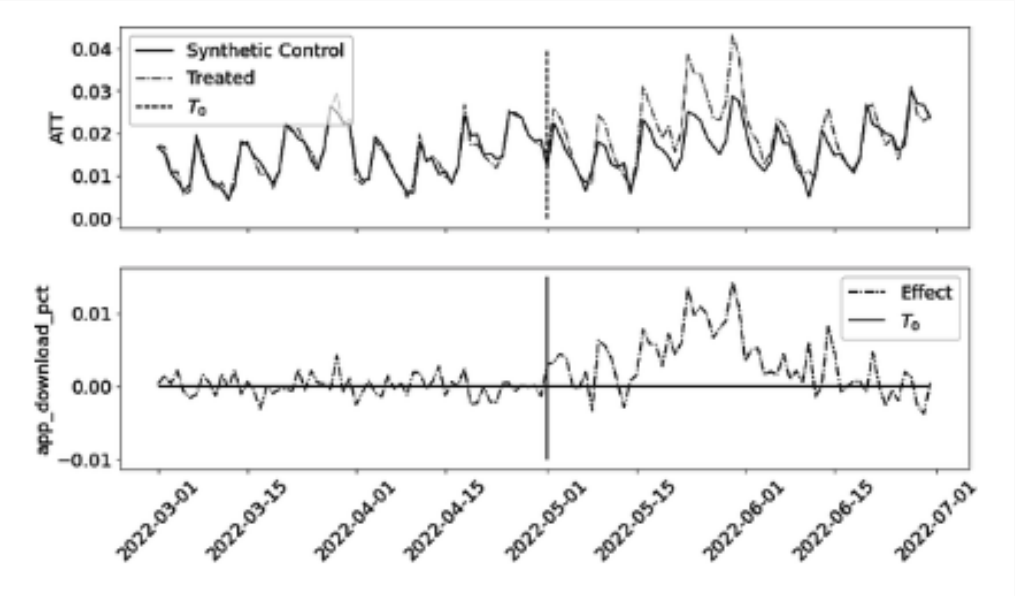


Synth Control: Canonical approach

$$\hat{\omega}^{SC} = \arg\min_{\omega} ||\bar{y}_{pre,tr} - Y_{pre,co}\omega_{co}||^2$$

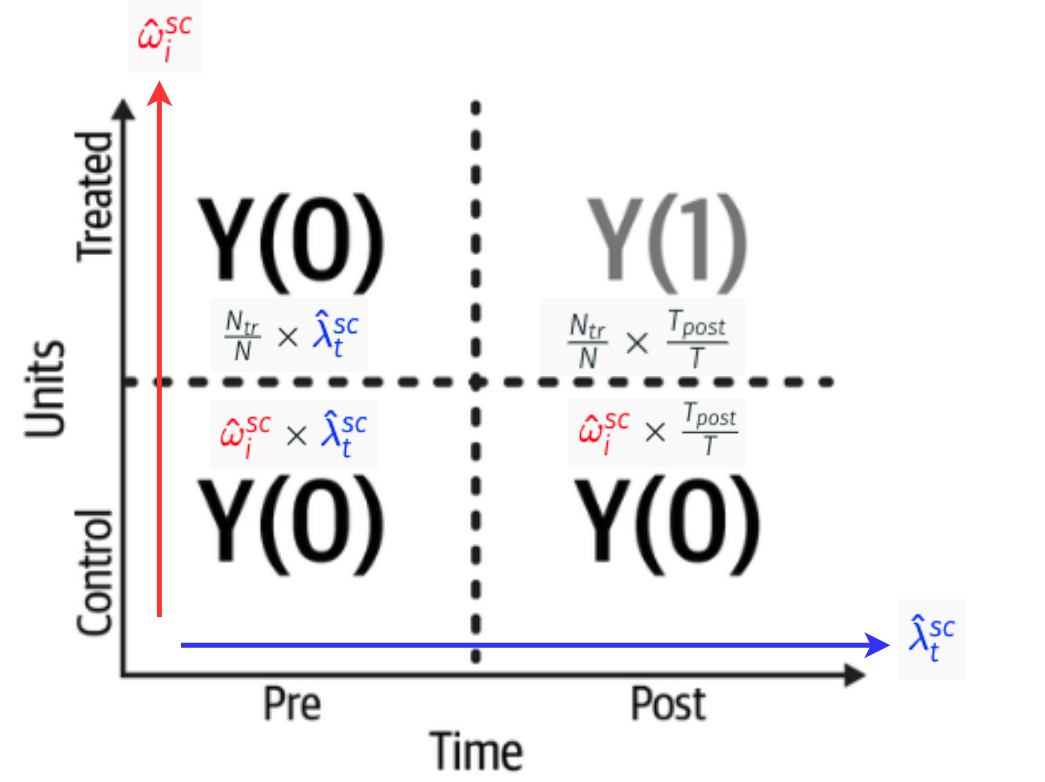
Where $\sum \omega_i = 1$ and $\omega_i > 0 \quad \forall \quad i$

- Avoid extrapolation
  - If the treatments units are very different from the ones in the control group you shouldn't even try
- Regularization



Synthetic Diff-in-Diff (SDID) 💊[2]

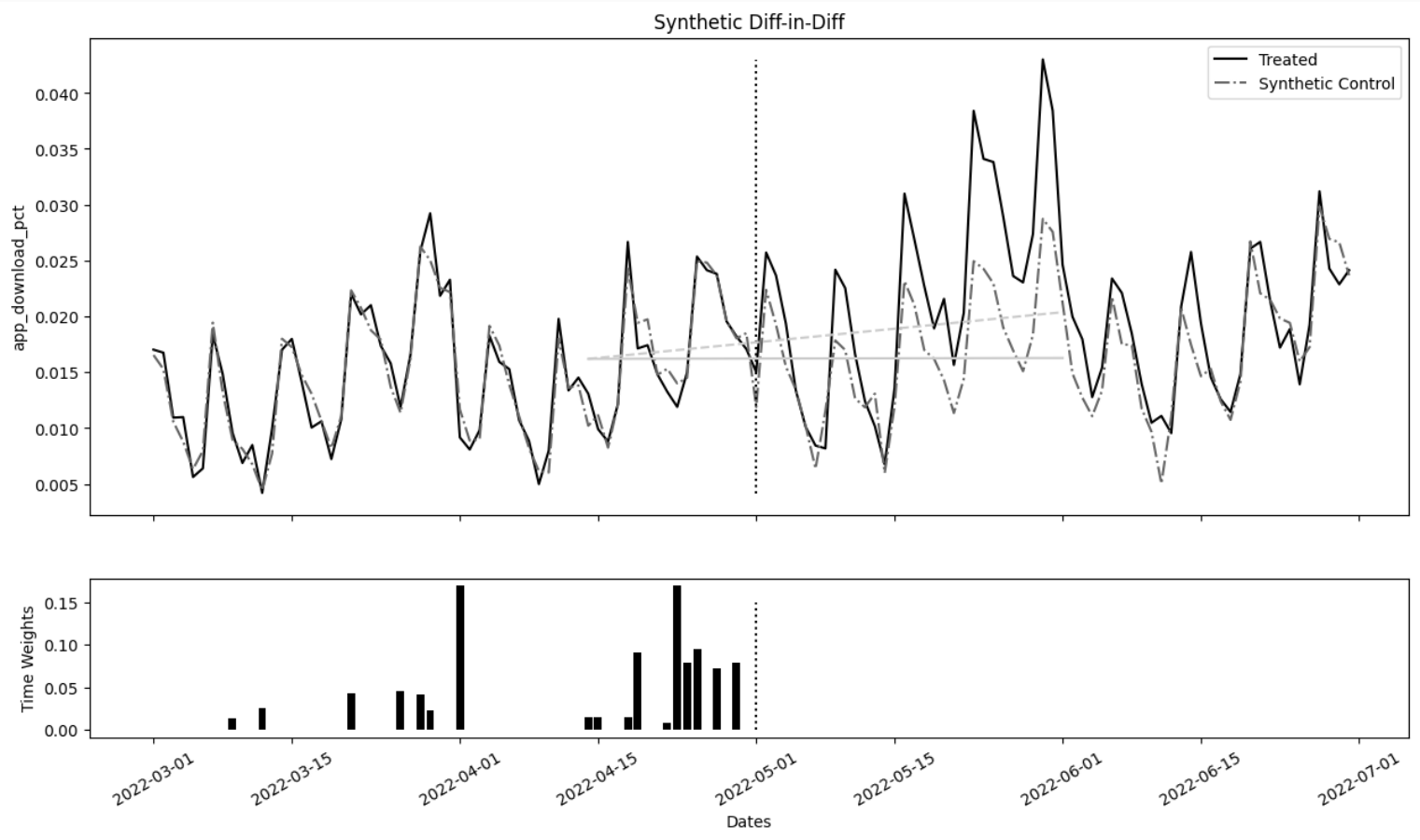
- First, we create a synthetic control
  - Unit weights $\hat{\omega}_i^{sc}$ (pre-treatment for control and treated)
- Second, we create time weights $\hat{\lambda}_t^{sc}$ (pre and post treatment only for control)
- Then, we multiple weights



Synthetic Diff-in-Diff (SDID) 💊²

Finally, we estimate DID using **weighted least squares** (WLS) regression:

$$\min \sum_i w_i(y_i - \beta_0 - \beta_1 treated_i - \beta_2 post_i - \beta_3 treated_i \times post_i)^2$$

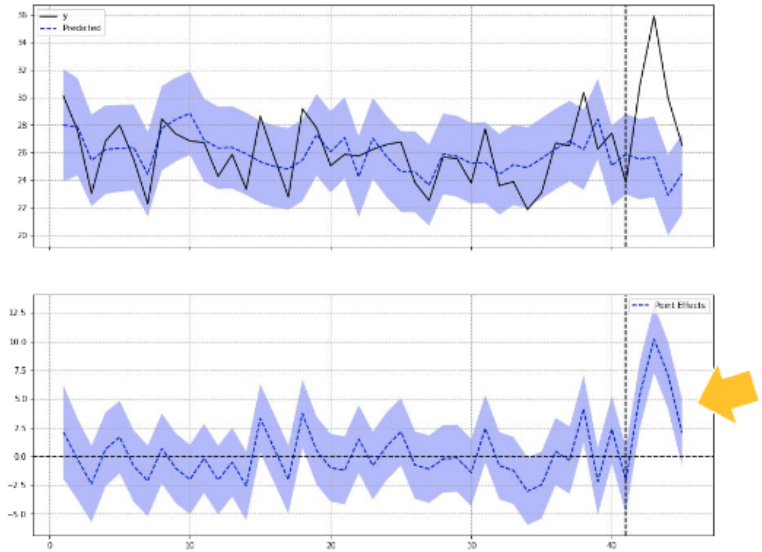


Synthetic Diff-in-Diff (SDID): Example

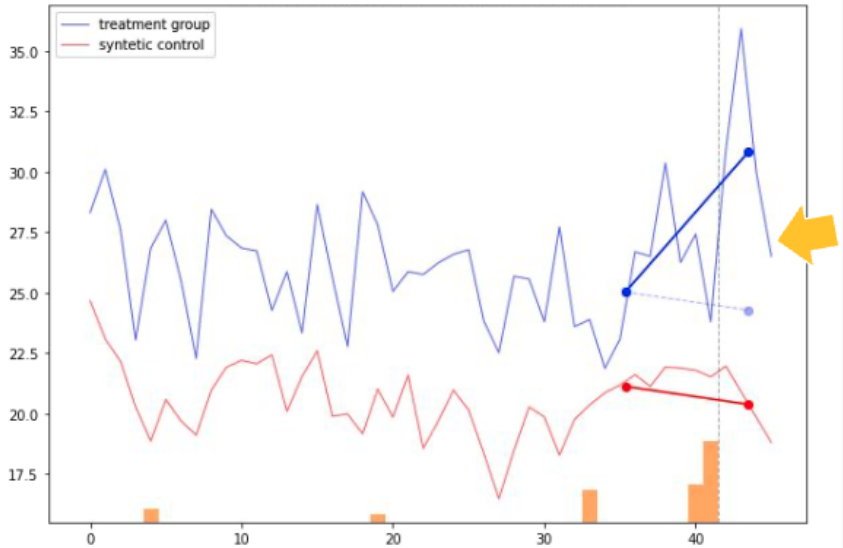
BSTS: 6.18 [4.75, 7.62]
DID: 5.32 [1.63, 9.00]
SDID: 6.55 [3.03, 10.06]

Contribution margin
estimated effects for Belgium

BSTS



SDID



Synthetic Diff-in-Diff (SDID) 💊[2]

- The SC makes the DID's parallel trend assumption more plausible
- SDID tends to have lower bias than DID and SC
- SDID tends to have lower variance

Geo-experiments

- We can treat entire markets, such as a city or a state, while leaving others as control
- This approach provide us with panel data, but as given
- How to best select the treated and control markets?

Geo-experiments

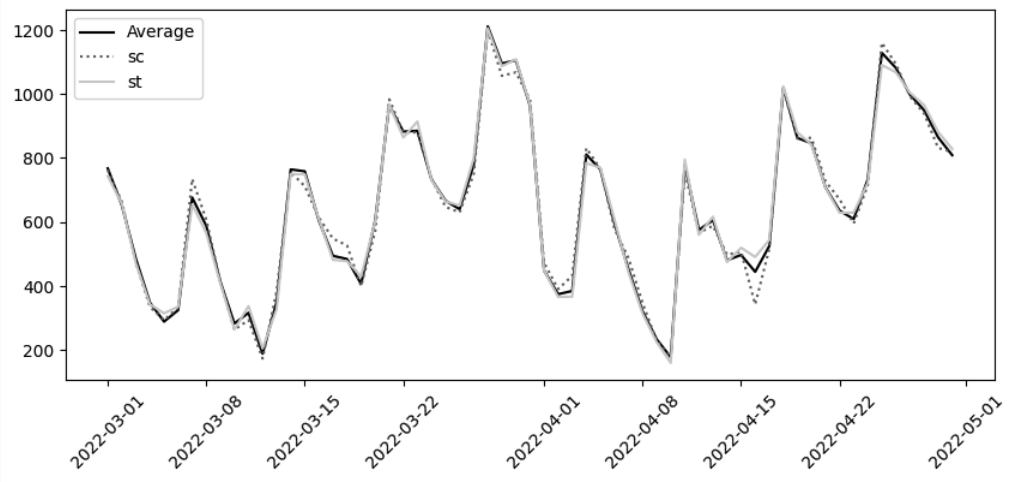
Goal: Select a group of cities (treatment and control) that is representative of the total market

That way, if you treat that group, you will get an idea of what would happen if the entire market (i.e., the country) is treated
$\sim$ maximize the external validity of the intervention

- Simple A/B estimating if we have lots of geographical units $\sim$ power analysis!
- Synthetic control design

Geo-experiments: Simple idea

- Define a number of treatment units m
- Iterate over random combinations of cities m + remaining cities
- Minimize the `loss` from the Synthetic Cohort function for each set of cities
- Retrieve the cities in control and treatment



Resources

- SDID Python: https://github.com/MasaAsami/pysynthdid
- SDID R: https://github.com/synth-inference/synthdid/
- Google causal impact (BSTS): https://github.com/WillianFuks/tfcausalimpact
- Chapter 8 and 10 from the book Causal Inference in Python