Universiteit Utrecht

Estimating causal effects of policy interventions

Workshop ODISSEI

Erik-Jan van Kesteren (& Oisín Ryan)

Today we use R. You like python better?

There's a book for that:

Causal Inference for The Brave and True

https://is.gd/sicsscausal

About me

SoDa

Erik-Jan van Kesteren

- Team lead ODISSEI SoDa team
- Background in statistics / social science
- Assistant professor @ Methods & Statistics UU

Some stuff I work on:

Latent variables, high-dimensional data, optimization, regularization, visualisation, Bayesian statistics, multilevel models, spatial data, generalized linear models, privacy, synthetic data, high-performance computing, software development, open science & reproducibility

Today's Goal

A brief practical introduction on evaluating the causal effects of policy interventions

The plan

- Introduction
  - Policy Interventions and Causal Inference
  - Pre-Post Analyses and Difference-in-Difference
- Practical
- Break
- Interrupted Time Series & Synthetic Control
- Practical

Policy Evaluations

Evaluating what the **effect** of implementing a particular **policy** or **intervention** was on some outcome of interest

Examples:

- What was the effect of raising the maximum speed limit on road deaths?
- What effect did introducing student loans have on post-graduation debt levels?
- Did introducing an after-school programme in disadvantaged neighbourhoods lead to improved educational outcomes in children from that neighbourhood?

Policy evaluations

Register data is great for this purpose!

- Historical data availability
- Wide range of variables to create outcome of interest
- Many options to create, inspect, and match potential control units (e.g., other schools, neighbourhoods)

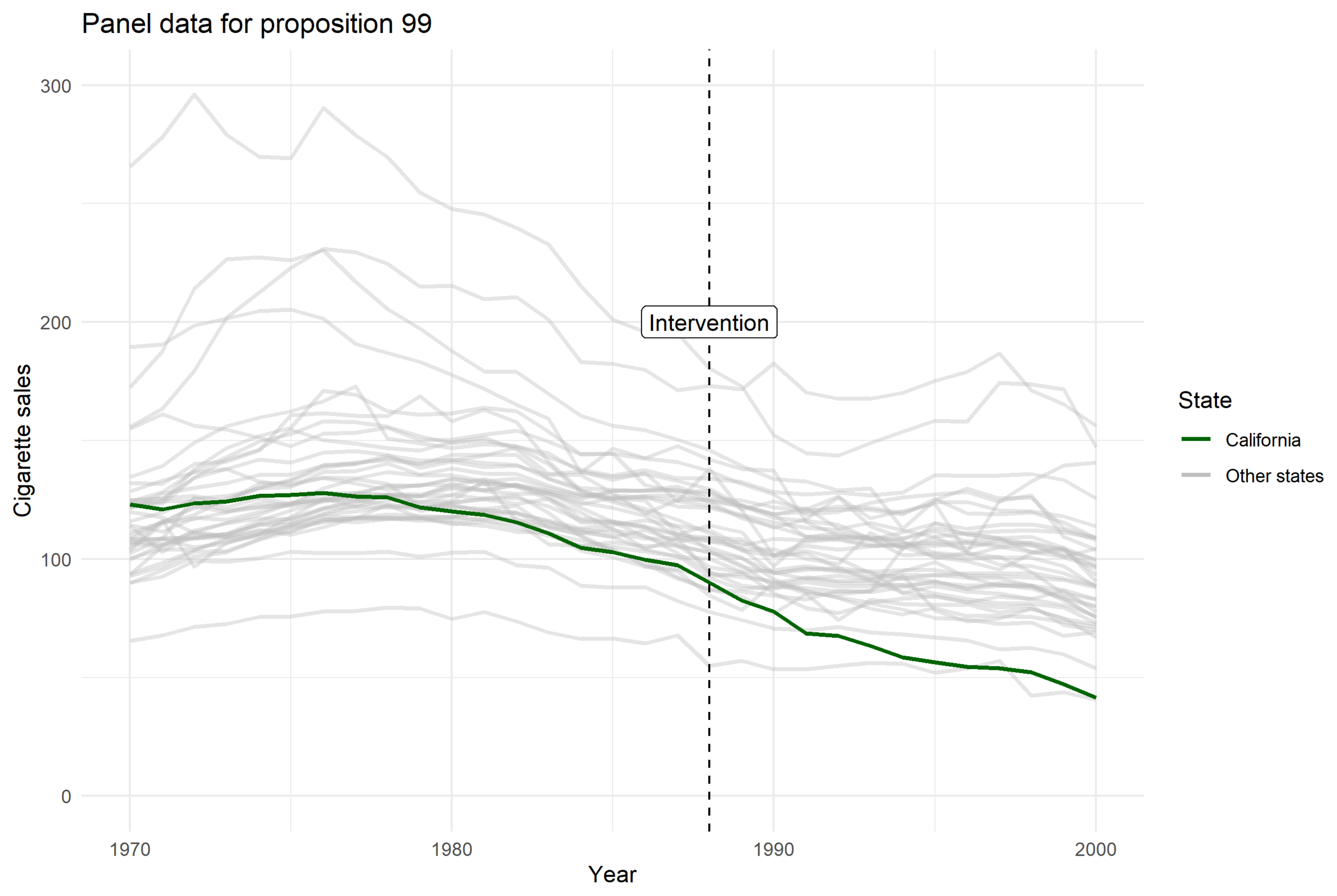
Running Example: Proposition 99

Proposition 99

- A famous example in causal inference literature

> *Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies:* ***Estimating the effect of California's tobacco control program****. Journal of the American statistical Association, 105(490), 493-505.*

- In 1988, the state of California imposed a 25% tax on tobacco cigarettes
- Total savings in personal health care expenditure until 2004 is $86 billion (Lightwood et al., 2008)



Panel data for proposition 99
Intervention
Cigarette sales
0
100
200
300
1970
1980
1990
2000
Year
State
California
Other states

Methods for Policy Evaluation

Many different methods have been developed to answer these types of research questions

Differing in:

- The **amount** and **type** of information they use
  - Number of time-points and potential “control” units
- The specific **statistical approach** they take
- The types of **assumptions** they make

Control Units

# Time-Points	None	Some
Two	Pre-Post	Diff-in-Diff
Many	Interrupted Time-Series (ITS)	Synthetic Control

Causal Inference: A primer

Potential Outcomes

Causal inference is (broadly) concerned with using **data** to estimate what the effect is of **intervening or changing** the value of one or more **variables**.

Using the **potential outcomes** framework, we can define causal inference as a *missing data problem*

Potential Outcomes: notation

- Let Y_i represent your headache level (high is bad)
- Let A_i be whether you take aspirin or not ($A_i = 1$ you take it, $A_i = 0$ you don't)

There are **two possible versions** of the outcome variable

- Y_i^1 your headache level **if you would take aspirin**
- Y_i^0 your headache level **if you would not take aspirin**

Causal Effects

We can define the **causal effect** of taking aspirin on your headache levels as the difference in potential outcomes

$$CE_i = Y_i^1 - Y_i^0$$

The **fundamental problem of causal inference:** You only ever observe one of the potential outcomes!

Data and Potential Outcomes

ID	Y	A
1	7	0
2	9	0
3	6	0
4	5	0
5	6	0
6	2	1
7	3	1
8	1	1
…	…	…
I	2	1

Data and Potential Outcomes

ID	Y	A	Y^0	Y^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	$\dots$	…
I	2	1	NA	2

Data and Potential Outcomes

ID	Y	A	Y^0	Y^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	…	…
I	2	1	NA	2

Causal Inference

In cross-sectional settings, we typically aim to make inferences about the **average causal effect.** This is known as a **causal estimand:**

$$ACE = E[Y^1] - E[Y^0]$$

In a **Randomized Controlled Trial,** we often use the difference in treated and untreated groups as an **estimator** of this causal effect:

$$\widehat{ACE} = E[Y|\, A = 1] - E[Y\,|A = 0]$$

Causal Inference

ID	Y	A	Y^0	Y^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	$\dots$	…
I	2	1	NA	2

Causal Inference

In cross-sectional settings, we typically aim to make inferences about the **average causal effect.** This is known as a **causal estimand:**

$$ACE = E[Y^1] - E[Y^0]$$

In a **Randomized Controlled Trial,** we often use the (sample) difference in treated and untreated groups as an **estimator** of this causal effect:

$$\widehat{ACE} = E[Y|\,A = 1] - E[Y\,|A = 0]$$

Causal Inference

ID	Y	A	Y^0	Y^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	…	…
I	2	1	NA	2

Causal Inference Assumptions

This type of **inference** about causal effects from **observed data** is only possible under certain **conditions** or **assumptions**

Exchangeability

- If we were to reverse treatment assignment we would observe the same group differences. Information is exchangeable between groups
- Basically: absence of **confounder variables**
  - E.g., People who have bad headaches choose to take the aspirin
- **RCTs** are powerful because **randomization** ensures exchangeability. But in principle this kind of inference is possible from non-RCT designs
- In practice we need **conditional exchangeability**; to control for **confounders!**

Causal Inference Assumptions

This type of **inference** about causal effects from **observed data** is only possible under certain **conditions** or **assumptions**

Stable Unit Treatment Value (also known as SUTVA)

- No Interference: The potential outcomes of one unit does not depend on the treatment assigned to another unit.
  - No “spillover”: My taking an aspirin does not influence your headache levels
- Consistency: Only one version of treatment, treatment is unambiguous
- I can directly observe one of the potential outcomes. If you receive treatment, then for you I observe $Y_i = Y_i^1$

Causal Inference Assumptions

- These two often appear in causal inference
- Need to deal with **confounders** and **no interference**

NB:

- **Other assumptions or conditions** may also be needed
- Depends on **design** and **analytic approach you take**

Causal inference for policies

Policy evaluation is a special case of causal inference:

- Usually: **one unit** observed **repeatedly over time**
- At some point in time (T_0) an **intervention** takes place

Pre-intervention we observe Y_t^0 and **post-intervention** Y_t^1

$Time$	Y_t	A_t
1	7	0
2	9	0
3	6	0
4	5	0
5	6	0
6	2	1
7	3	1
8	1	1
…	…	…
T	2	1

$Time$	Y_t	A_t	Y_t^0	Y_t^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
...	...	...	$\cdots$	...
T	2	1	NA	2

Causal Effects of Policies

Estimate the **causal effect of the policy intervention** as difference between:

(a) the **observed outcome** <u>after</u> the policy was introduced

(b) What the outcome **would have been** without the intervention

$$CE_t = Y_t^1 - Y_t^0$$

where $t > T_0$ (i.e., the post-intervention time period)

$Time$	Y_t	A_t	Y_t^0	Y_t^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	$\cdots$	…
T	2	1	NA	2

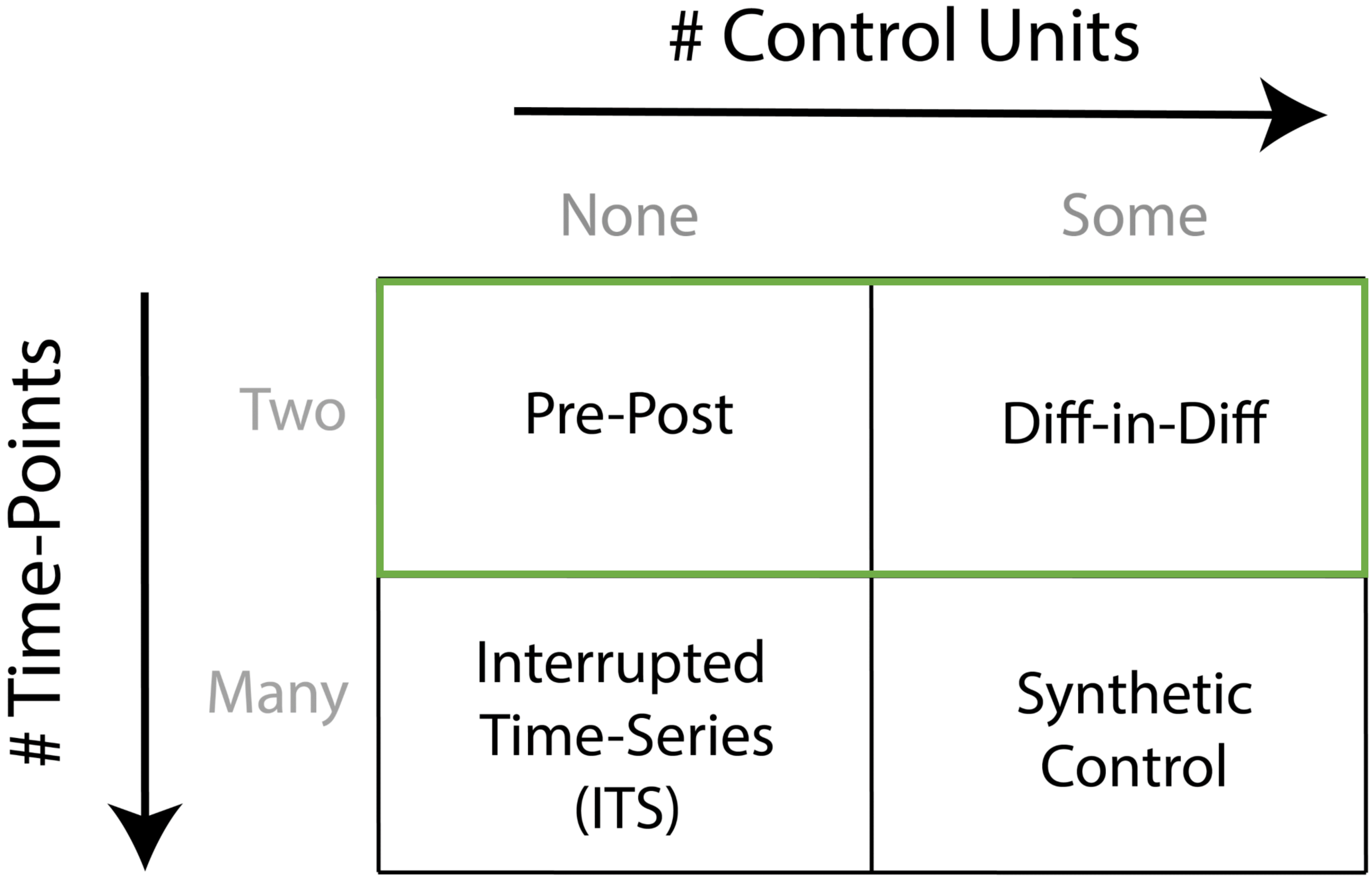
$Time$	Y_t	A_t	Y_t^0	Y_t^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	…	…
T	2	1	NA	2

The problem of estimating the effect of a policy intervention is equivalent to the problem of estimating Y_t^0

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. Journal of Economic Literature, 59(2), 391-425.

Estimating the causal effect

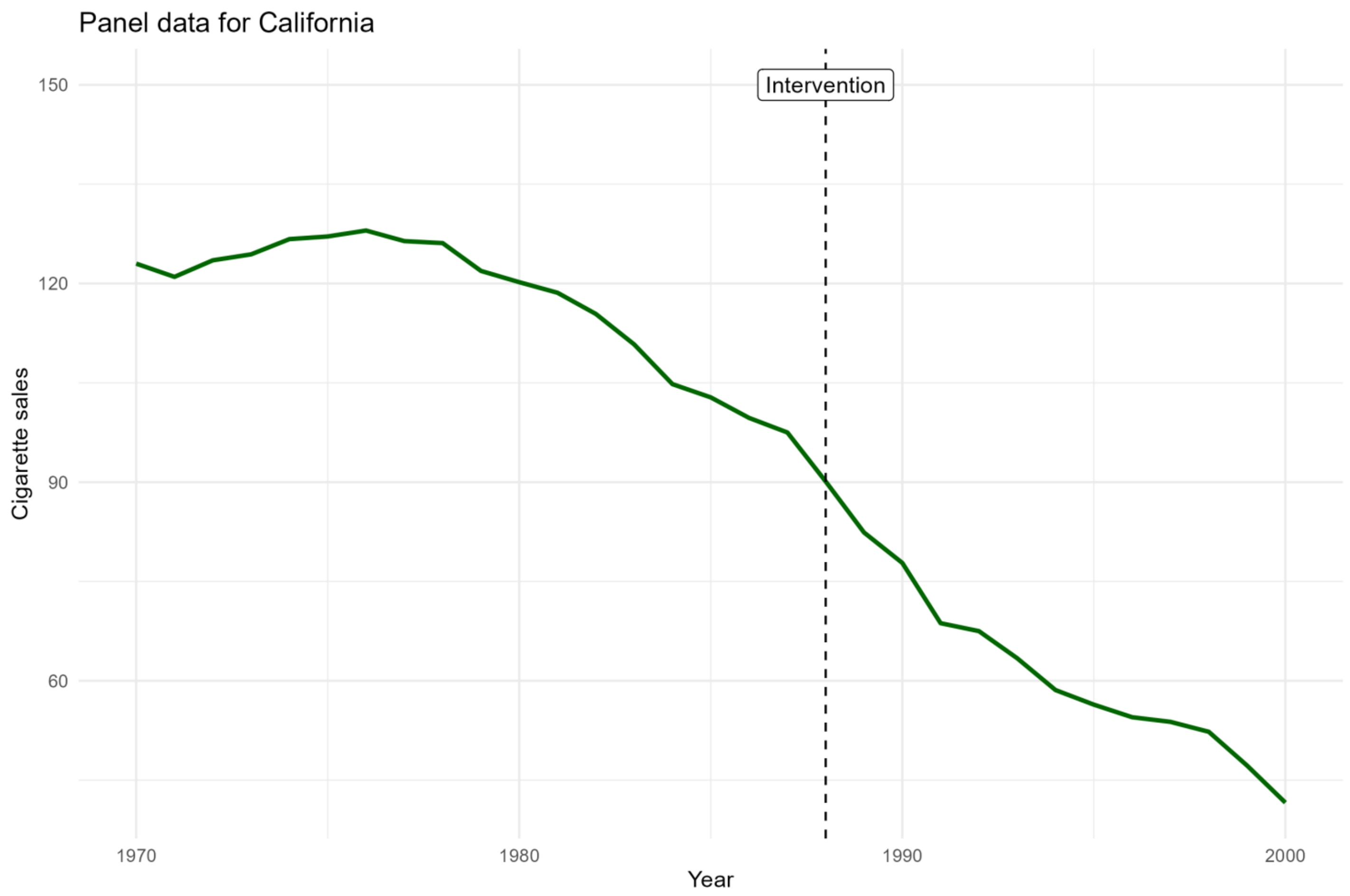
Basic methods


Control Units
None
Some
Time-Points
Two
Many
Pre-Post
Diff-in-Diff
Interrupted Time-Series (ITS)
Synthetic Control

Pre-Post Estimator

Pre-post estimator

We use only the cigarette sales time series for California



Pre-post estimator

- We want to estimate the following quantity:

$$\overline{CE}_{post} = \bar{Y}^{1}_{post} - \bar{Y}^{0}_{post}$$

- But we cannot observe $\bar{Y}^{0}_{post}$!
- Solution: replace $\bar{Y}^{0}_{\boldsymbol{post}}$ by $\bar{Y}^{0}_{\boldsymbol{pre}}$, which is observable

$$\overline{CE}_{post} = \bar{Y}^{1}_{post} - \bar{Y}^{0}_{pre}$$

Pre – Post analysis

$Time$	Y_t	A_t	Y_t^0	Y_t^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	…	…
T	2	1	NA	2

$\bar{Y}_{pre}^{0}$

$\bar{Y}_{post}^{1}$

Pre – Post analysis

$Time$	Y_t	A_t	Y_t^0	Y_t^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	…	…
T	2	1	NA	2

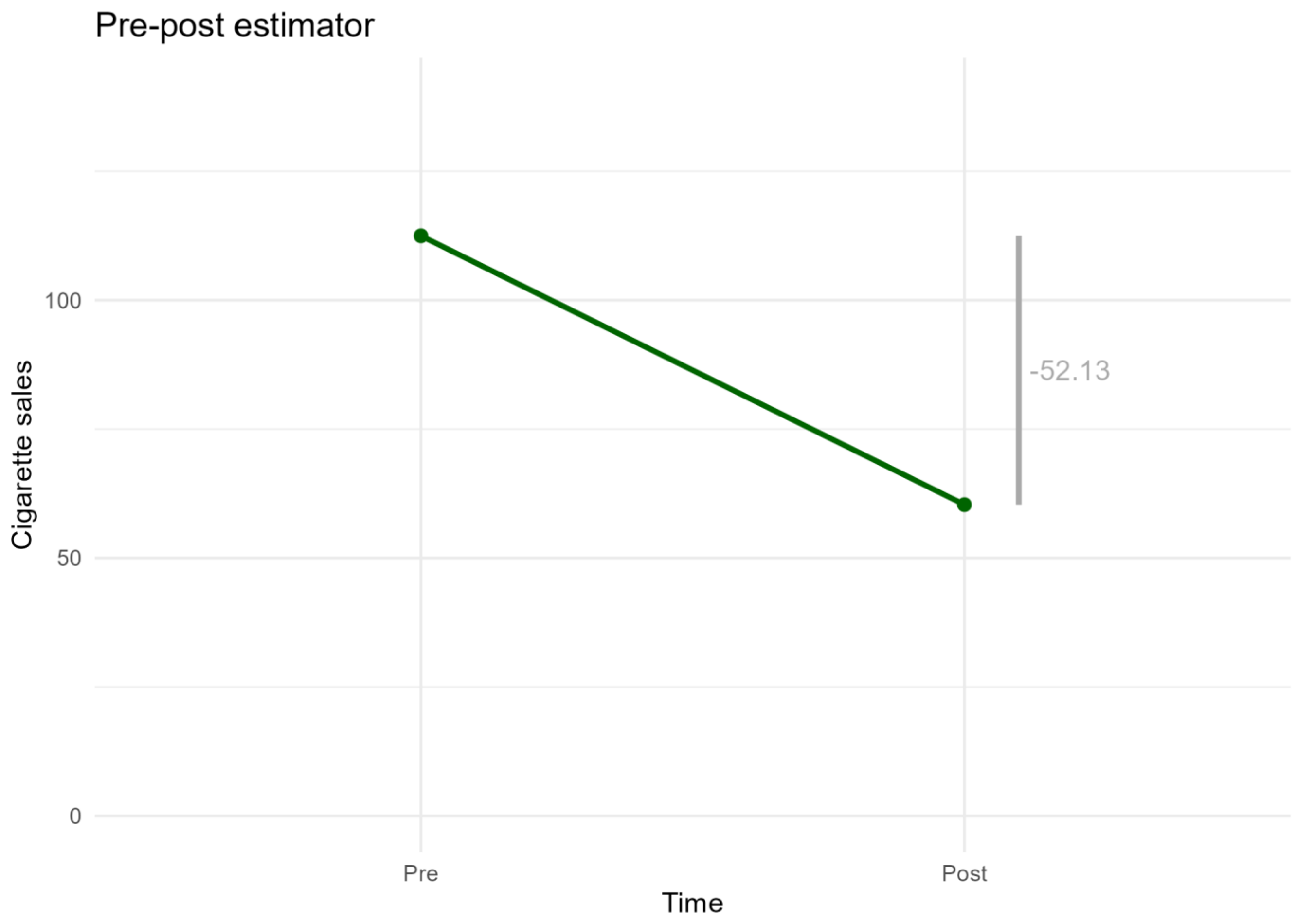
$\bar{Y}^0_{pre}$

Assume equal to

$\bar{Y}^1_{post} - \bar{Y}^0_{post}$

$$\overline{CE}_{post} = \bar{Y}^1_{post} - \bar{Y}^0_{post}$$

Pre-post estimator



Pre-post estimator

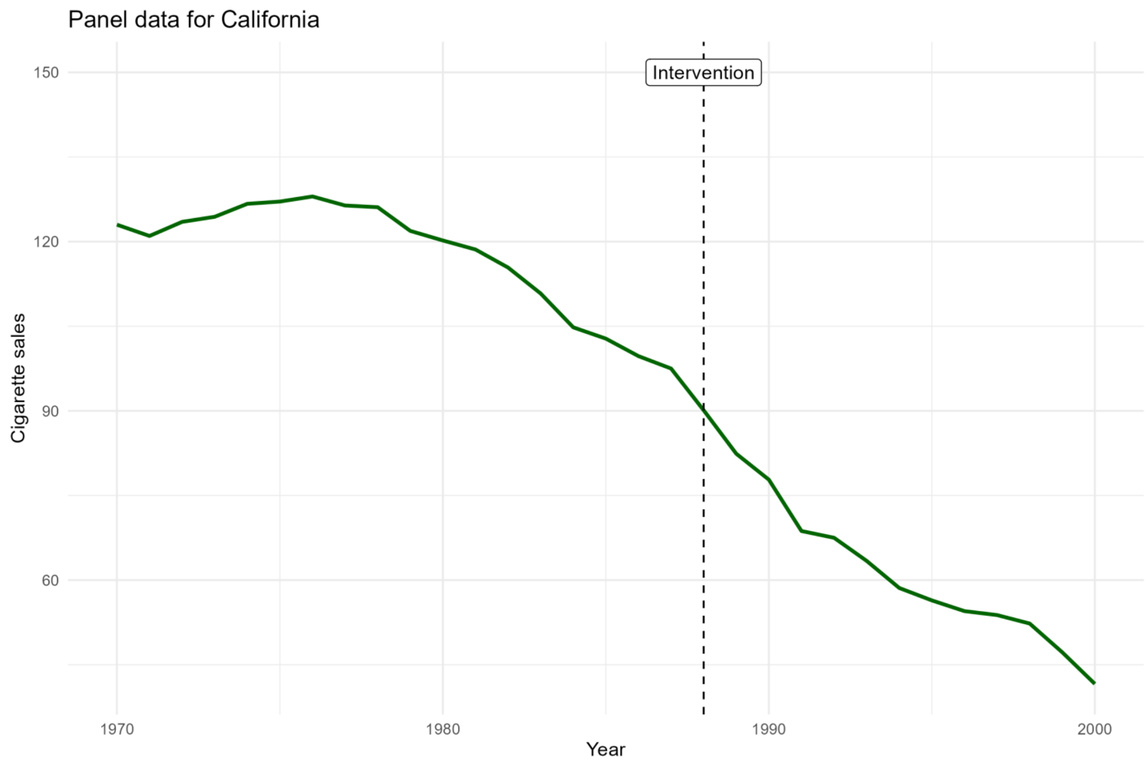
Most important / strict assumption:

No trend in time

- Remember: we assumed $\bar{Y}^0_{post} = \bar{Y}^0_{pre}$
- We assume the pre-post difference is caused by intervention **only**
- If trend exists, then the effect of trend and of intervention cannot be distinguished

Pre-post estimator

- Is there a trend in time, independent of the intervention?
- How much of pre-post difference is caused by intervention?



Difference-in-Differences

Difference-in-differences

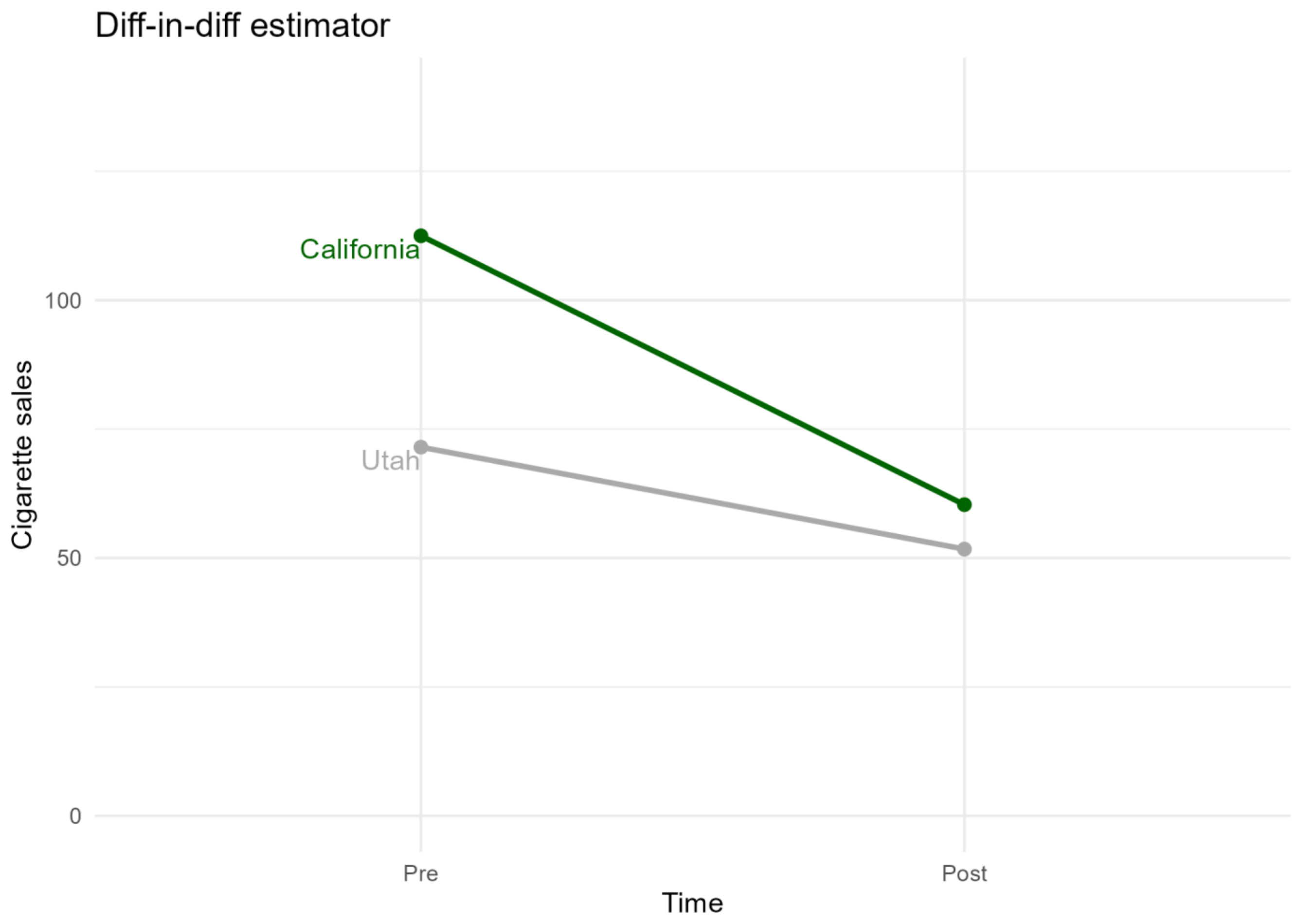
„transparent and often at least superficially plausible"

Angrist, J. D. and Krueger, A. B. (1999). Empirical strategies in labor economics. In Handbook of labor economics, volume 3, pages 1277–1366. Elsevier.

- Used a lot in economics
- There is a lot of discussion around this topic
- We will explain the basic method here
- There are a lot of possible extensions!

$Time$	Y_t	A_t	Y_t^0	Y_t^1	C_{1t}
1	7	0	7	NA	2
2	9	0	9	NA	6
3	6	0	6	NA	4
4	5	0	5	NA	2
5	6	0	6	NA	1
6	2	1	NA	2	3
7	3	1	NA	3	2
8	1	1	NA	1	4
…	…	…	…	…	…
T	2	1	NA	2	3

Difference-in-differences



Pre-post estimator

- Like before, we estimate the following quantity:

$$\overline{CE}_{post} = \bar{Y}^{1}_{post} - \bar{Y}^{0}_{post}$$

- Now, we assume there is an effect of time: $\beta \cdot Time$
- We can represent unobservable $\bar{Y}^{0}_{post}$ as

$$\bar{Y}^{0}_{post} = \bar{Y}^{0}_{pre} + \beta \cdot Time$$

Pre-post estimator

- But the trend $\beta \cdot Time$ is also unobservable!
- Solution: assume equal trends for Utah and California

$$\beta \cdot Time = (\bar{C}^0_{post} - \bar{C}^0_{pre})$$

- Thus, our model for the counterfactual is:

$$\bar{Y}^0_{post} = \bar{Y}^0_{pre} + (\bar{C}^0_{post} - \bar{C}^0_{pre})$$

Pre-post estimator

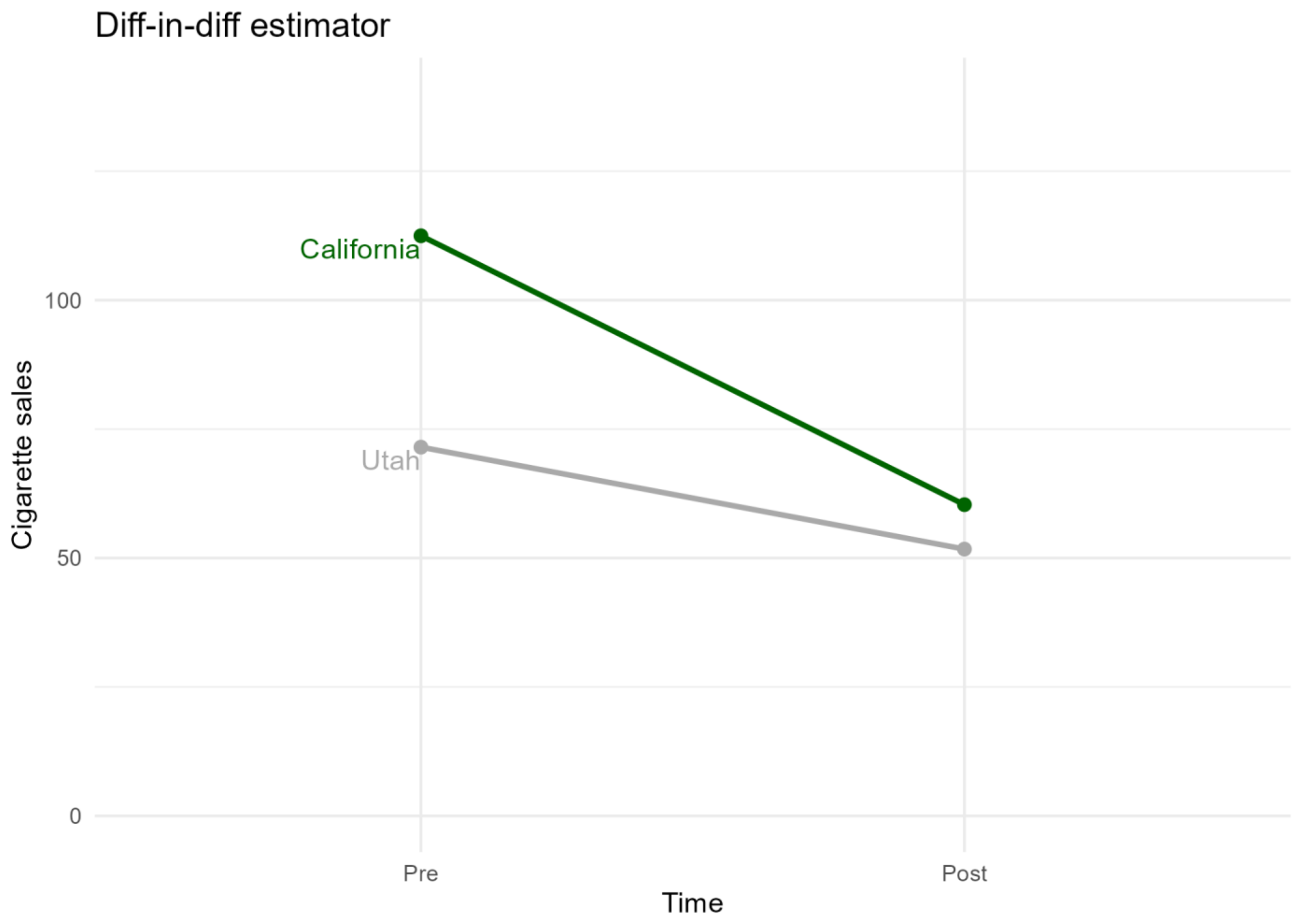
- Plugging this into the causal effect equation:

$$\overline{CE}_{post} = \left(\bar{Y}_{post}^{1} - \bar{Y}_{pre}^{0}\right) - \left(\bar{C}_{post}^{0} - \bar{C}_{pre}^{0}\right)$$

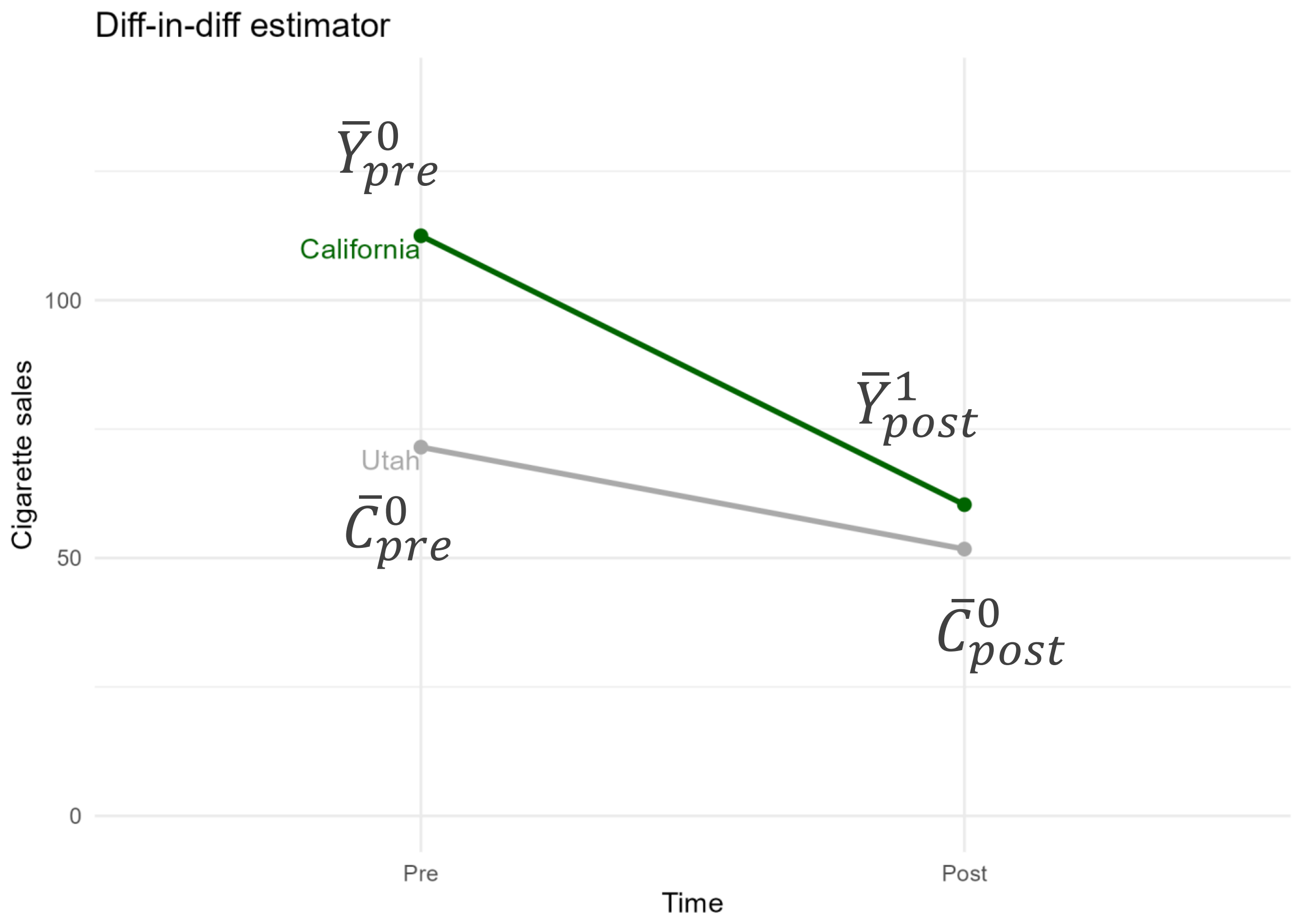
- Difference in differences!

$$\widehat{CE}_{post} = \left(\bar{Y}_{post} - \bar{Y}_{pre}\right) - \left(\bar{C}_{post} - \bar{C}_{pre}\right)$$

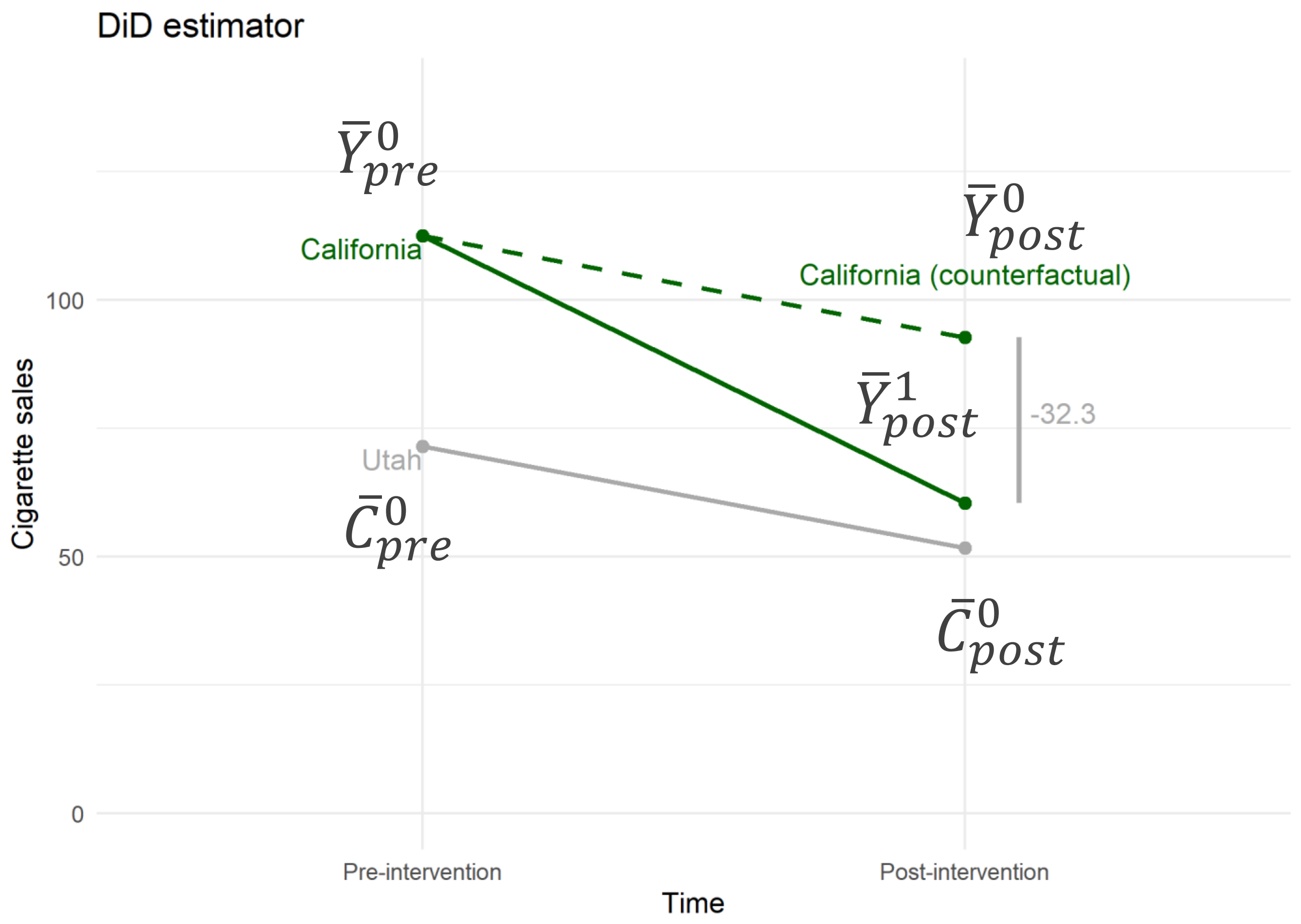
Difference-in-differences



Difference-in-differences



Difference-in-differences



Most important assumptions

Parallel trends

$$\beta \cdot Time = (\bar{C}^0_{post} - \bar{C}^0_{pre})$$

Time effect is the same for the treated and the control unit

No interference / spillover

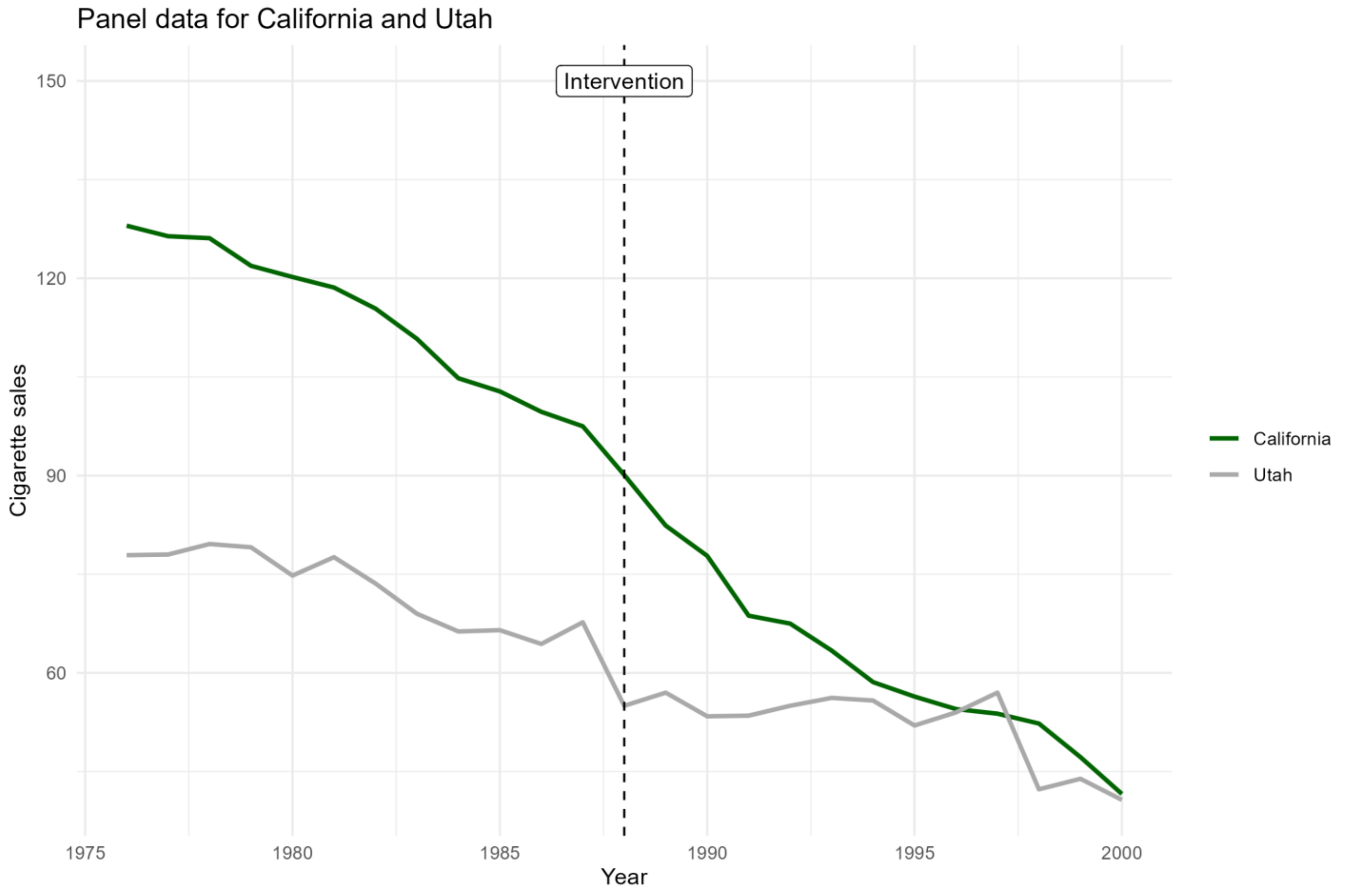
$$\bar{C}_{post} = \bar{C}^0_{post}$$

The control does not receive any intervention effect

Most important assumptions

- Can we assume parallel trends?

- At least superficially plausible ☺



Practical

Work in pairs/groups!

https://is.gd/sicsscausal

Break