Difference-in-differences

Filippo Biscarini

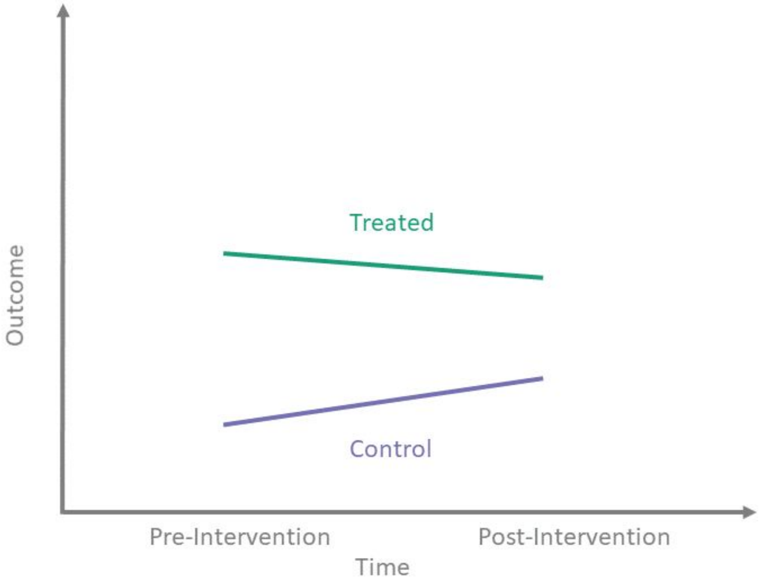
(Biostatistician, bioinformatician and quantitative geneticist) **CNR**, Milan (Italy)

Motivation and characteristics

diff-in-diff

- in some (many) settings, RCT (randomized “clinical” trials) are not possible
- we need to use observational data or quasi-experimental settings
- e.g. policies in economics, politics, public health etc. (also Mendelian randomization)

Motivation and characteristics



[From: https://diff.healthpolicydatascience.org/]

- for diff-in-diff, we need observations on subjects exposed (treated) and not (control) to the intervention, both before and after the intervention
- **treatment** and **time** components (sounds familiar? ;-))
- **difference before and after intervention (treatment)**
- compared with before-after difference in the **control group** (no treatment) → corrects for trend (differences due to other reasons)

Diff-in-diff: calculations

diff-in-diff = (treatment_post - treatment_pre) - (control_post - control_pre)

	pre	**post**	**diff**
treated	70	83	13
control	68	76	8
diff	2	7	**5**

diff-in-diff

Example*:

- survival of cancer patients (expected life span)
- treatment: latest-generation cancer treatments
- control: increased lifespan due to increased quality of life in the general population

*artificial example

Diff-in-diff: statistical model

$$
\begin{aligned}
y = \mu + \beta_1 \,\text{treatment} + \beta_2 \,\text{time} \\
+ \beta_3 \,(\text{treatment x time}) + e
\end{aligned}
$$

- **Interaction!:** outcome was observed in the **treatment group AND** it was observed **after the intervention** (different -or -reversed- slope vs control group)
- **grouped data**: always compare treatment and control groups
- **coefficients**: represent group means and their differences

Diff-in-diff: statistical model

$$y = \mu + \beta_1 \, \text{treatment} + \beta_2 \, \text{time} + \beta_3 \, (\text{treatment x time}) + e$$

- **β₁**: treatment - control (conditioned/independent on/of time: before applying the treatment)
- **β₂**: after - before (without treatment: in the control group; trend independent of treatment)
- **β₃**: (treat_after - treat_before) - (ctrl_after - ctrl_before)

diff-in-diff

Diff-in-diff: statistical model

$$y = \mu + \beta_1 \text{treatment} + \beta_2 \text{time}$$
$$+ \beta_3 (\text{treatment x time}) + e$$

- **β**$_1$: treatment - control (conditioned/independent on/of time: before applying the treatment)
- **β**$_2$: after - before (without treatment: in the control group; trend independent of treatment)
- **β**$_3$: (treat_after - treat_before) - (ctrl_after - ctrl_before)

how much the average outcome of the treatment group has changed in the period after the treatment, compared to **what would have happened had the intervention not occurred**

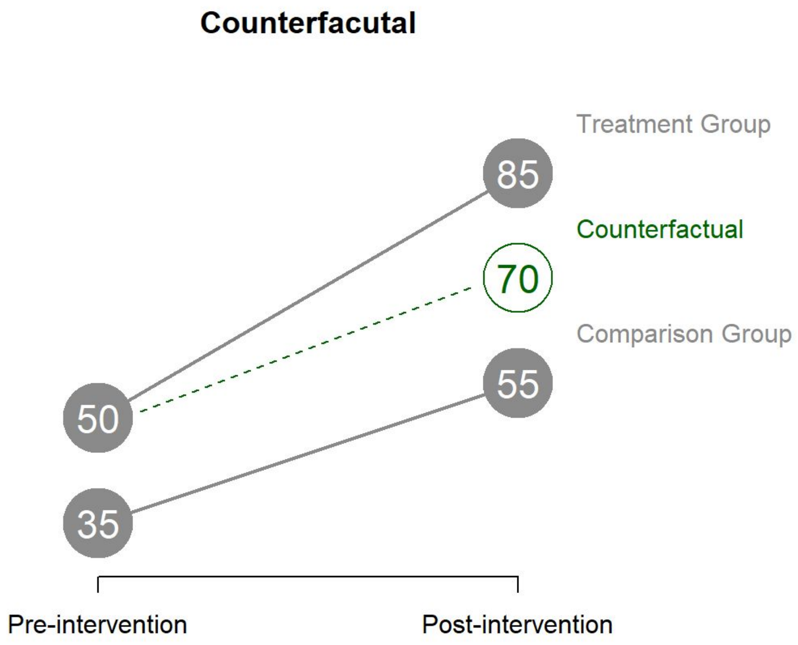
if **β**$_3$= 0 → the treatment had no effect

counterfactual!

Diff-in-diff: hypotheses

- **β_0**: intercept → **is the average outcome of the control group before the treatment ≠ 0?**
- **β_1**: (treatment - control | time) → **is the difference between treatment and control before the treatment ≠ 0?**
- **β_2**: (after - before | treatment) → **is the difference before and after the treatment in the control group ≠ 0?**
- **β_3**: (treat_after - treat_before) - (ctrl_after - ctrl_before): **is diff-in-diff ≠ 0?** (Does the treatment have an effect?)

Diff-in-diff: counterfactual



[From: https://ds4ps.org/PROG-EVAL-III/DiffInDiff.html]

- counterfactual: what would have occured to **y** had the intervention not happened
- in the diff-in-diff model, the counterfactual is the outcome of the treated group, had the intervention not occured (extrapolated from the control trend)
- β_3 represents the difference between the counterfactual and the average actual outcome of the treatment group after the treatment

Diff-in-diff: concluding remarks

- we presented here a super-simplified introduction to the difference-in-differences methodology
- diff-in-diff can be mistaken for a “quick and easy” way to answer causal questions (it is actually much more complex than that …)
- synthetic controls: if you don’t actually have control observations, you can create a synthetic control group from existing data (e.g. covid-19 vaccination policy applied in country A: compare with similar countries that did not apply vaccination / or different vaccination policy → many options to construct the control group)