Introduction to Causal Inference and Causal Data Science

Estimating Causal Effects of Policy Interventions

Oisin Ryan

UMC Utrecht

Context: “Policy Evaluations”

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

Examples:

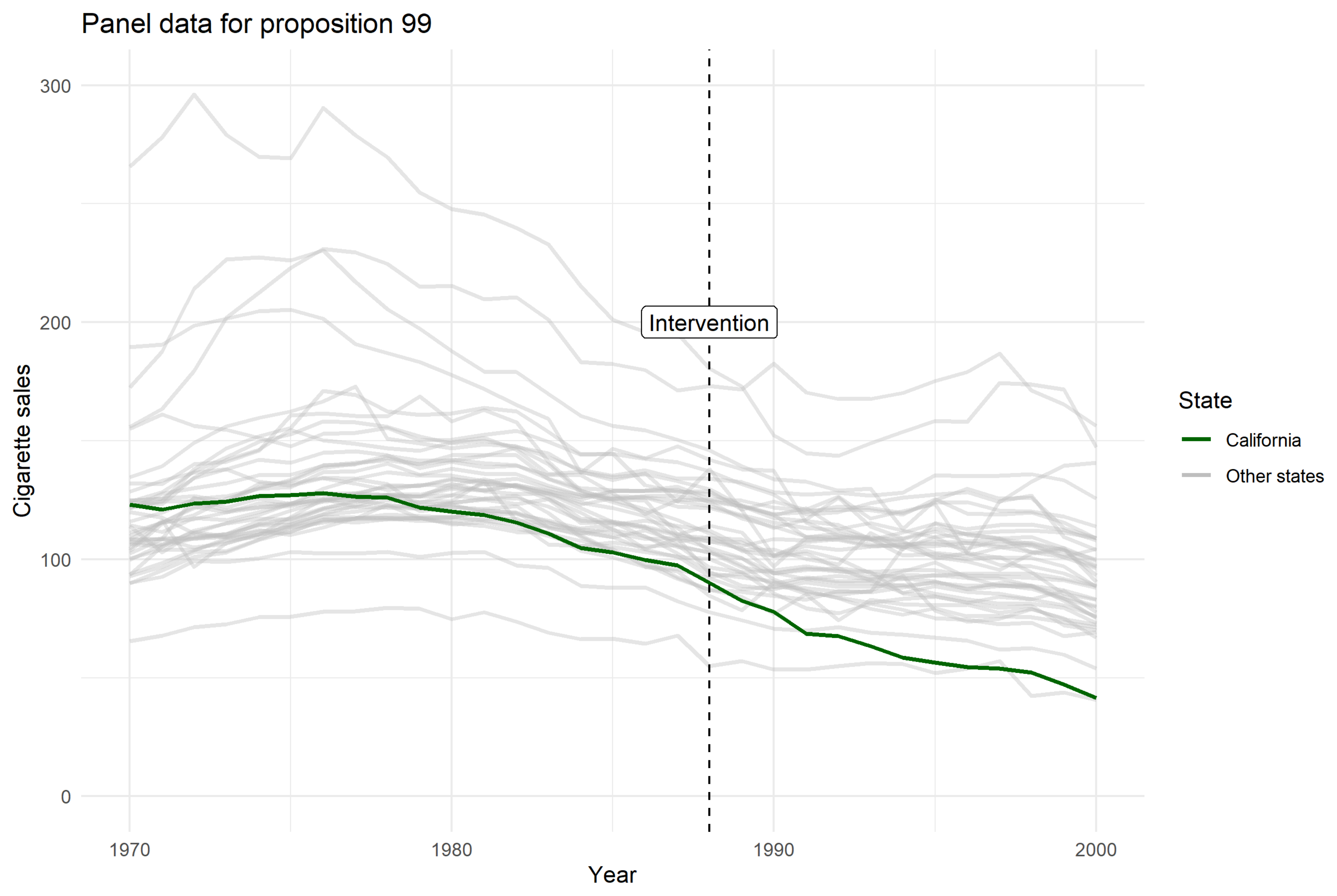
- What was the effect of raising the maximum speed limit on motorways in the Netherlands on road deaths?
- Did introducing an after-school programme in disadvantaged neighbourhoods lead to improved educational outcomes in children from that neighbourhood?

Running Example: Proposition 99

- A famous example in policy evaluation 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
- Did this intervention successfully reduce cigarette sales in California?


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

Causal Policy Evaluation

Basic Structure:

- We have one (or more) **unit(s)** which we observe **before** and **after** some intervention or action
- Did the intervention produce a change in the outcome for that unit?

Causal Policy Evaluation

Basic Structure:

- We have one (or more) **unit(s)** which we observe **before** and **after** some intervention or action
- Did the intervention produce a change in the outcome for that unit?

Many different methods/approaches which differ in:

- The **amount** and **type** of information they use
  - Amount of time-points and amount of potential “control” units
- The specific **statistical approach** they take
- The types of **assumptions** they rely on for identification

Control Units

Time-Points

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

This Lecture

Policy evaluation through the lens of potential outcomes

We will consider the case where:

- we have **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

We want to estimate the **causal effect of the policy intervention**

We think about this as the difference between

(a) the **observed outcome** after 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
…	…	…	…	…
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

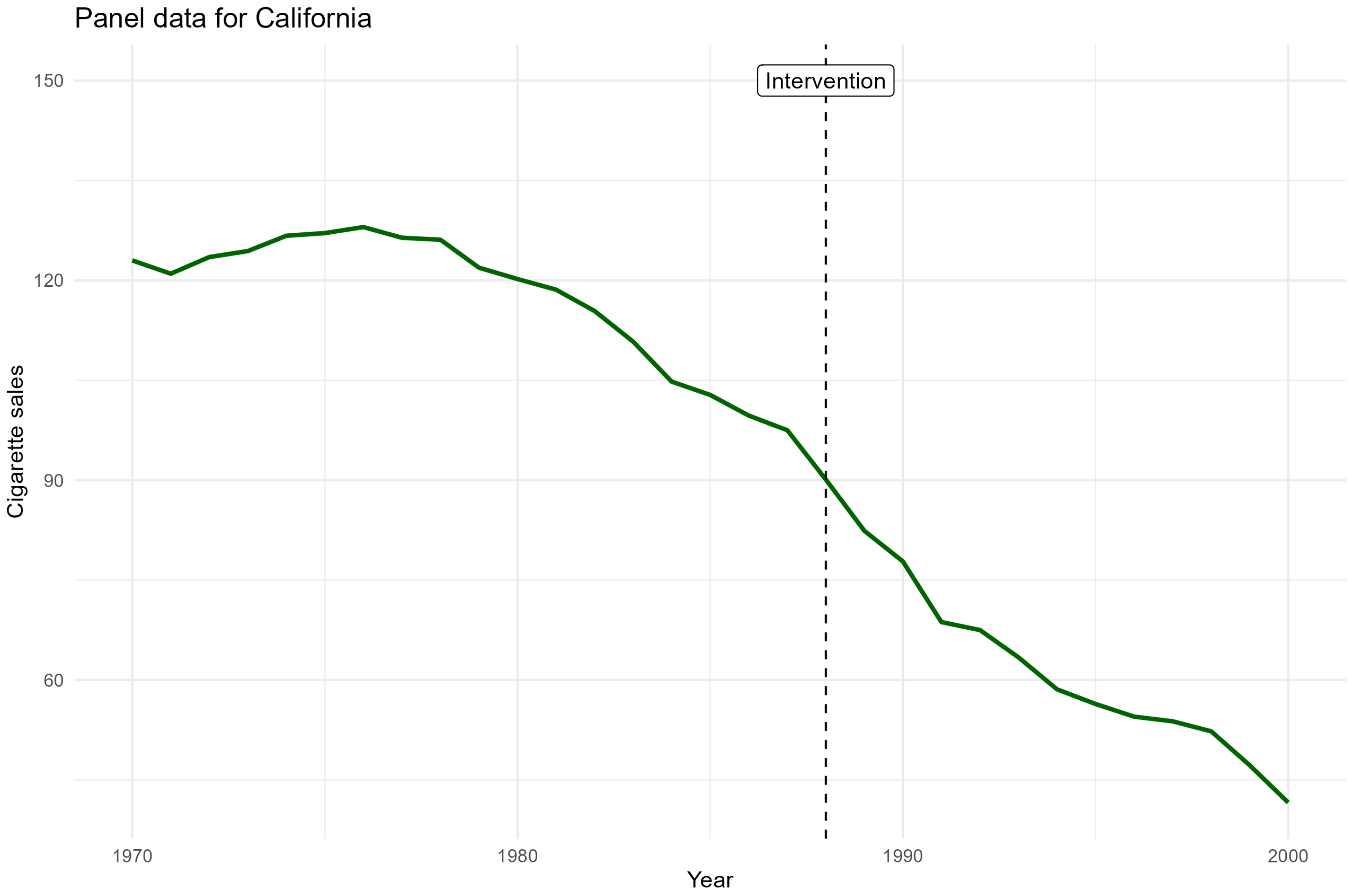
The problem of estimating the effect of a policy intervention is equivalent to the problem of estimating Y_t^0

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

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:

$$CE_{post} = E[Y^1_{post}] - E[Y^0_{post}]$$

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

$$CE_{post} = E[Y^1_{post}] - E[Y^0_{pre}]$$

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
…	…	…	$\dots$	…
T	2	1	NA	2

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

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

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

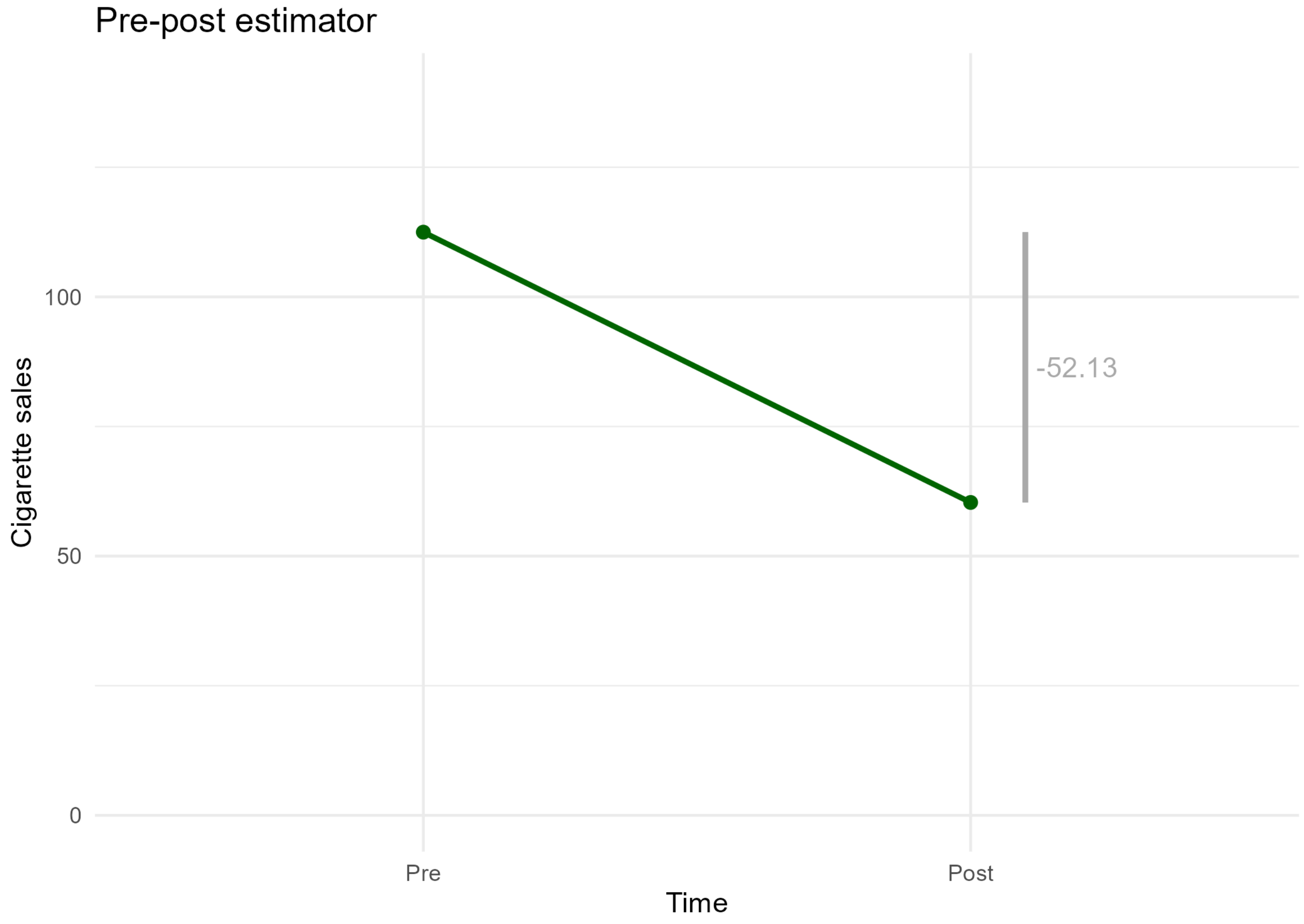
$\bar{Y}^0_{pre}$

Assume equal to

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

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

Pre-post estimator



Can estimate uncertainty by using a regression model on the disaggregated data, correcting SEs for unmodelled autocorrelation

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 **<u>only</u>**
- If trend exists, then the effect of trend and of intervention cannot be distinguished

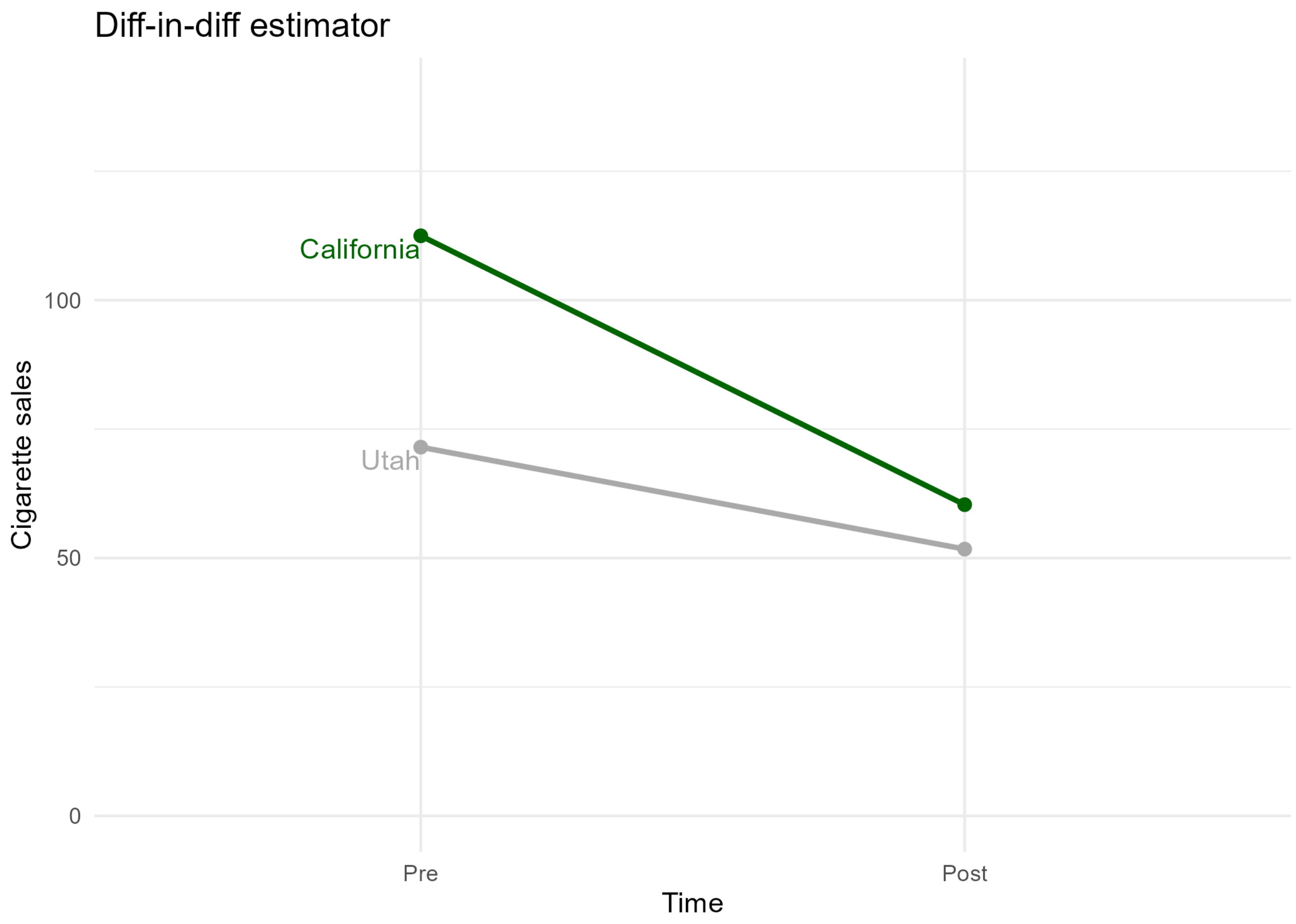
Difference-in-Differences

„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.

$Time$	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



Difference-in-differences

- Like before, we want to 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

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

Difference-in-differences

- 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:

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

Difference-in-differences

- Plugging this into the causal effect equation:

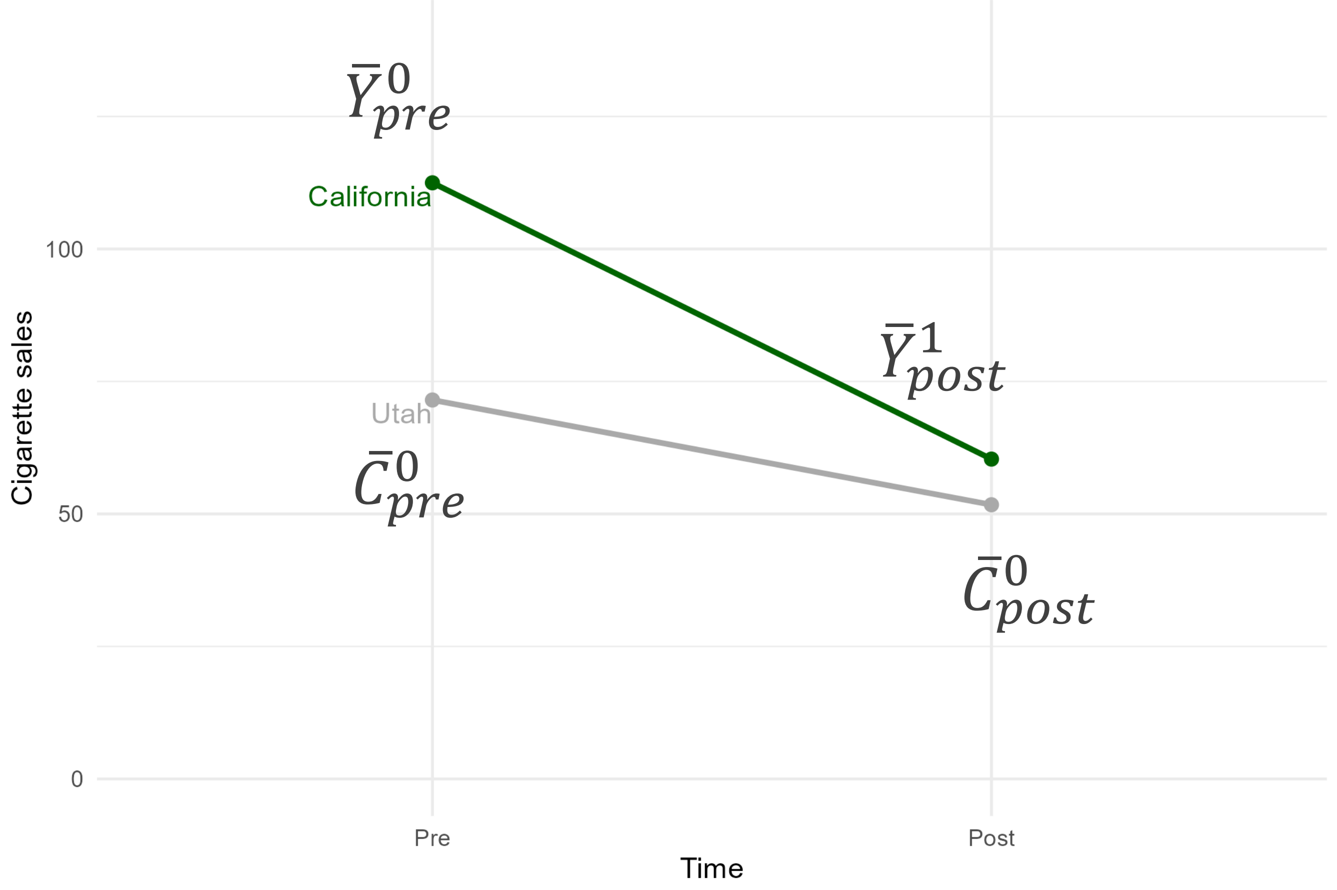
$$\overline{CE}_{post} = \left(\bar{Y}^1_{post} - \bar{Y}^0_{pre}\right) - \left(\bar{C}^0_{post} - \bar{C}^0_{pre}\right)$$

- Difference in differences!

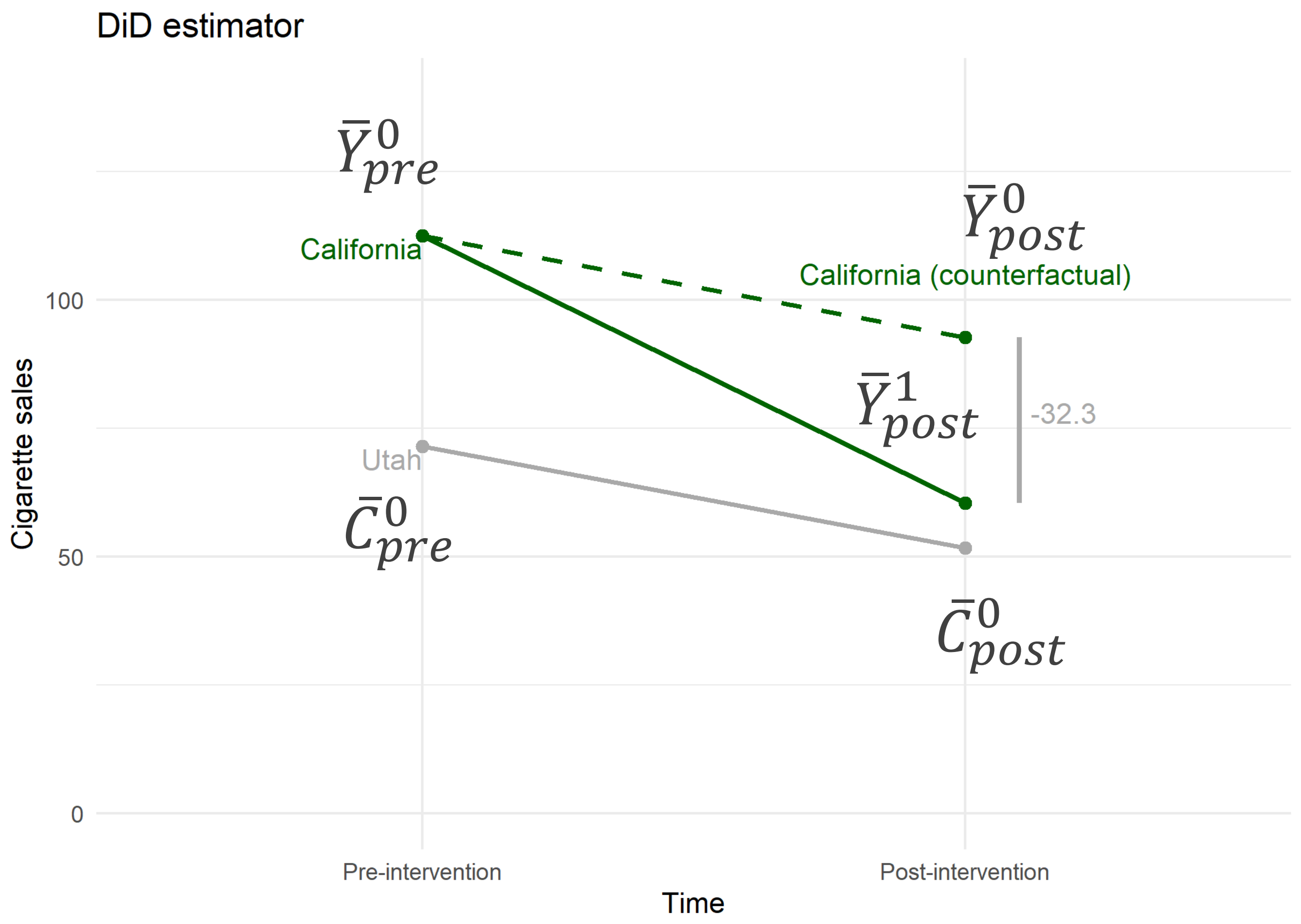
$$\widehat{CE}_{post} = \left(\bar{Y}_{post} - \bar{Y}_{pre}\right) - \left(\bar{C}_{post} - \bar{C}_{pre}\right)$$

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

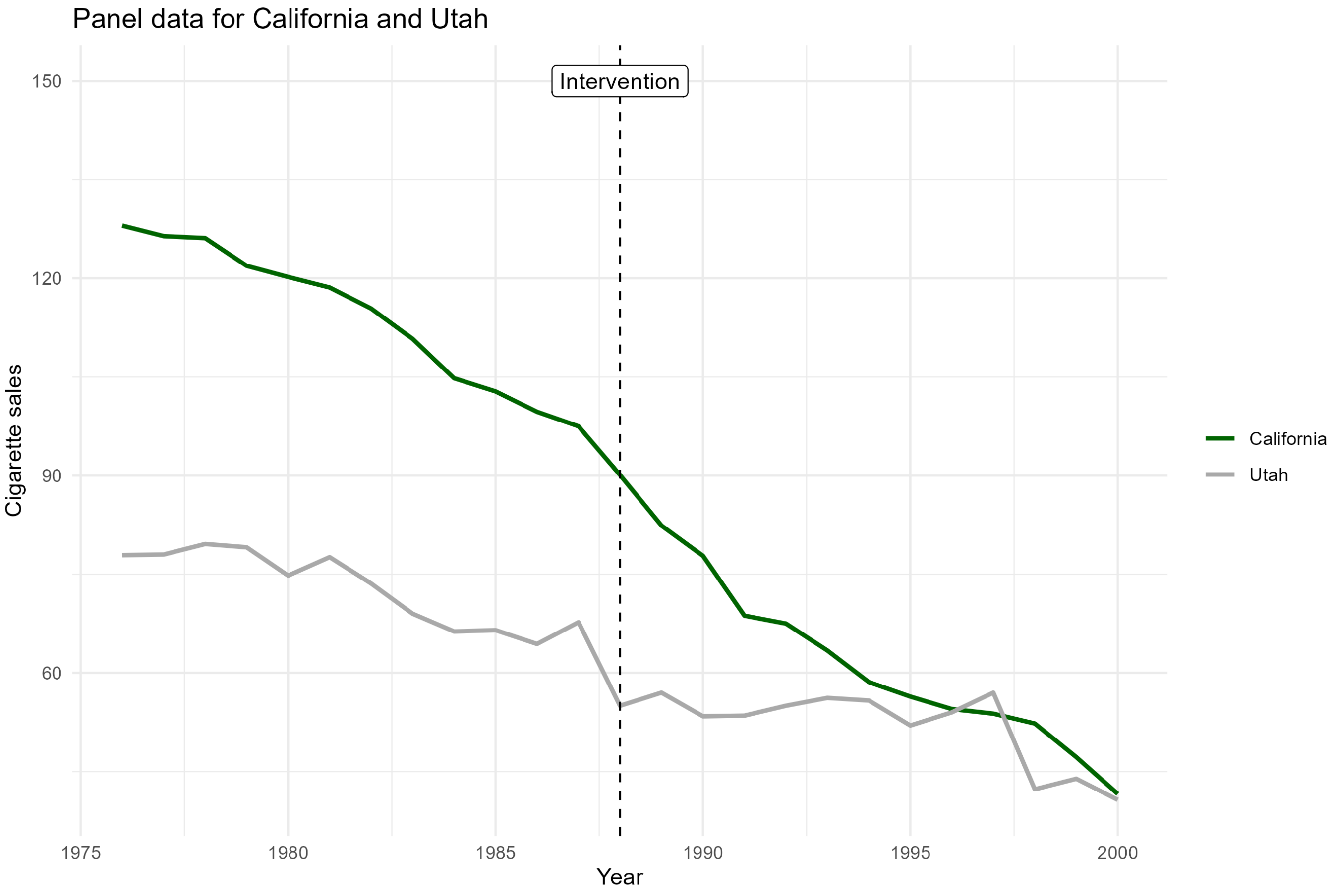
$\bar{C}_{post} = \bar{C}^0_{post}$

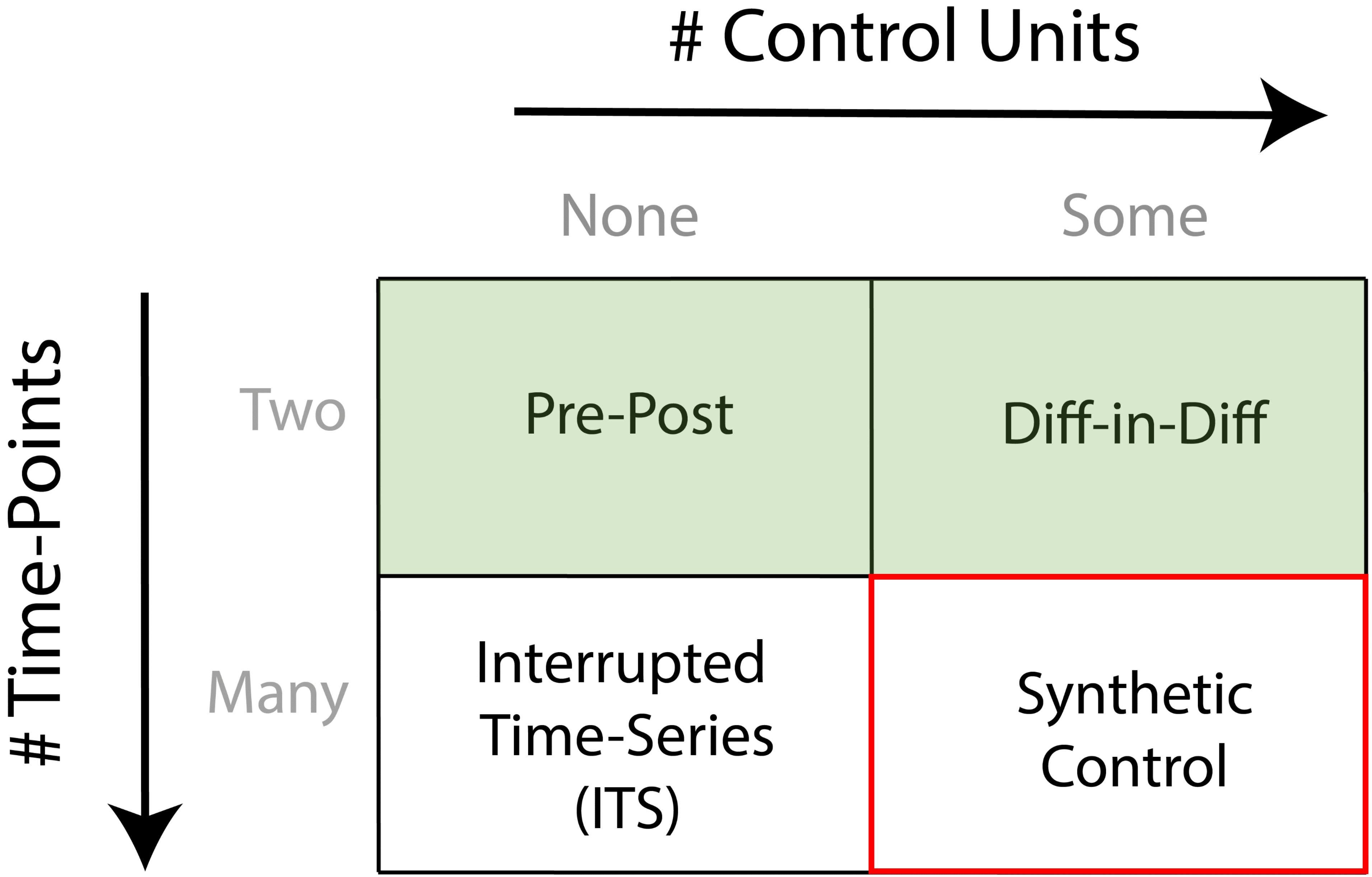
The control unit is not affected by the intervention

Most important assumptions

- Can we assume parallel trends?

- Superficially plausible?





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

Synthetic Control

„arguably the most important innovation in the policy evaluation literature in the last 15 years”

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. Journal of Economic perspectives, 31(2), 3-32.

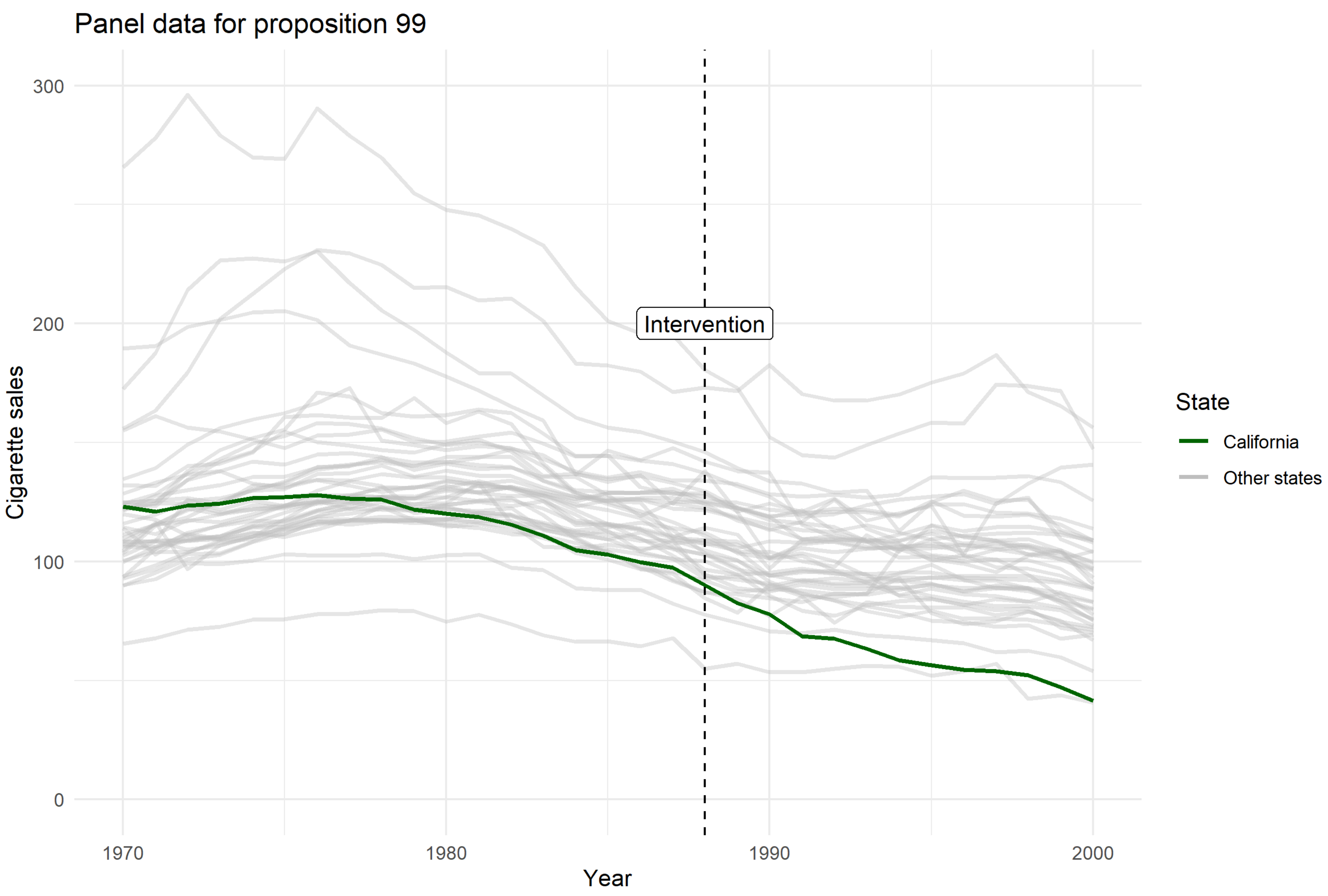
Basic idea

With diff-in-diff we used a control unit to attempt a correction for unmeasured time-varying confounders (e.g., macroeconomic situation in U.S.A.)

- You need a good control unit to make the parallel trends assumption at least superficially plausible
- But how much is Utah really like California?

We can instead use a weighted average of a **donor pool** of control units to create a **synthetic control** unit

- Choose the weights such that control is like California


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

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

Synthetic Control

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

Estimate Weights

$$Y_t = \sum_{j=1}^{J} \widehat{w_j} C_{jt} \qquad t < T_0$$

Synthetic Control

$Time$	Y_t	A_t	Y_t^0	Y_t^1	C_{1t}^0	C_{2t}^0	…	C_{jt}^0
1	7	0	7	NA	2	9	…	6
2	9	0	9	NA	6	9	…	8
3	6	0	6	NA	4	3	…	5
4	5	0	5	NA	2	1	…	4
5	6	0	6	NA	1	2	…	7
6	2	1	$\widehat{Y_6^0}$	2	3	6	…	7
7	3	1	$\widehat{Y_7^0}$	3	2	5	…	6
8	1	1	$\widehat{Y_8^0}$	1	4	6	…	5
…	…	…	…	…	…	…	…	4
T	2	1	$\widehat{Y_T^0}$	2	3	4	…	6

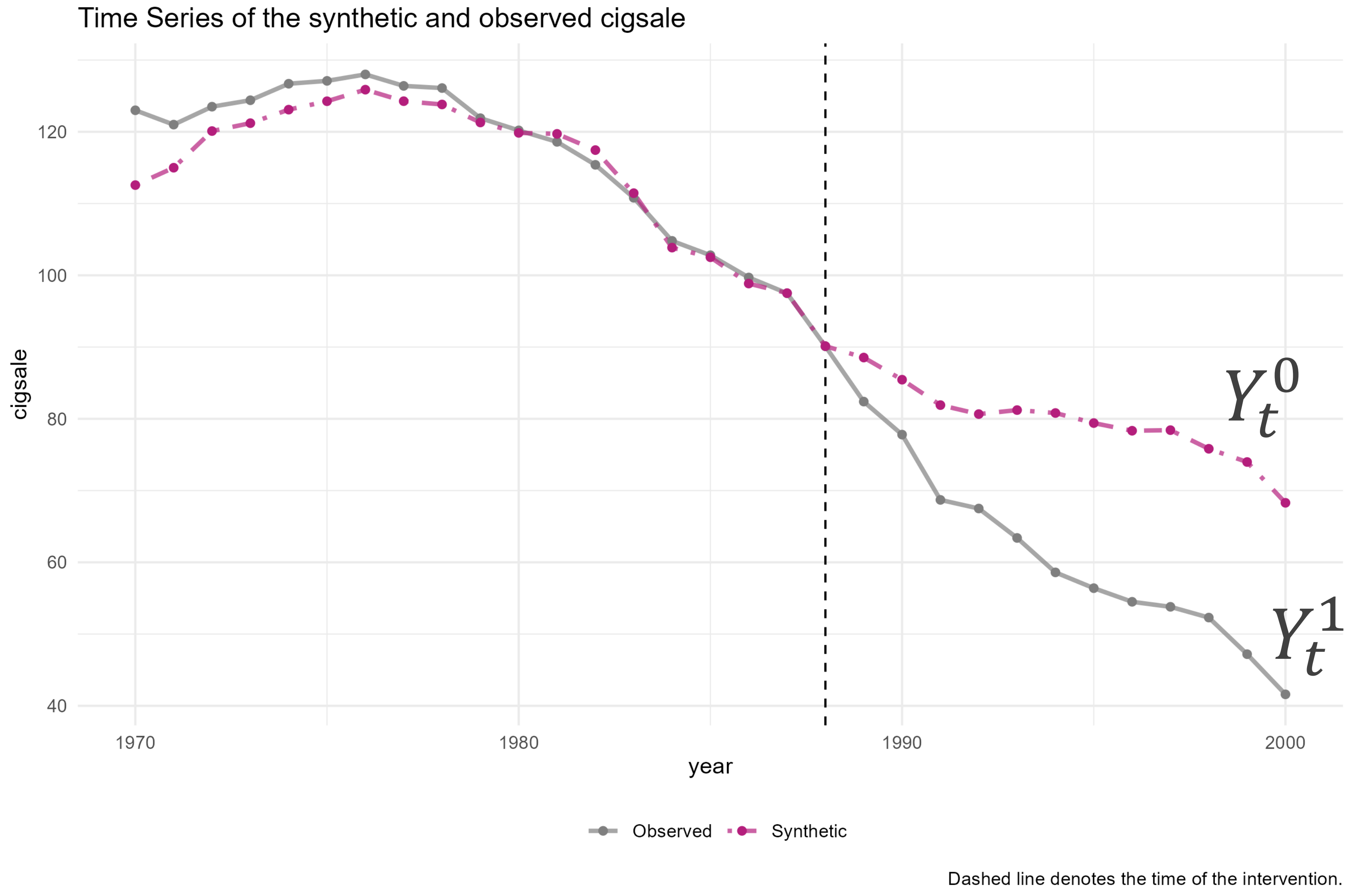
Estimate Weights

$$Y_t = \sum_{j=1}^{J} \widehat{w_j} C_{jt} \qquad t < T_0$$

$$\widehat{Y_t^0} = \sum_{j=1}^{J} \widehat{w_j} C_{jt} \qquad t > T_0$$

Impute counterfactual

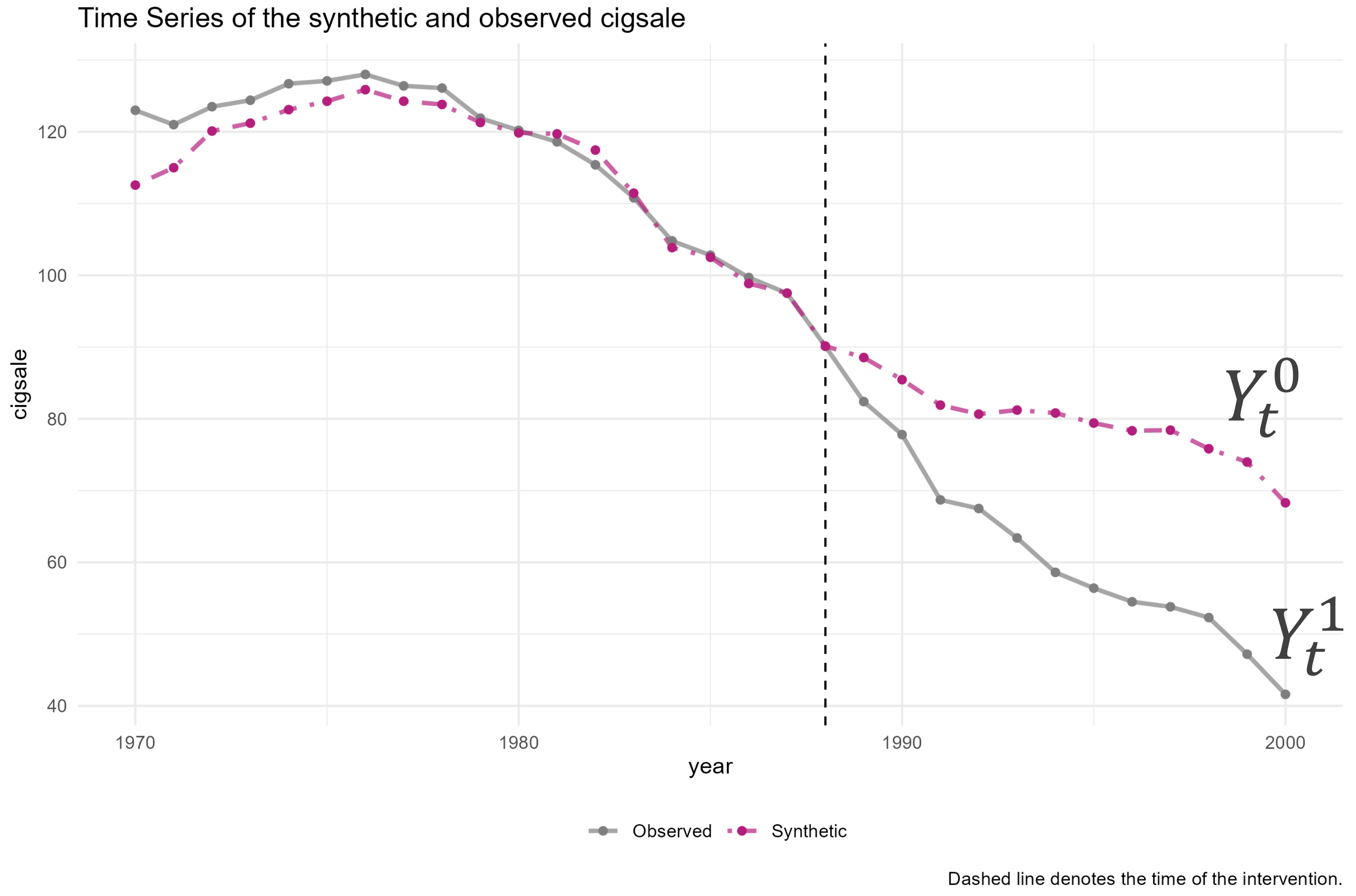
$$\widehat{CE_t} = Y_t^1 - \widehat{Y_t^0}$$

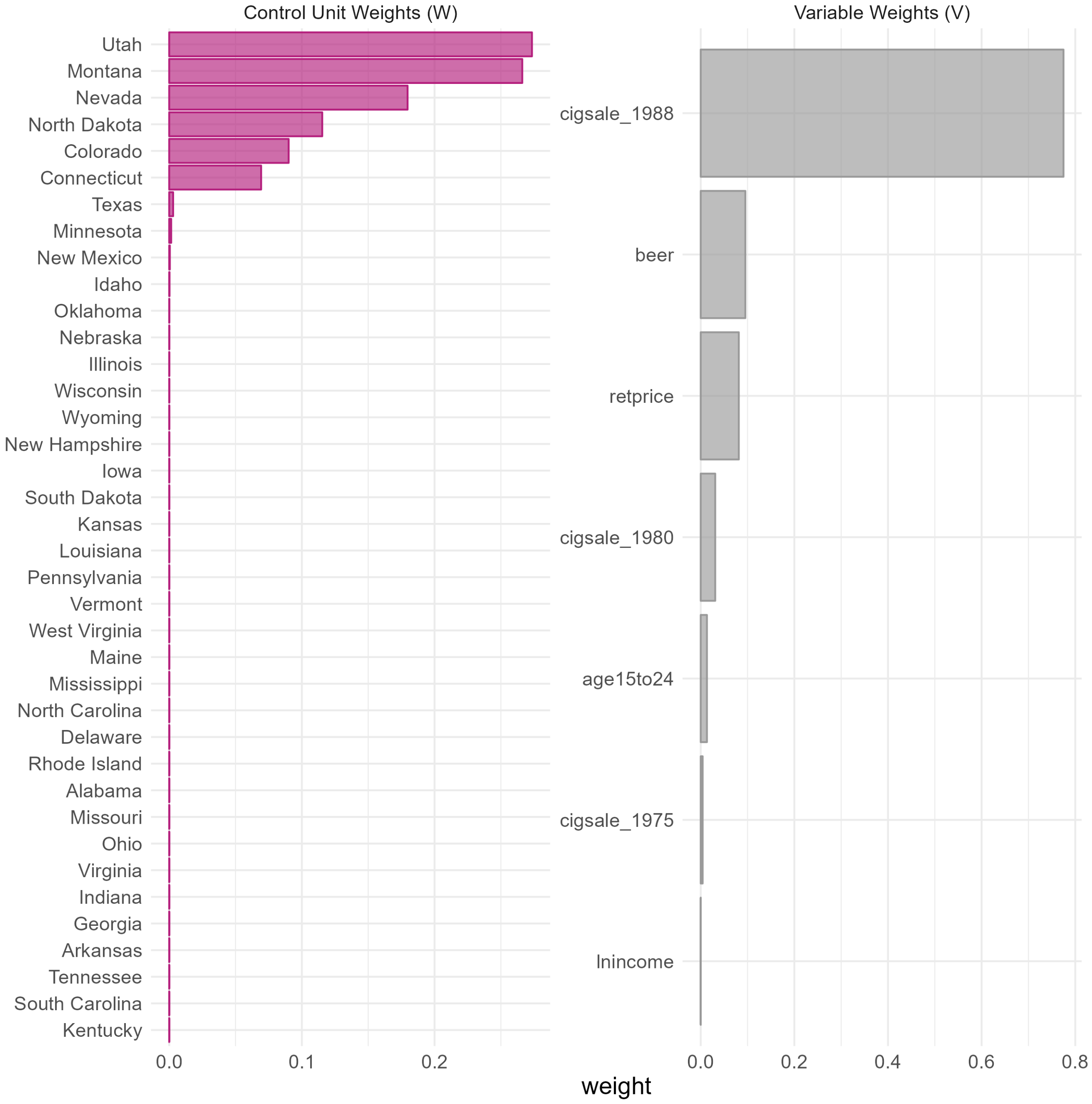


Time Series of the synthetic and observed cigsale
cigsale
year
120
100
80
60
40
1970
1980
1990
2000
Y_t^0
Y_t^1
Observed
Synthetic
Dashed line denotes the time of the intervention.

Estimating weights

- Choose weights such that the synthetic control **looks like** the treated unit
- Use only pre-intervention data for this
- Weights should be positive and sum to one
  Interpolation constraint / convex hull


Time Series of the synthetic and observed cigsale
cigsale
120
100
80
60
40
1970
1980
1990
2000
year
Y_t^0
Y_t^1
Observed
Synthetic
Dashed line denotes the time of the intervention.


Control Unit Weights (W)
Variable Weights (V)
Utah
Montana
Nevada
North Dakota
Colorado
Connecticut
Texas
Minnesota
New Mexico
Idaho
Oklahoma
Nebraska
Illinois
Wisconsin
Wyoming
New Hampshire
Iowa
South Dakota
Kansas
Louisiana
Pennsylvania
Vermont
West Virginia
Maine
Mississippi
North Carolina
Delaware
Rhode Island
Alabama
Missouri
Ohio
Virginia
Indiana
Georgia
Arkansas
Tennessee
South Carolina
Kentucky
cigsale_1988
beer
retprice
cigsale_1980
age15to24
cigsale_1975
lnincome
0.0
0.1
0.2
0.0
0.2
0.4
0.6
0.8
weight

Key Assumption

No interference / spillover:

The donor pool units do not receive any intervention effect

Example spillover effects

- Californians living near the border may buy their cigarettes in states across the border
- Other states may pass laws similar to on California

There are many choices

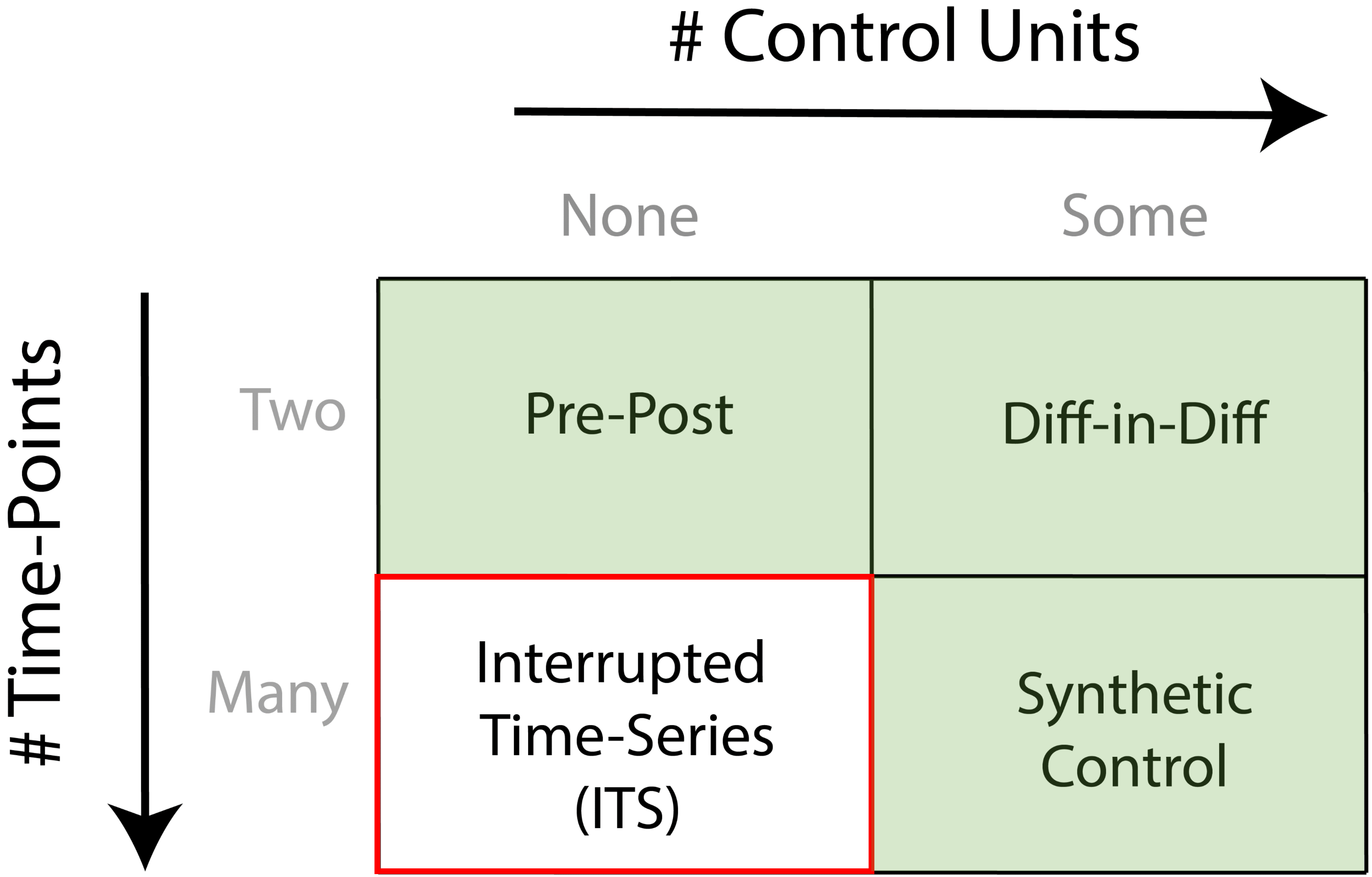
- Which units in the donor pool?
- Which control variables?
- What should my weights optimize?
- How many nonzero unit weights should I get?
- What settings do I give to the nonlinear optimizer?

“researcher degrees of freedom”

There are many choices

- These choices influence your causal estimate $\widehat{CE}_t$
- Think of your causal estimate as "conditional" on the "model" (choices)
- Investigate the impact of different choices through robustness checks / sensitivity analysis

Interrupted Time Series



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

Interrupted Time Series

Another popular method which is built on **predicting** the counterfactual

Here the prediction is based on a **time-forward** or **forecasting** model

- I.e. we use **past** pre-intervention data to **impute** the missing counterfactual at each point in time

Interrupted Time Series

$Time$	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	$\widehat{Y_6^0}$	2
7	3	1	$\widehat{Y_7^0}$	3
8	1	1	$\widehat{Y_8^0}$	1
…	…	…	…	…
T	2	1	$\widehat{Y_T^0}$	2

Train a forecasting model, e.g.

$$\widehat{Y}_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots \beta * Time$$

Make forecasts

$$\widehat{CE}_t = Y_t^1 - \widehat{Y_t^0}$$

Building a forecasting model

Much of the challenge of this approach is in choosing an appropriate **forecasting** model

These can be very simple or very complex, e.g.:

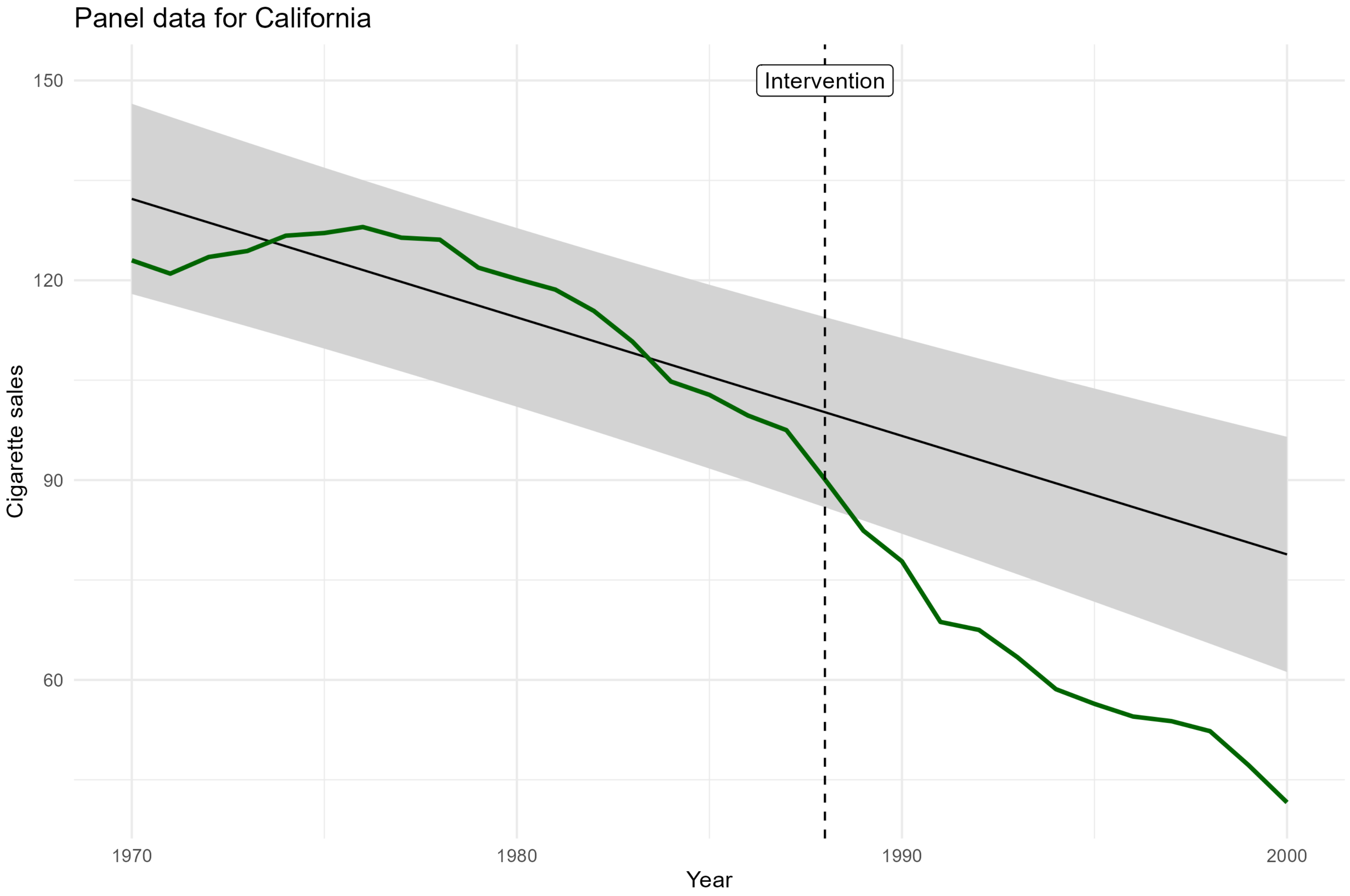
- We can forecast by fitting a **growth curve** which would model the overall time trend

$$Y_t = \beta_0 + \beta_1\, Time + e_t$$

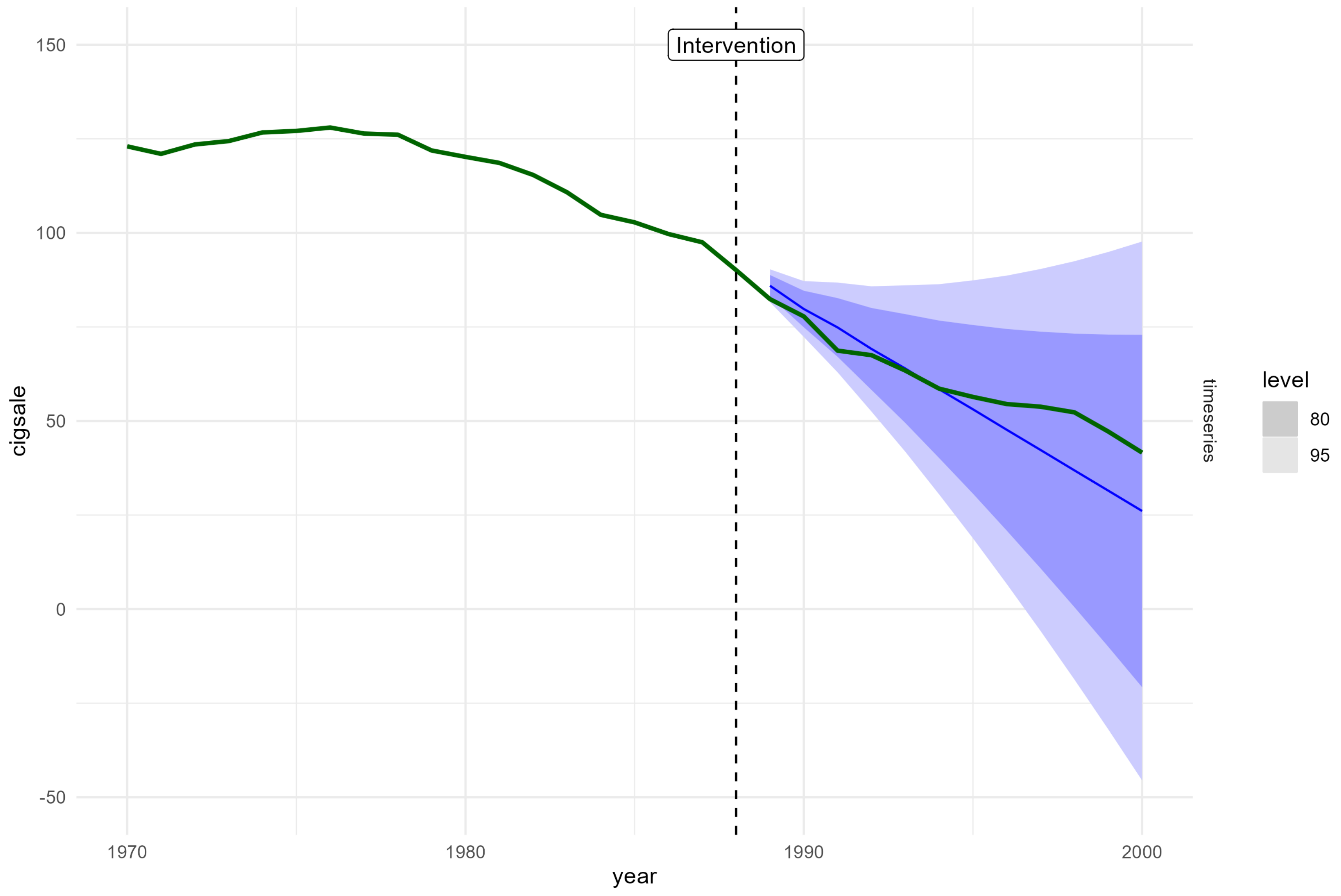
- We can forecast by using **time-series models** that model ***autocorrelation***

$$Y_t = \phi_1 Y_{t-1} + e_t \qquad Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t \qquad Y_t - Y_{t-1} = \gamma e_{t-1} + e_t$$

e.g. ARIMA models can account for autocorrelation and time trends



Panel data for California
Intervention
Cigarette sales
150
120
90
60
1970
1980
1990
2000
Year



Intervention
150
100
50
0
-50
cigsale
1970
1980
1990
2000
year
timeseries
level
80
95

Key Assumptions

Compared to pre-post, we do not assume away the trend, but instead model it directly

But our inferences about the causal effect are entirely dependent on being able to fit **an appropriate forecasting model**

- i.e. one that correctly captures the trend(s) and autocorrelation structures in the data

In practice, this may be **<u>very difficult</u>**

Discussion

Summary

Different policy evaluation methods have been developed across a variety of different disciplines

- Economics, Epidemiology/Public health, Social Sciences

These methods differ in terms of:

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

Using potential outcomes, we can more readily understand and compare these different methods

Control Units

Time-Points

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

So which method is best?

The answer **in part** depends on what type and amount of data you have

- But this is the **easy part**

The answer in practice depends on **domain knowledge**

- The **hard part** is to figure out which **assumptions** you need for causal inference and whether they are reasonable in your particular use case
- <u>It may simply not be possible in some cases!</u>
- E.g. DiD won't work if trends are not parallel; synthetic control won't work if there is interference between units (no matter how much data you have!)
- Often, methods which are "data hungry" can relax some assumptions, but:

<u>There is no free lunch!</u>

Useful References

Difference in Differences

Angrist, J. D., & Krueger, A. B. (1999). Empirical strategies in labor economics. In Handbook of labor economics (Vol. 3, pp. 1277-1366). Elsevier.

Angrist, J. D., & Pischke, J. S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton university press.

Caniglia, E. C., & Murray, E. J. (2020). Difference-in-difference in the time of cholera: a gentle introduction for epidemiologists. *Current epidemiology reports*, *7*, 203-211.

Interrupted Time Series

Bernal, J. L., Cummins, S., & Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: a tutorial. International journal of epidemiology, 46(1), 348-355.

Bernal, J.L, Cummins, S., & Gasparrini, A. (2019). Difference in difference, controlled interrupted time series and synthetic controls. International journal of epidemiology, 48(6), 2062-2063.

Useful References

Synthetic Control

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.

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

CausalImpact

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. The Annals of Applied Statistics, 247-274.

Linden, A. (2018). Combining synthetic controls and interrupted time series analysis to improve causal inference in program evaluation. Journal of evaluation in clinical practice, 24(2), 447-453.

http://google.github.io/CausalImpact/CausalImpact.html

Useful References

Synthetic DiD

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, *111*(12), 4088-4118.

More on Causal Policy Evaluation

Free online course materials made by Andrew Heiss

Program Evaluation for Public Service

https://evalf22.classes.andrewheiss.com/content/

Introduction to Causal Inference and Causal Data Science

Q&A

Oisin Ryan & Bas Penning de Vries