# Machine Learning for econometrics

## Event studies: Causal methods for pannel data

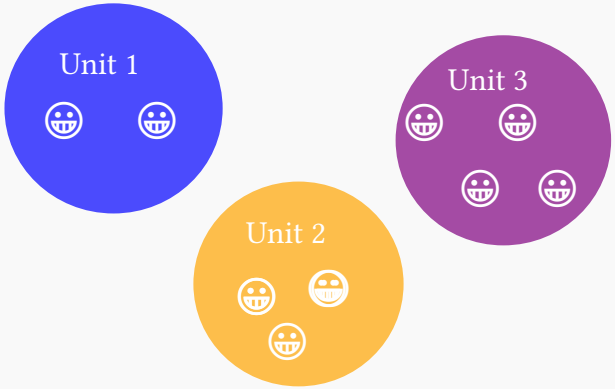
Matthieu Doutreligne

February, 11th, 2025

# Motivation

**Estimation of the effect of a treatment when data is:**

Aggregated: country-level data such as employment rate, GDP…

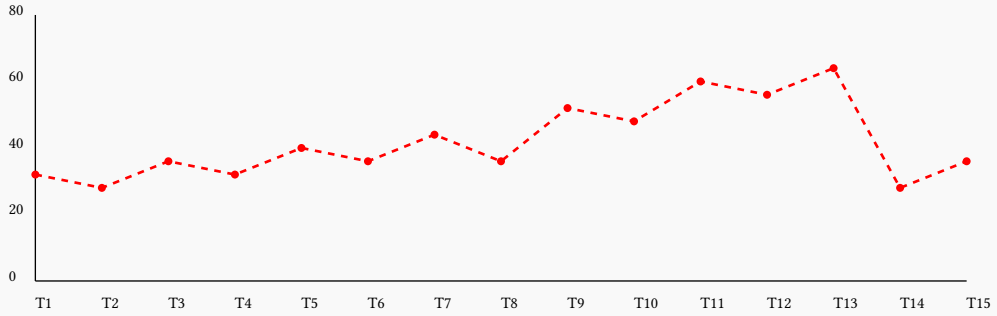


# Setup: event studies

## Estimation of the effect of a treatment when data is:

Aggregated: country-level data such as employment rate, GDP…

Longitudinal: multiple time periods (or repeated cross-sections)…



# Setup: event studies

## Estimation of the effect of a treatment when data is:

Aggregated: country-level data such as employment rate, GDP...

Longitudinal: multiple time periods (or repeated cross-sections)...

With multiple aggregated units: countries, firms, geographical regions...

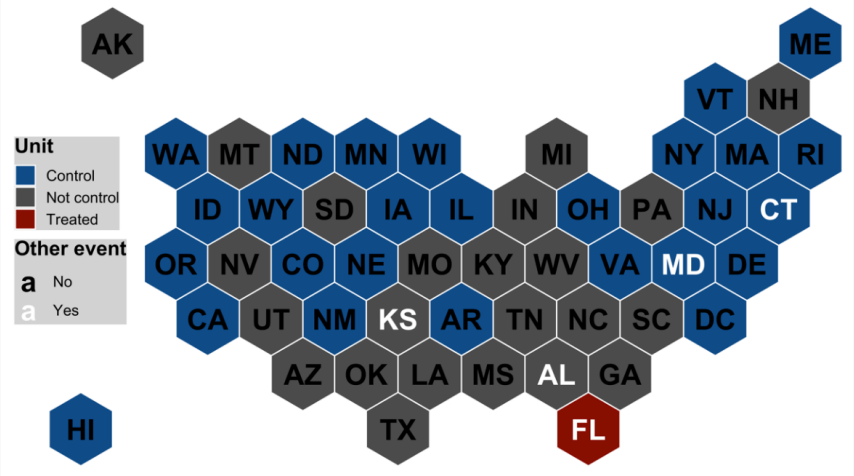



Figure from (Degli Esposti et al., 2020)

# Setup: event studies

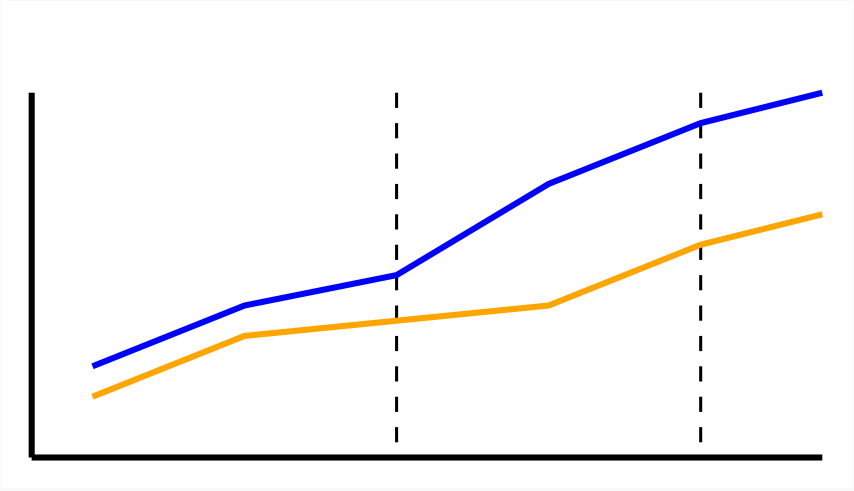
**Estimation of the effect of a treatment when data is:**

Aggregated: country-level data such as employment rate, GDP…

Longitudinal: multiple time periods (or repeated cross-sections)…

With multiple aggregated units: countries, firms, geographical regions…

Staggered adoption of the treatment: units adopt the policy/treatment at different times…



# Setup: event studies

**Estimation of the effect of a treatment when data is:**

Aggregated: country-level data such as employment rate, GDP…

Longitudinal: multiple time periods (or repeated cross-sections)…

With multiple aggregated units: countries, firms, geographical regions…

Staggered adoption of the treatment: units adopt the policy/treatment at different times…

This setup is known as

**Panel data, event studies, longitudinal data, time-series data.**

# Examples of event studies

**Archetypal questions**

- Did the new marketing campaign had an effect on the sales of a product?
- Did the new tax policy had an effect on the consumption of a specific product?
- Did the guidelines on the prescription of a specific drug had an effect on the practices?

# Examples of event studies

**Archetypal questions**

- Did the new marketing campaign had an effect on the sales of a product?
- Did the new tax policy had an effect on the consumption of a specific product?
- Did the guidelines on the prescription of a specific drug had an effect on the practices?

**Modern examples**

- What is the effect of the extension of Medicaid on mortality? (Miller et al., 2019)
- What is the effect of Europe's protected area policies (*Natura 2000*) on vegetation cover and on economic activity? (Grupp et al., 2023)
- Which policies achieved major carbon emission reductions? (Stechemesser et al., 2024)

# Setup: event studies are quasi-experiment

**Quasi-experiment**

A situation where the treatment is not randomly assigned by the researcher but by nature or society.
It should introduce *some* randomness in the treatment assignment: enforcing treatment exogeneity, ie. ignorability (ie. unconfoundedness).

# Setup: event studies are quasi-experiment

**Quasi-experiment**

A situation where the treatment is not randomly assigned by the researcher but by nature or society.
It should introduce *some* randomness in the treatment assignment: enforcing treatment exogeneity, ie. ignorability (ie. unconfoundedness).

## Other quasi-experiment designs

- Instrumental variables: a variable that is correlated with the treatment but not with the outcome.
- Regression discontinuity design: the treatment is assigned based on a threshold of a continuous variable.

# Table of contents

# Reminder on difference-in-differences

# Difference-in-differences

## History

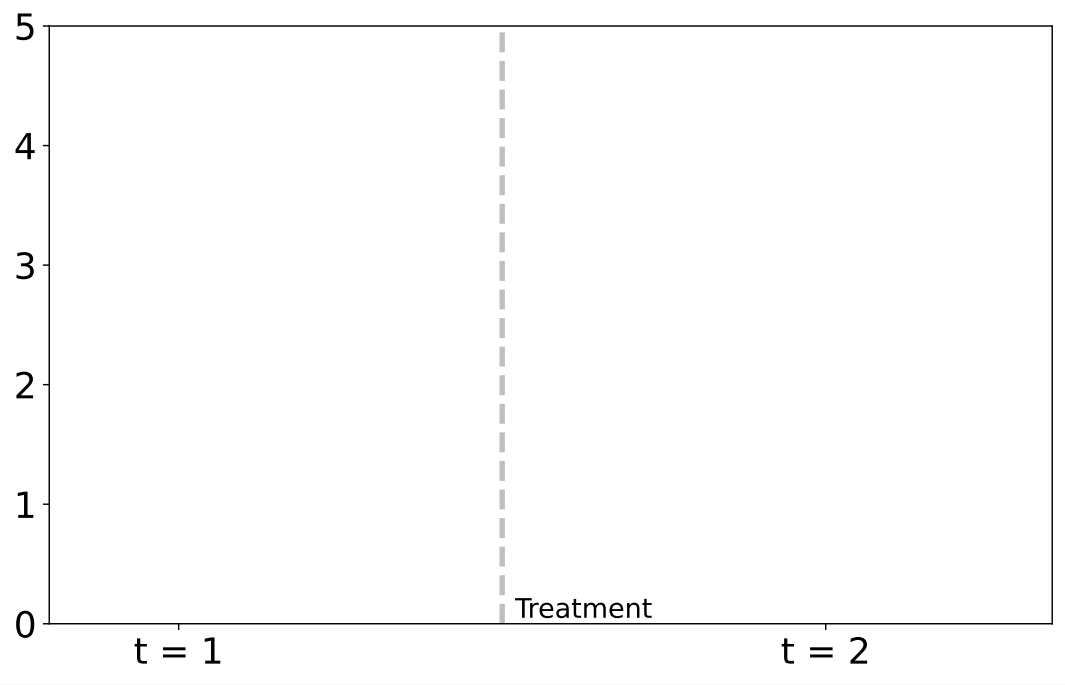
- First documented example (though not formalized): John Snow showing how cholera spread through the water in London (Snow, 1855)[1]
- Modern usage introduced formally by (Ashenfelter, 1978), applied to labor economics

## Idea

- Contrast the temporal effect of the treated unit with the control unit temporal effect:
- The difference between the two differences is the treatment effect

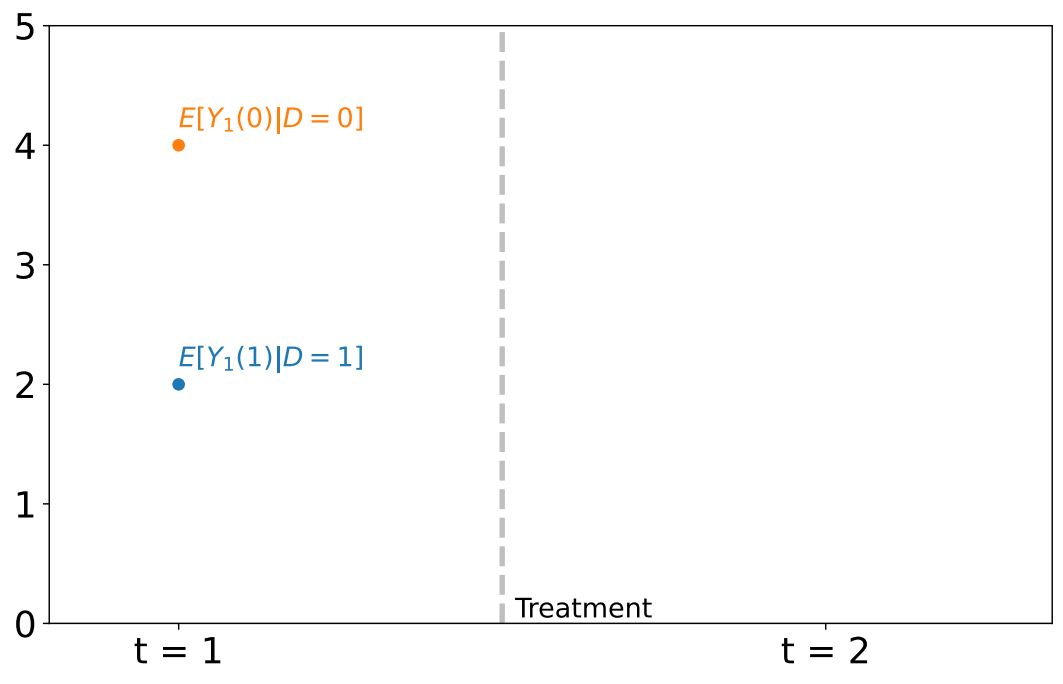
[1] Good description: https://mixtape.scunning.com/09-difference_in_differences#john-snows-cholera-hypothesis

## Two period of times: t=1, t=2



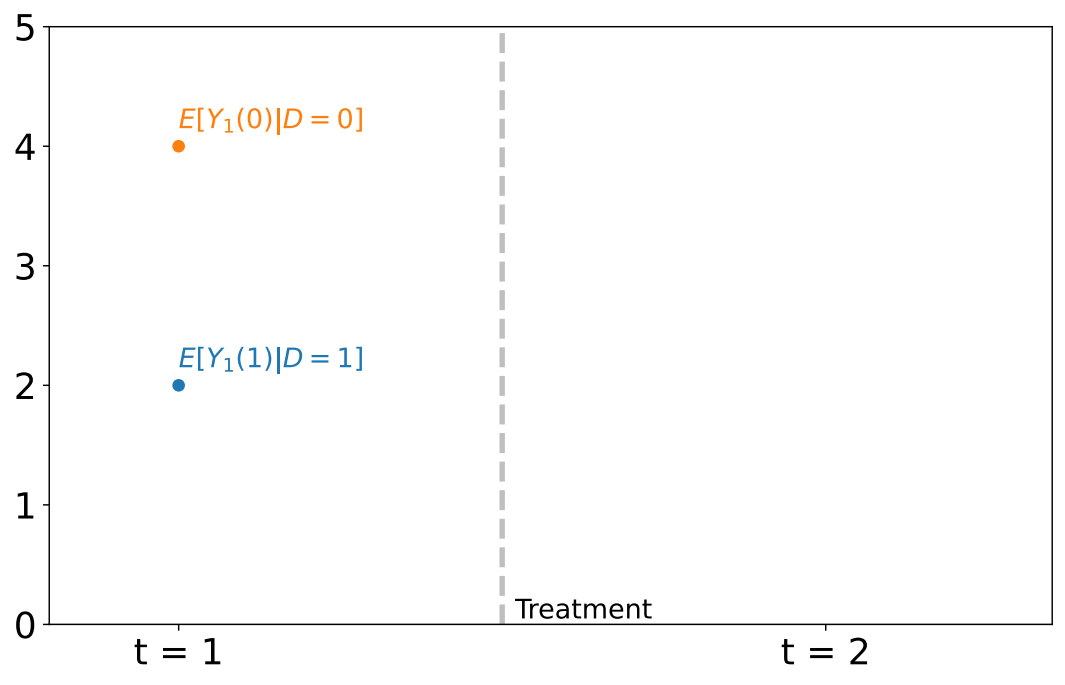
# Difference-in-differences framework

**Potential outcomes: $Y_t(d)$ where $d = \{0, 1\}$ is the treatment at period 2**



# Difference-in-differences framework

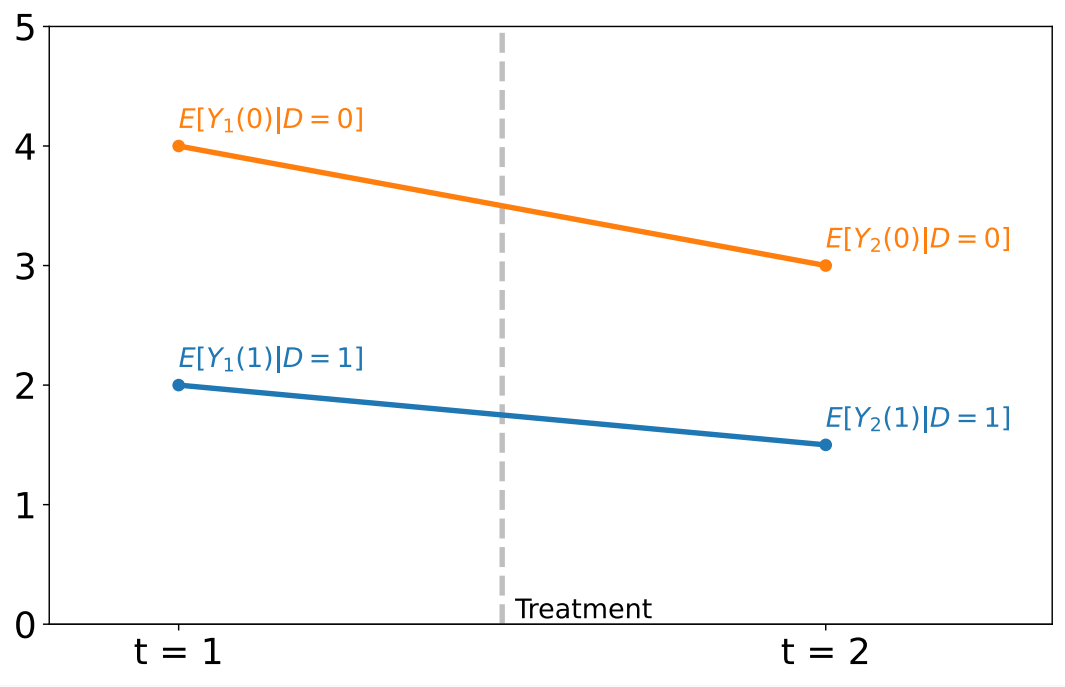
**Potential outcomes: $Y_t(d)$ where $d = \{0, 1\}$ is the treatment at period 2**




⚠️ $\mathbb{E}[Y_1(1)] = \mathbb{E}[Y_1(1) \mid D = 1]\mathbb{P}(D = 1) + \mathbb{E}[Y_1(1) \mid D = 0]\mathbb{P}(D = 0)$
but we only observe $\mathbb{E}[Y_1(1) \mid D = 1]$

# Difference-in-differences framework

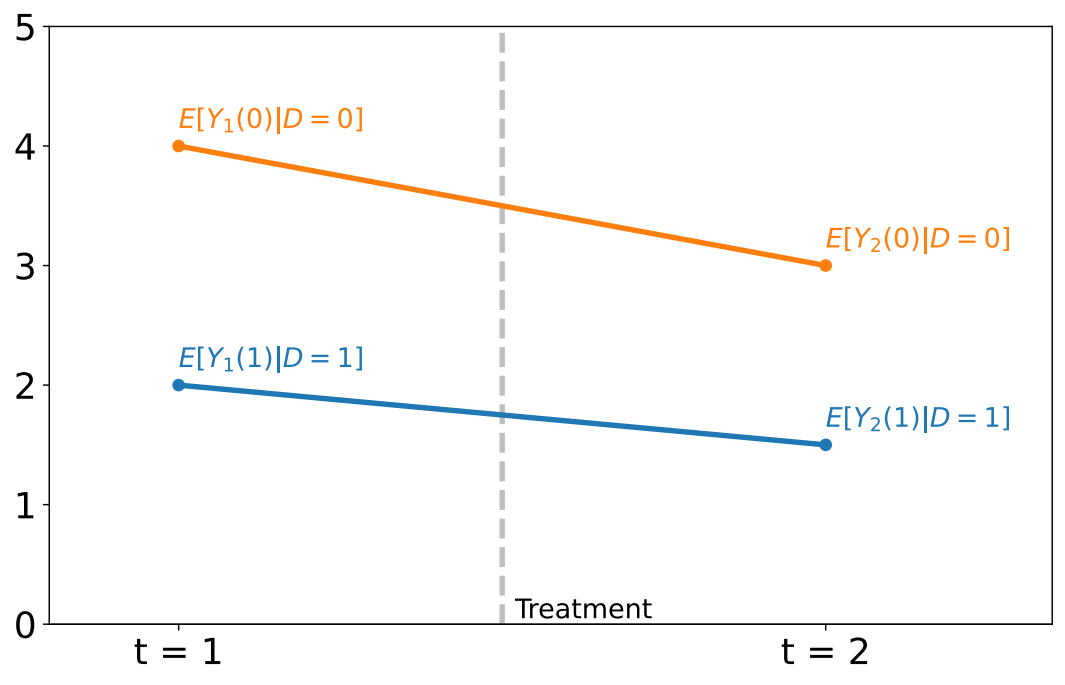
**Our target is the average treatment effect on the treated (ATT)**




$$\tau_{\text{ATT}} = \mathbb{E}[Y_2(1)|\ D=1] - \mathbb{E}[Y_2(0)|\ D=1]$$

# Difference-in-differences framework

**Our target is the average treatment effect on the treated (ATT)**

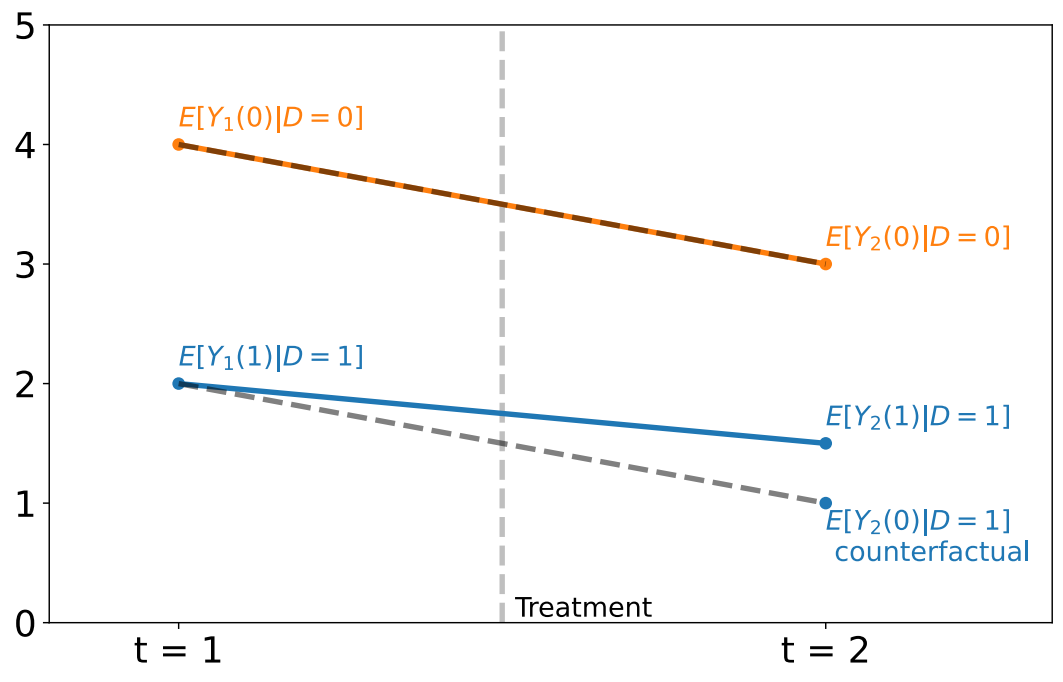



$$\tau_{\text{ATT}} = \underbrace{[Y_2(1)|\ D=1]}_{\text{treated outcome for t=2}} - \underbrace{\mathbb{E}[Y_2(0)|\ D=1]}_{\text{unobserved counterfactual}}$$

# Difference-in-differences framework

**First assumption, parallel trends**

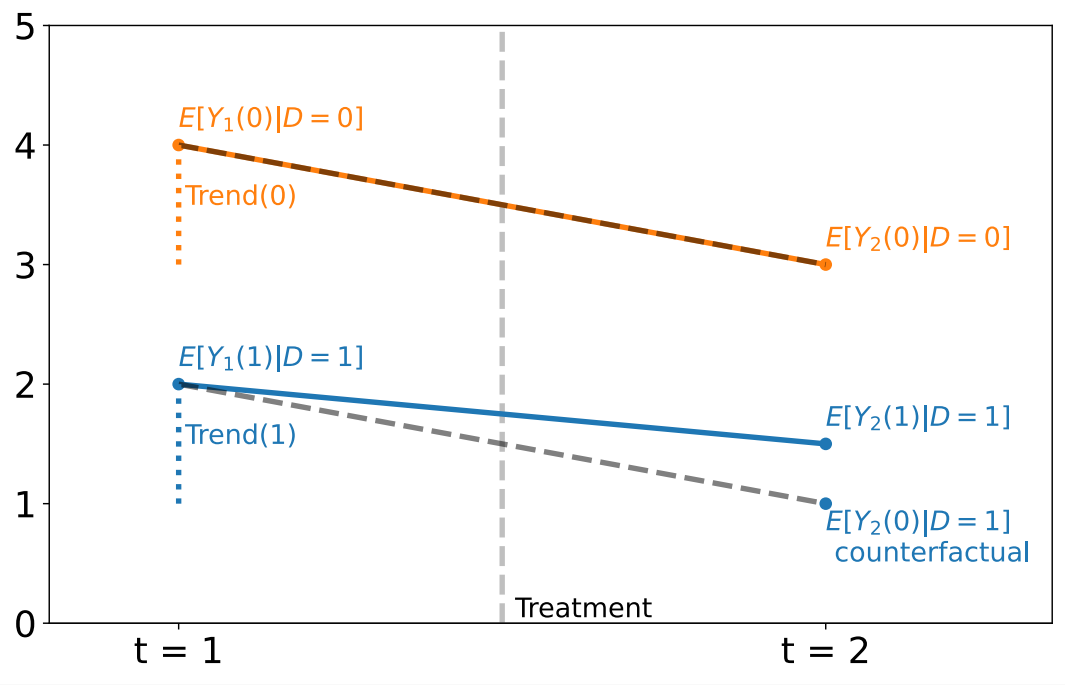
$$\mathbb{E}[Y_2(0) - Y_1(0) \mid D = 1] = \mathbb{E}[Y_2(0) - Y_1(0) \mid D = 0]$$



# Difference-in-differences framework

## First assumption, parallel trends

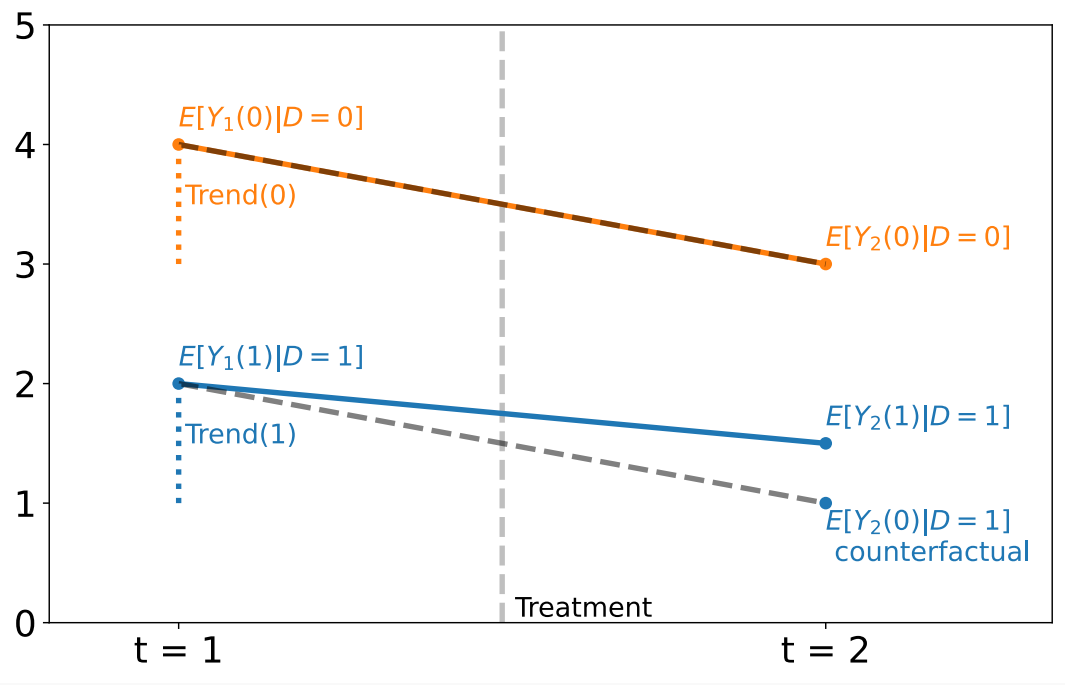
$$\underbrace{[Y_2(0) - Y_1(0) \mid D = 1]}_{\text{Trend}(1)} = \underbrace{\mathbb{E}[Y_2(0) - Y_1(0) \mid D = 0]}_{\text{Trend}(0)}$$



# Difference-in-differences framework

## First assumption, parallel trends

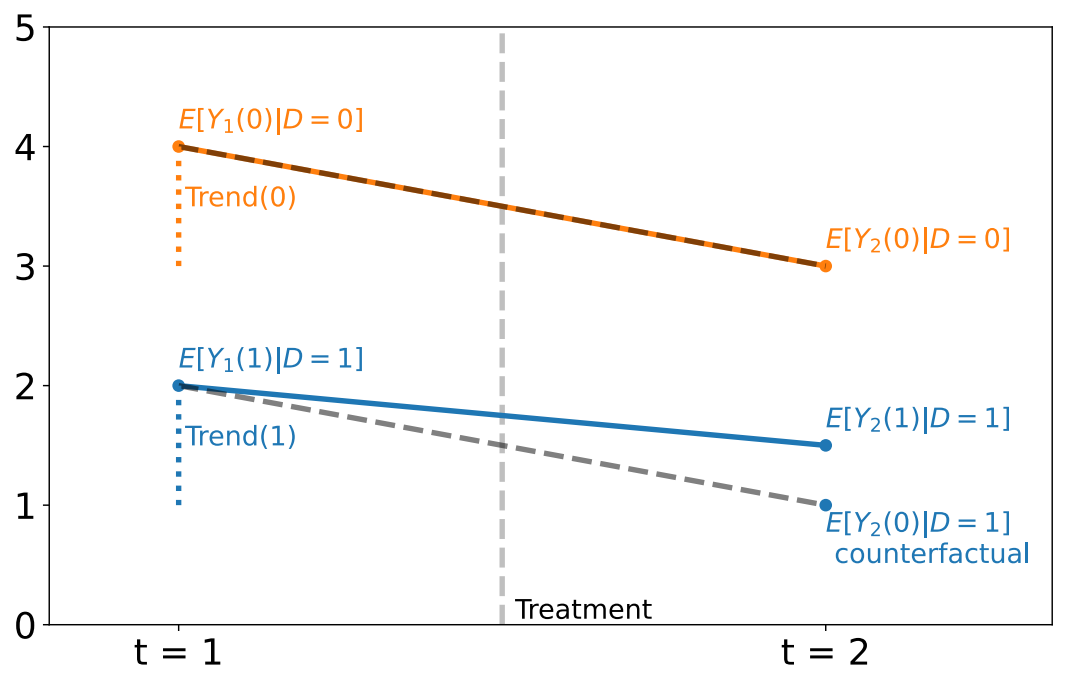
$$\mathbb{E}[Y_2(0) \mid D = 1] = \mathbb{E}[Y_1(0) \mid D = 1] + \mathbb{E}[Y_2(0) - Y_1(0) \mid D = 0]$$



# Difference-in-differences framework

## First assumption, parallel trends

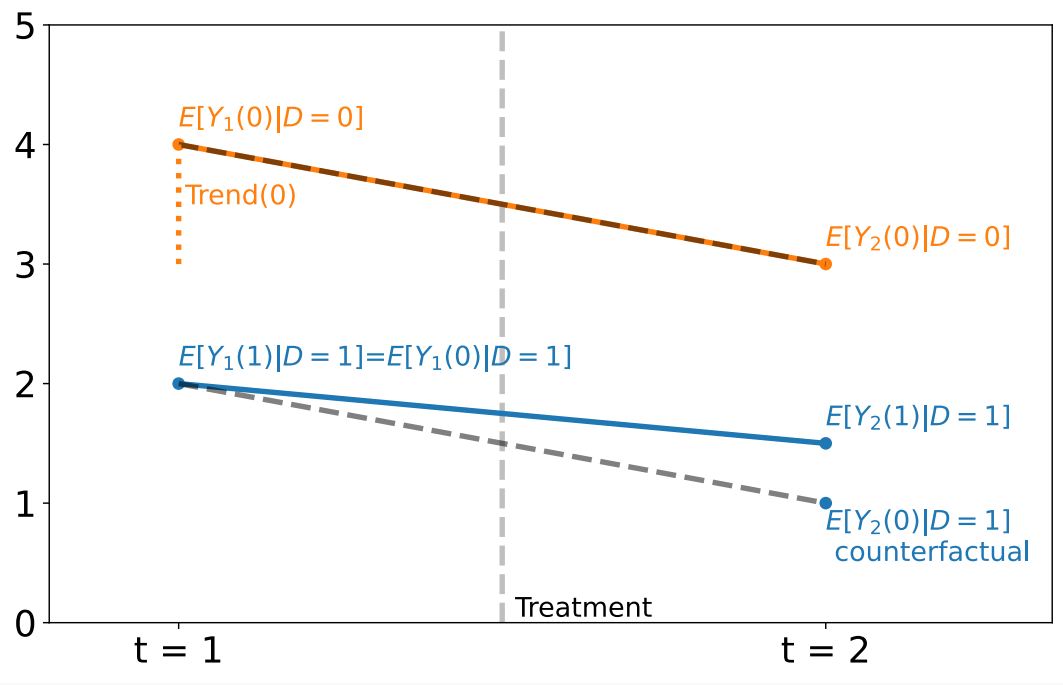
$$\mathbb{E}[Y_2(0) \mid D = 1] = \underbrace{[Y_1(0) \mid D = 1]}_{\text{unobserved counterfactual}} + \mathbb{E}[Y_2(0) - Y_1(0) \mid D = 0]$$



# Difference-in-differences framework

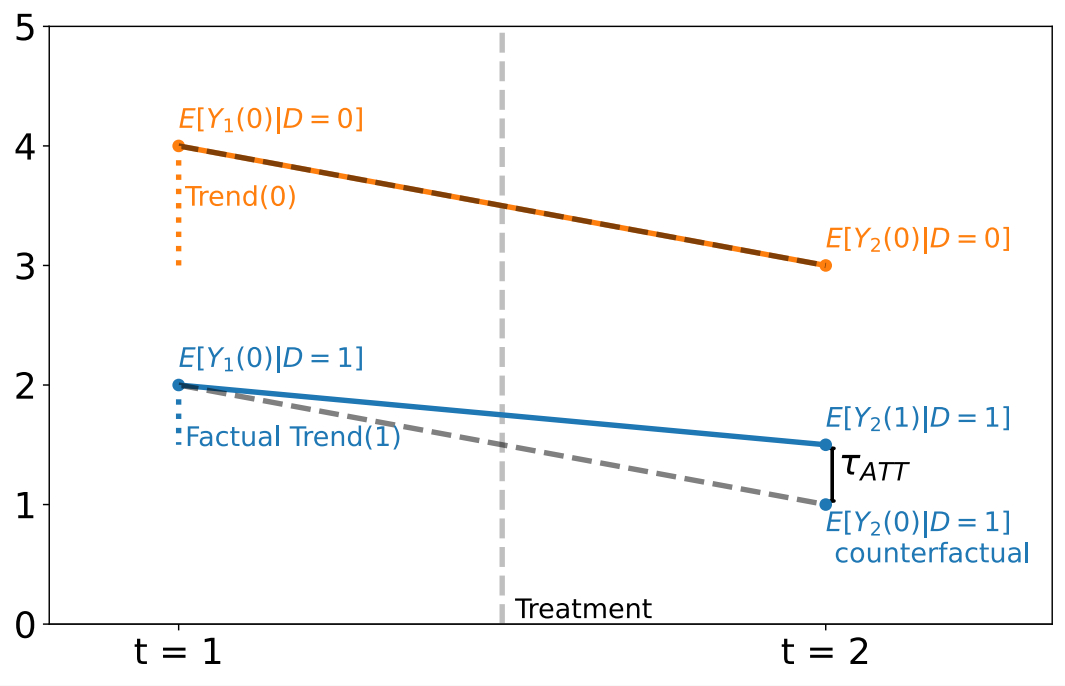
**Second assumption, no anticipation of the treatment**

$$E[Y_1(1)|D=1] = E[Y_1(0)|D=1]$$



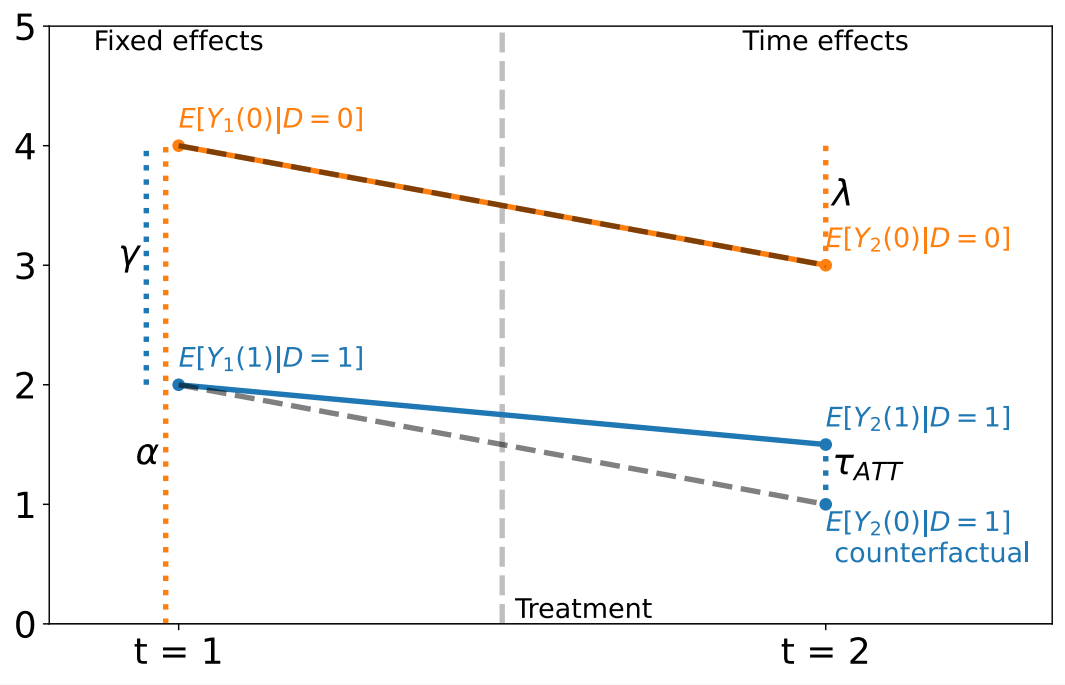
# Difference-in-differences framework: identification of ATT

$$\tau_{\text{ATT}} = \mathbb{E}[Y_2(1)|\ D=1] - \mathbb{E}[Y_2(0)|\ D=1]$$
$$= \underbrace{\mathbb{E}[Y_2(1)|\ D=1] - \mathbb{E}[Y_1(0)|D=1]}_{\text{Factual Trend}(1)} - \underbrace{\mathbb{E}[Y_2(0)|D=0] - \mathbb{E}[Y_1(0)|D=0]}_{\text{Trend}(0)}$$



# Estimation: link with two way fixed effect (TWFE)

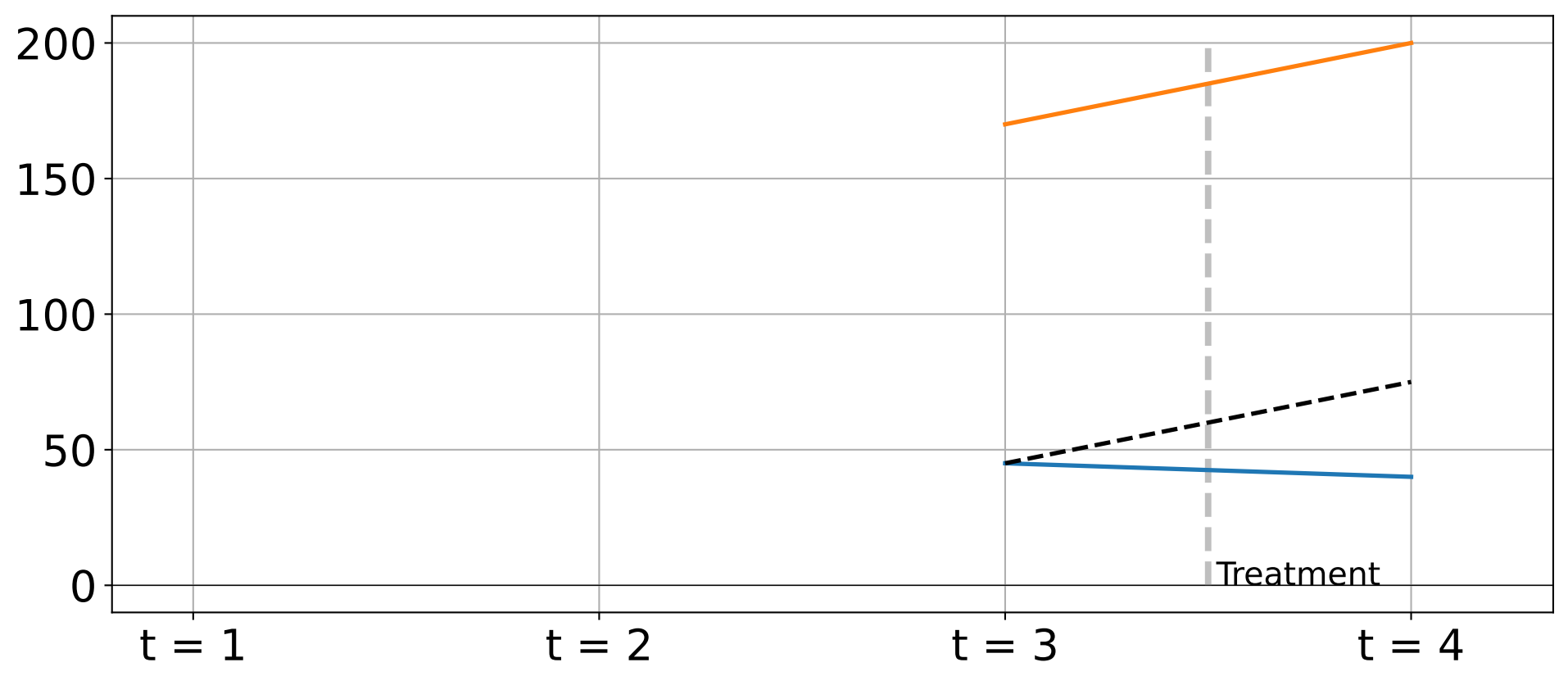
$$Y = \alpha + \gamma D + \lambda \mathbb{1}(t = 2) + \tau_{\text{ATT}} D \mathbb{1}(t = 2)$$




⚠️ Mechanic link: works only under parallel trends and no anticipation assumptions.

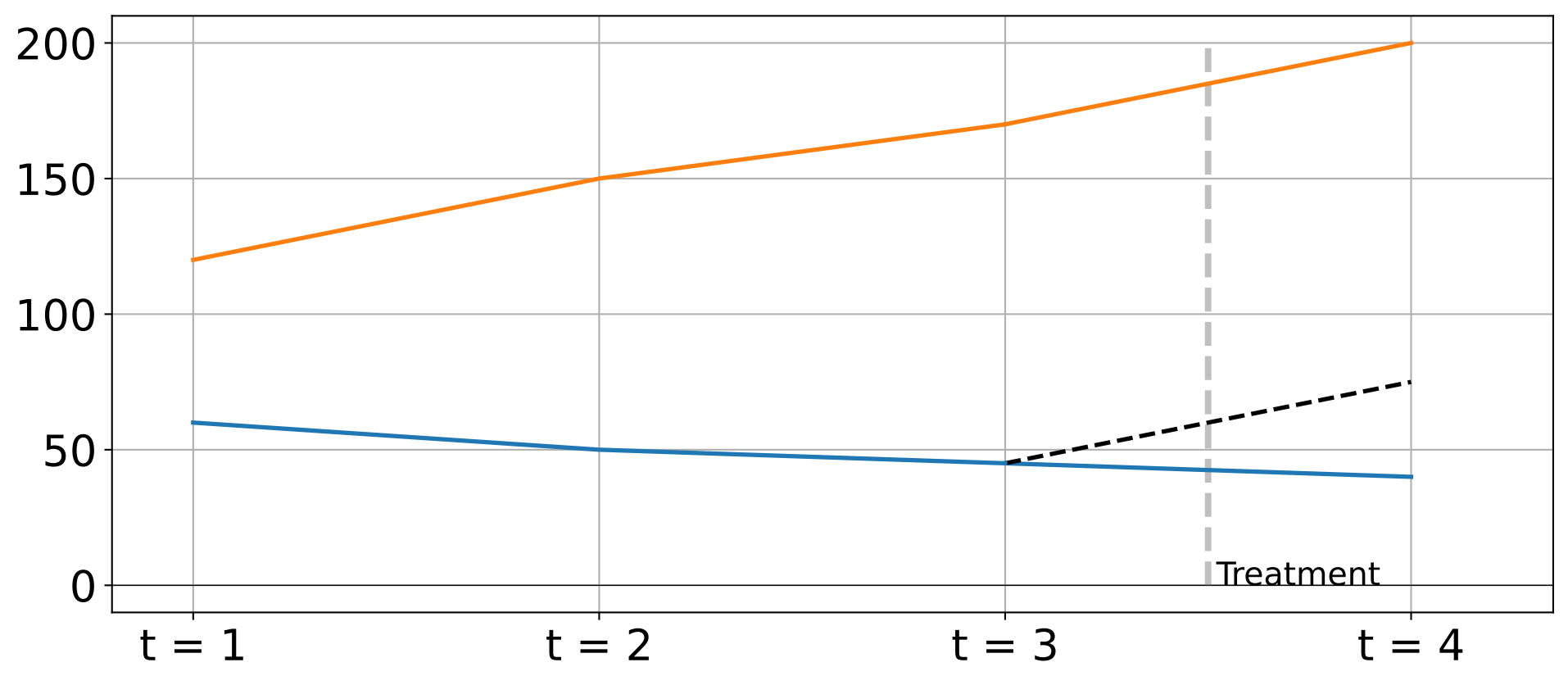
# Failure of the parallel trend assumption

## Seems like the treatment decreases the outcome!



# Failure of the parallel trend assumption

## Oups...



# DID estimator for more than two time units

**Target estimand: sample average treatment effect on the treated (SATT)**

$$\tau_{\text{SATT}} = \frac{1}{|\{i:D_i=1\}|}\sum_{i:D_i=1}\frac{1}{T-H}\sum_{t=H+1}^{T} Y_{it}(1) - Y_{it}(0)$$

**DID estimator**

$$\widehat{\tau_{\text{DID}}} = \frac{1}{|\{i:D_i=1\}|}\sum_{i:D_i=1}\left[\frac{1}{T-H}\sum_{t=H+1}^{T} Y_{it} - \frac{1}{H}\sum_{t=1}^{H} Y_{it}\right] - \frac{1}{|\{i:D_i=0\}|}\sum_{i:D_i=0}\left[\frac{1}{T-H}\sum_{t=H+1}^{T} Y_{it} - \frac{1}{H}\sum_{t=1}^{H} Y_{it}\right]$$

Assumption

**No anticipation of the treatment:** $Y_{it}(0) = Y_{it}(1) \forall t = 1, ..., H.$

**Parallel trend:** $\mathbb{E}[Y_{it}(0, \infty) - Y_{i1}(0, \infty)] = \beta_t, t = 2, ..., T.$

See (Wager, 2024) for a clear proof of consistancy.

# DID: Take-away

## Pros

- Extremely common in economics and quite simple to implement.
- Can be extended to (Wager, 2024)
  - more than two time periods: exact same formulation
  - staggered adoption of the treatment: a bit more complex

## Cons

- Very strong assumptions: parallel trends and no anticipation.
- Does not account for heterogeneity of treatment effect over time (De Chaisemartin & d'Haultfoeuille, 2020).

# DID: Take-away

## Pros

- Extremely common in economics and quite simple to implement.
- Can be extended to (Wager, 2024)
  - more than two time periods: exact same formulation
  - staggered adoption of the treatment: a bit more complex

## Cons

- Very strong assumptions: parallel trends and no anticipation.
- Does not account for heterogeneity of treatment effect over time (De Chaisemartin & d'Haultfoeuille, 2020).

## Can we do better: ie. robust to the parallel trend assumption?

# Synthetic controls

# Synthetic controls

## References

Introduced by (Abadie & Gardeazabal, 2003) and (Abadie et al., 2010).

Quick introduction in (Bonander et al., 2021), technical overiew in (Abadie, 2021),

# Synthetic controls

**References**

Introduced by (Abadie & Gardeazabal, 2003) and (Abadie et al., 2010).

Quick introduction in (Bonander et al., 2021), technical overiew in (Abadie, 2021),

> The most important innovation in the policy evaluation literature in the last few years
>
> — (Athey & Imbens, 2017)

# Synthetic controls

**References**

Introduced by (Abadie & Gardeazabal, 2003) and (Abadie et al., 2010).

Quick introduction in (Bonander et al., 2021), technical overiew in (Abadie, 2021),

> The most important innovation in the policy evaluation literature in the last few years
>
> — (Athey & Imbens, 2017)

**Idea**

Find a weighted average of controls that predicts well the treated unit outcome before treatment.

## Example

What is the effect of tobacco tax on cigarettes sales? (Abadie et al., 2010)

# Examples of application of synthetic controls to epidemiology

- What is the effect of taxes on sugar-based product consumption (Puig-Codina et al., 2021)

# Synthetic control example: California's Proposition 99 (Abadie et al., 2010)

## Context

1988: 25-cent tax per pack of cigarettes, ban of on cigarette vending machines in public areas accessible by juveniles, and a ban on the individual sale of single cigarettes.

# Synthetic control example: California's Proposition 99 (Abadie et al., 2010)

## Context

1988: 25-cent tax per pack of cigarettes, ban of on cigarette vending machines in public areas accessible by juveniles, and a ban on the individual sale of single cigarettes.

## Setup

**Outcome, $Y_{j,t}$: cigarette sales per capita**

# Synthetic control example: California's Proposition 99 (Abadie et al., 2010)

## Context

1988: 25-cent tax per pack of cigarettes, ban of on cigarette vending machines in public areas accessible by juveniles, and a ban on the individual sale of single cigarettes.

## Setup

**Outcome, $Y_{j,t}$: cigarette sales per capita**

**Treated unit, $j = 1$: California as from 1988**

# Synthetic control example: California's Proposition 99 (Abadie et al., 2010)

## Context

1988: 25-cent tax per pack of cigarettes, ban of on cigarette vending machines in public areas accessible by juveniles, and a ban on the individual sale of single cigarettes.

## Setup

**Outcome, $Y_{j,t}$: cigarette sales per capita**

**Treated unit, $j = 1$: California as from 1988**

**Control units, $j \in \{2, ..J\}$: 39 other US states without similar policies**

# Synthetic control example: California's Proposition 99 (Abadie et al., 2010)

## Context

1988: 25-cent tax per pack of cigarettes, ban of on cigarette vending machines in public areas accessible by juveniles, and a ban on the individual sale of single cigarettes.

## Setup

**Outcome, $Y_{j,t}$: cigarette sales per capita**

**Treated unit, $j = 1$: California as from 1988**

**Control units, $j \in \{2, ..J\}$: 39 other US states without similar policies**

**Time period: $t \in \{1, ..T\} = \{1970, ..2000\}$ and treatment time $T_0 = 1988$**

# Synthetic control example: California's Proposition 99 (Abadie et al., 2010)

## Context

1988: 25-cent tax per pack of cigarettes, ban of on cigarette vending machines in public areas accessible by juveniles, and a ban on the individual sale of single cigarettes.

## Setup

**Outcome, $Y_{j,t}$: cigarette sales per capita**

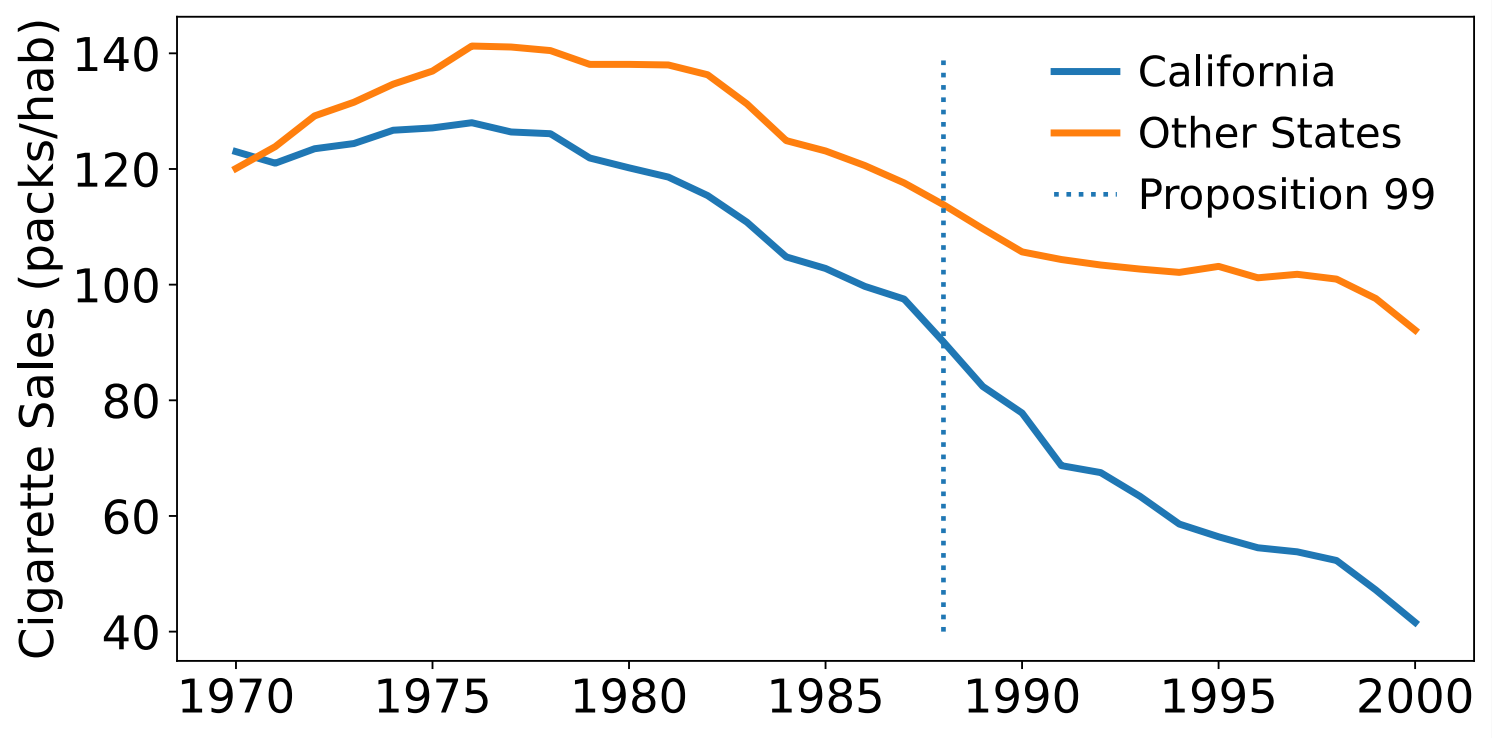
**Treated unit, $j = 1$: California as from 1988**

**Control units, $j \in \{2, ..J\}$: 39 other US states without similar policies**

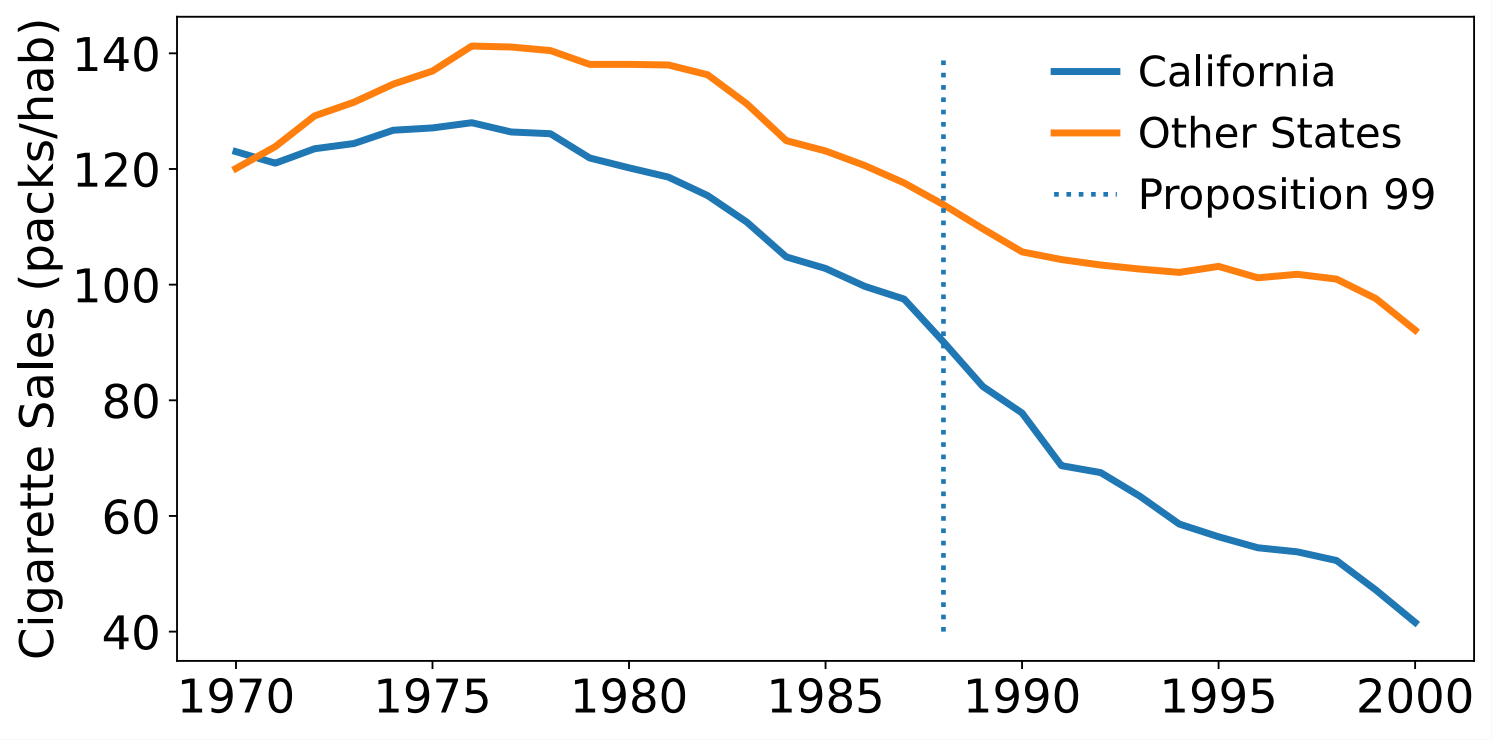
**Time period: $t \in \{1, ..T\} = \{1970, ..2000\}$ and treatment time $T_0 = 1988$**

**Covariates $X_{j,t}$: cigarette price, previous cigarette sales.**

# Synthetic control example: plot the data

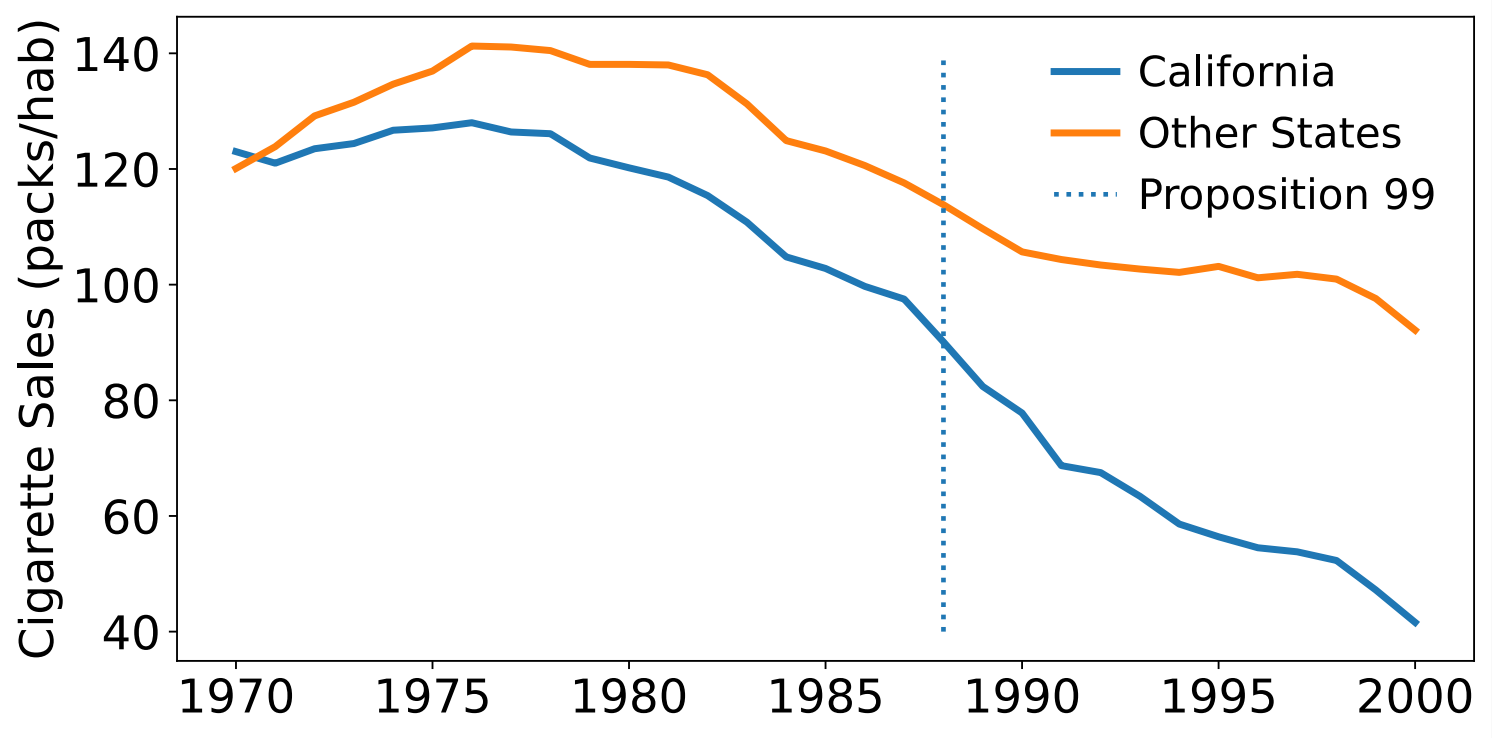


# Synthetic control example: plot the data




😲 Decrease in cigarette sales in California.

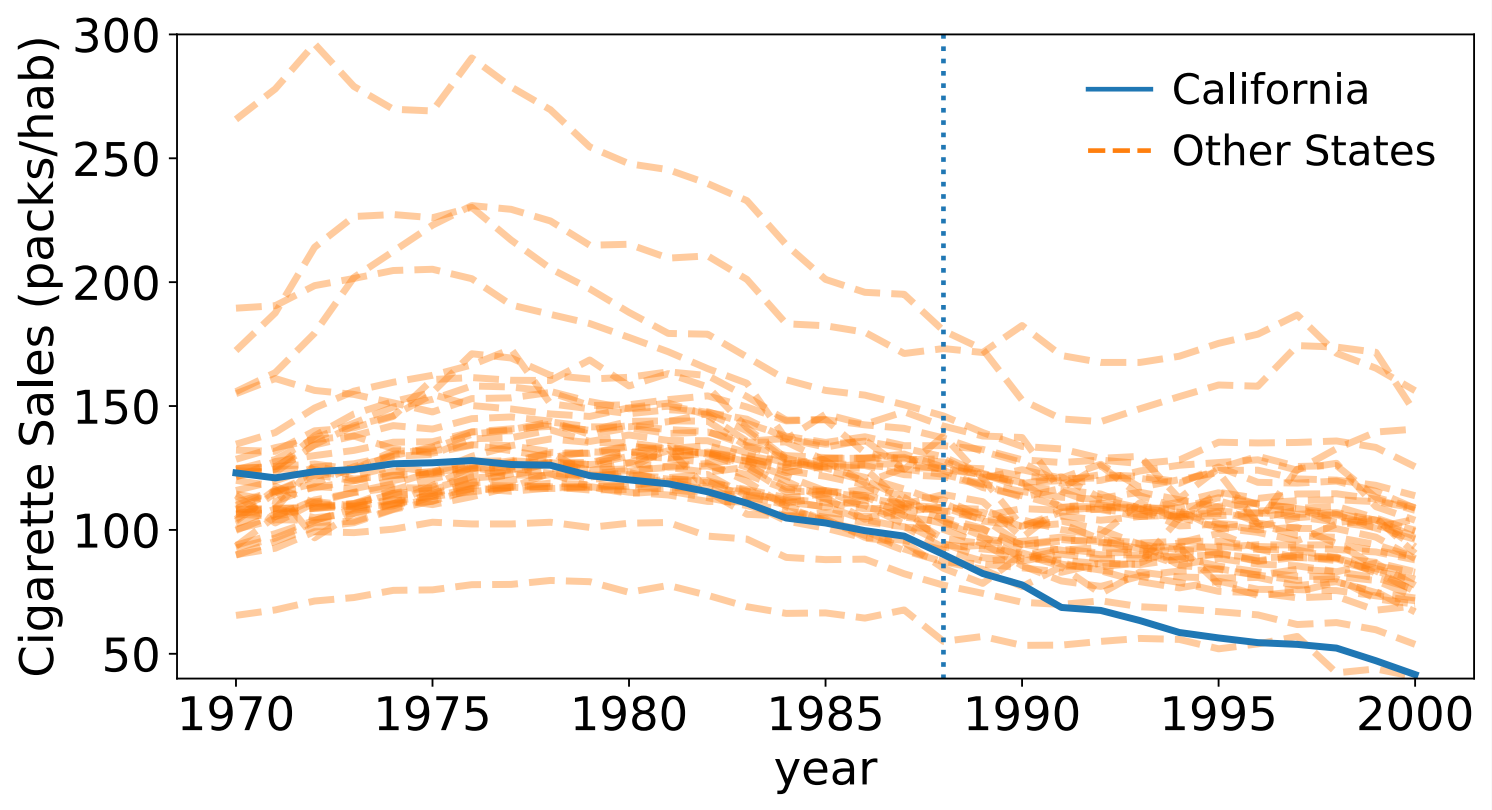
# Synthetic control example: plot the data



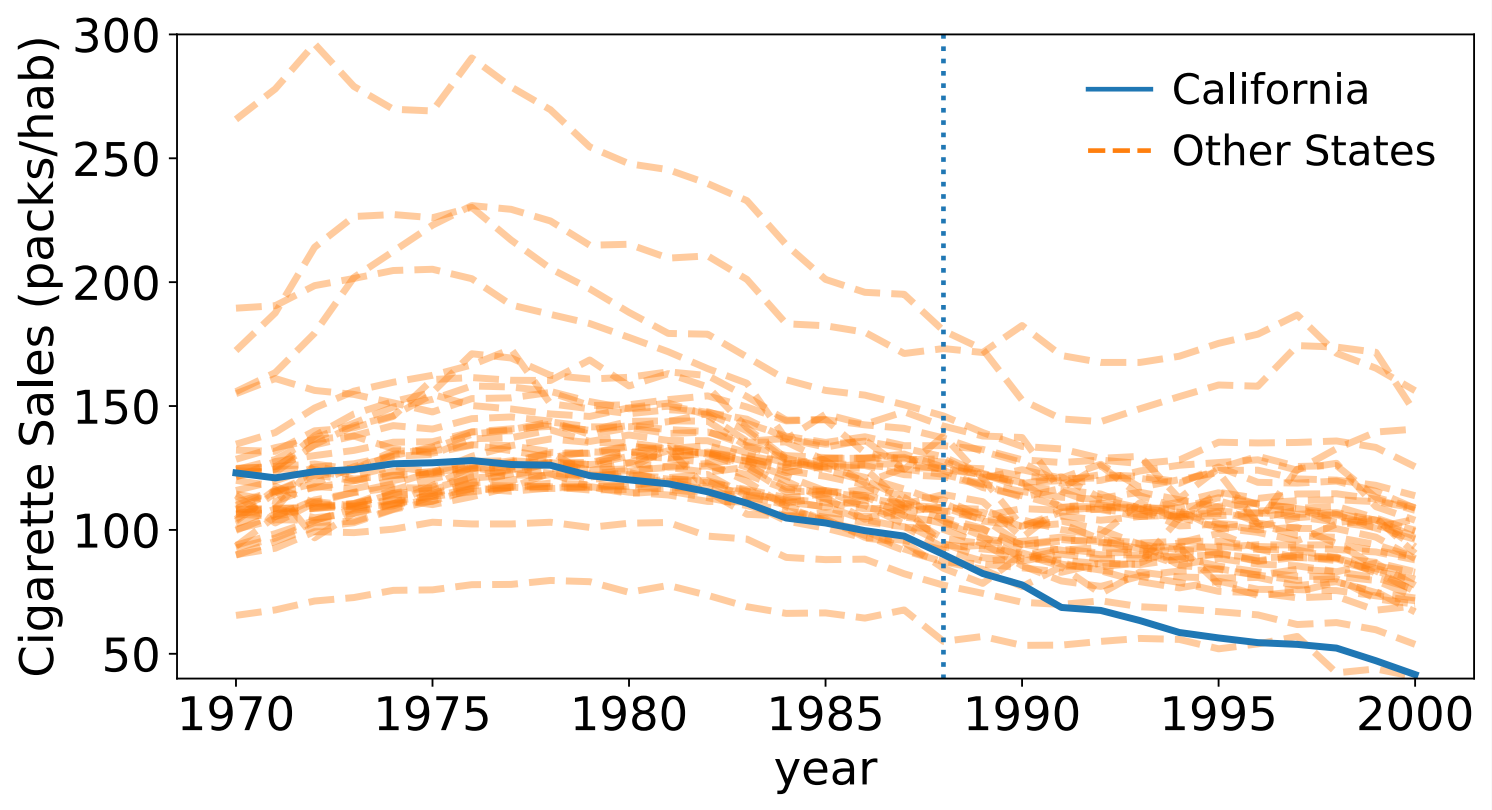

😲 Decrease in cigarette sales in California.

🤔 Decrease began before the treatment and occured also for other states.

# Synthetic control example: plot the data



# Synthetic control example: plot the data

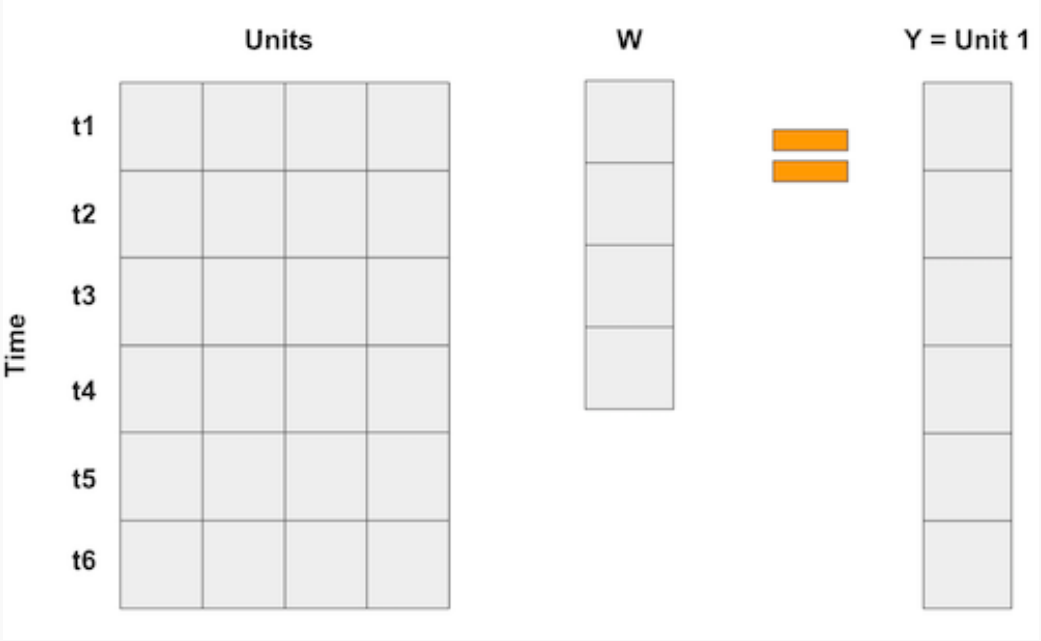



💡 Force parallel trends: Find a weighted average of other states that predicts well the pre-treatment trend of California (before $T_0 = 1988$).

# Synthetic control as weighted average of control outcomes

Build a predictor for $Y_{1,t}$ (California):

$$\hat{Y}_{1,t} = \sum_{j=2}^{n_0+1} \hat{w}_j Y_{j,t}$$



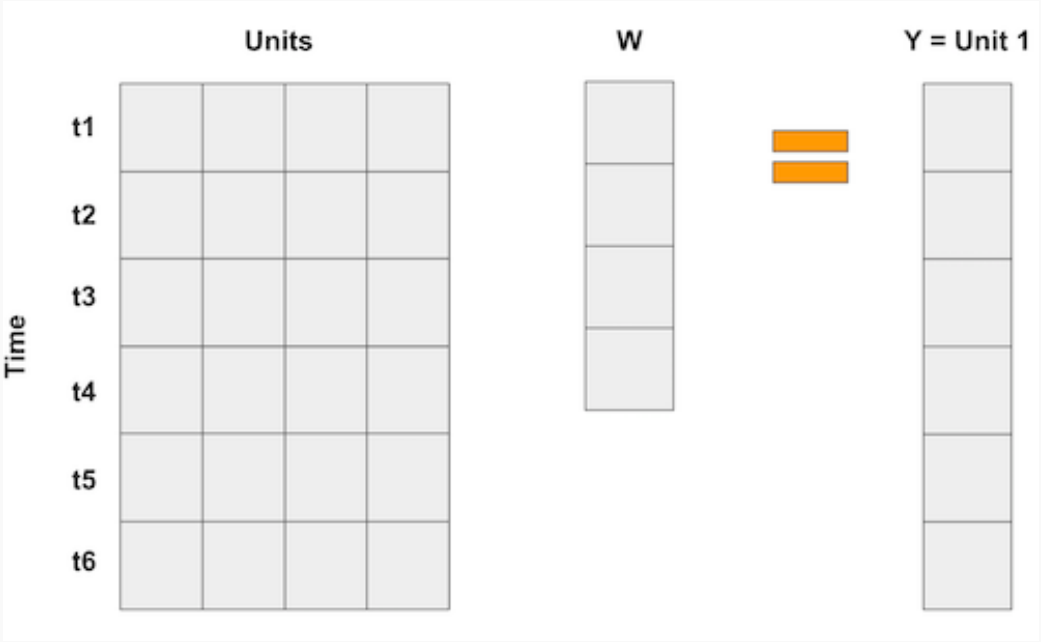
# Synthetic control as weighted average of control outcomes

Build a predictor for $Y_{1,t}$ (California):

$$\hat{Y}_{1,t} = \sum_{j=2}^{n_0+1} \hat{w}_j Y_{j,t}$$

🤔 How to choose the weights?

Minimize some distance between the treated and the controls.



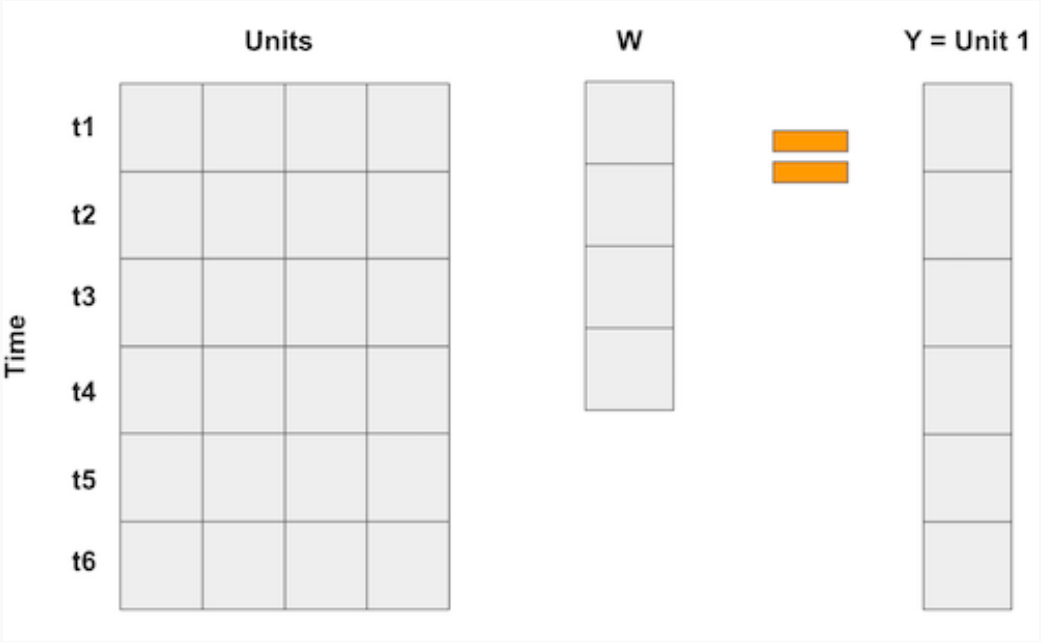
# Synthetic control as weighted average of control outcomes

Build a predictor for $Y_{1,t}$ (California):

$$\hat{Y}_{1,t} = \sum_{j=2}^{n_0+1} \hat{w}_j Y_{j,t}$$

🤔 How to choose the weights?

Minimize some distance between the treated and the controls.



# Synthetic control as weighted average of control outcomes

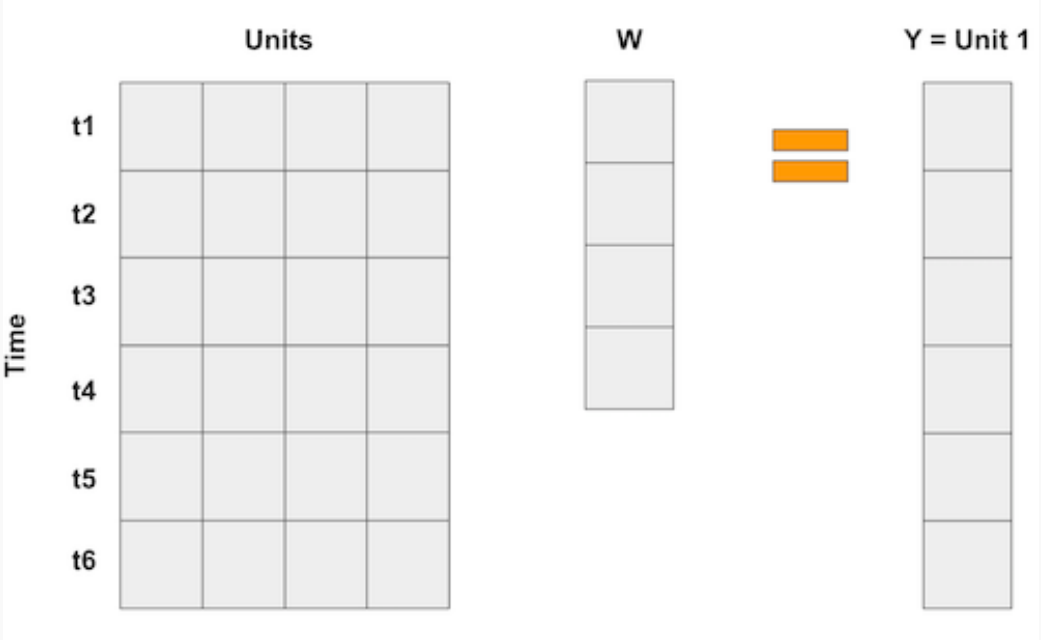
Build a predictor for $Y_{1,t}$ (California):

$$\hat{Y}_{1,t} = \sum_{j=2}^{n_0+1} \hat{w}_j Y_{j,t}$$

🤔 How to choose the weights?

Minimize some distance between the treated and the controls.

🤓 This is called a balancing estimator: kind of Inverse Probability Weighting (Wager, 2024, chapter 7)



# Synthetic controls: minimization problem

**Characteristics**

Pre-treatment characteristics concatenate pre-treatment outcomes and other pre-treatment predictors $Z_1$ eg. cigarette prices:

$$X_{\text{treat}} = X_1 = \begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ .. \\ Y_{1,T_0} \\ Z_1 \end{pmatrix} \in R^{p \times 1}$$

# Synthetic controls: minimization problem

**Characteristics**

Pre-treatment characteristics concatenate pre-treatment outcomes and other pre-treatment predictors $Z_1$ eg. cigarette prices:

$$X_{\text{treat}} = X_1 = \begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ .. \\ Y_{1,T_0} \\ Z_1 \end{pmatrix} \in R^{p \times 1}$$

Let the control pre-treatment characteristics be: $X_{\text{control}} = \left(X_2, .., X_{n_0+1}\right) \in R^{p \times n_0}$

**Minimization problem**

# Synthetic controls: minimization problem

**Characteristics**

Pre-treatment characteristics concatenate pre-treatment outcomes and other pre-treatment predictors $Z_1$ eg. cigarette prices:

$$X_{\text{treat}} = X_1 = \begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ .. \\ Y_{1,T_0} \\ Z_1 \end{pmatrix} \in R^{p \times 1}$$

Let the control pre-treatment characteristics be: $X_{\text{control}} = \left(X_2, .., X_{n_0+1}\right) \in R^{p \times n_0}$

**Minimization problem**

$$w^* = \text{argmin}_w \ \|X_{\text{treat}} - X_{\text{control}} w\|_V^2$$

# Synthetic controls: minimization problem

**Characteristics**

Pre-treatment characteristics concatenate pre-treatment outcomes and other pre-treatment predictors $Z_1$ eg. cigarette prices:

$$X_{\text{treat}} = X_1 = \begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ .. \\ Y_{1,T_0} \\ Z_1 \end{pmatrix} \in R^{p \times 1}$$

Let the control pre-treatment characteristics be: $X_{\text{control}} = \left(X_2, .., X_{n_0+1}\right) \in R^{p \times n_0}$

**Minimization problem**

$$w^* = \text{argmin}_w \ \|X_{\text{treat}} - X_{\text{control}} w\|_V^2$$

$$\text{where } \|X\|_V = \sqrt{X^T V X} \quad \text{with} \quad V \in \text{diag}(R^p)$$

This gives more importance to some features than others.

# Synthetic controls: minimization problem

**Characteristics**

Pre-treatment characteristics concatenate pre-treatment outcomes and other pre-treatment predictors $Z_1$ eg. cigarette prices:

$$X_{\text{treat}} = X_1 = \begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ .. \\ Y_{1,T_0} \\ Z_1 \end{pmatrix} \in R^{p \times 1}$$

Let the control pre-treatment characteristics be: $X_{\text{control}} = \left(X_2, .., X_{n_0+1}\right) \in R^{p \times n_0}$

**Minimization problem with constraints**

$$w^* = \text{argmin}_w \ \|X_{\text{treat}} - X_{\text{control}} w\|_V^2$$

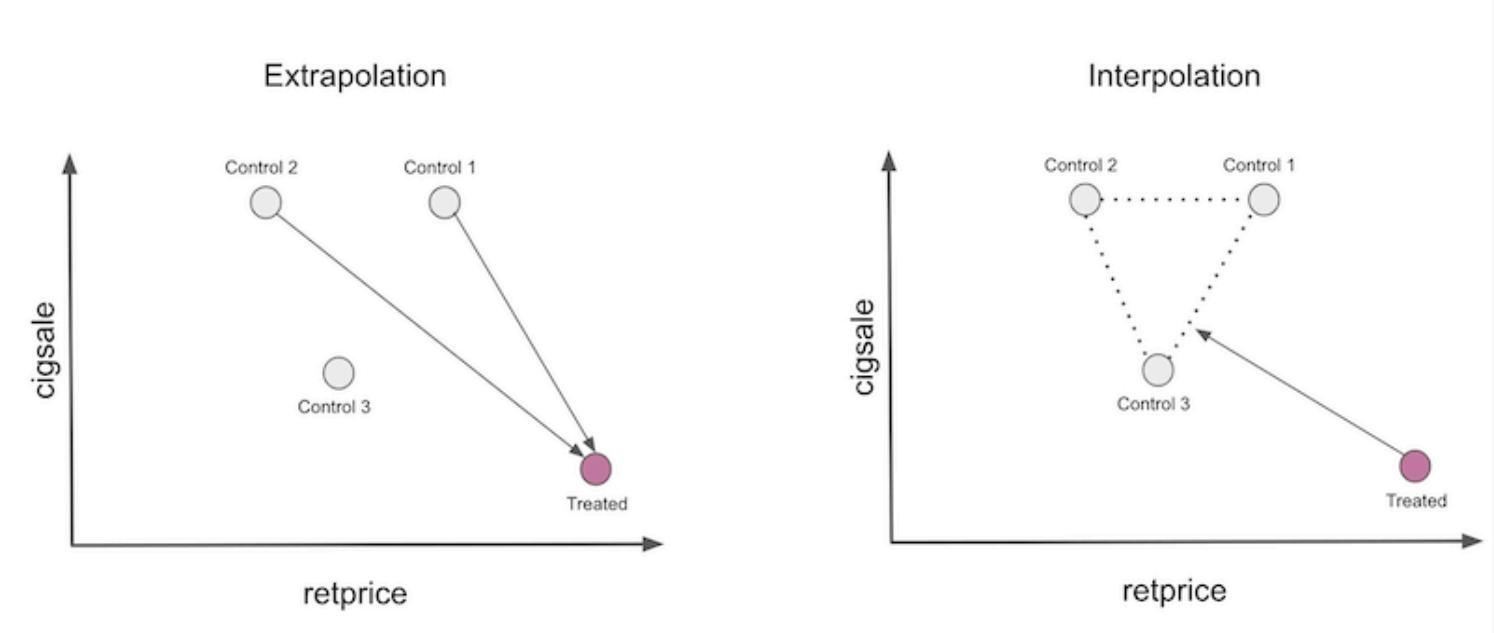
$$s.t. \ w_j \geq 0,$$

$$\sum_{j=2}^{n_0+1} w_j = 1$$

# Synthetic controls: Why choose positive weights summing to one?

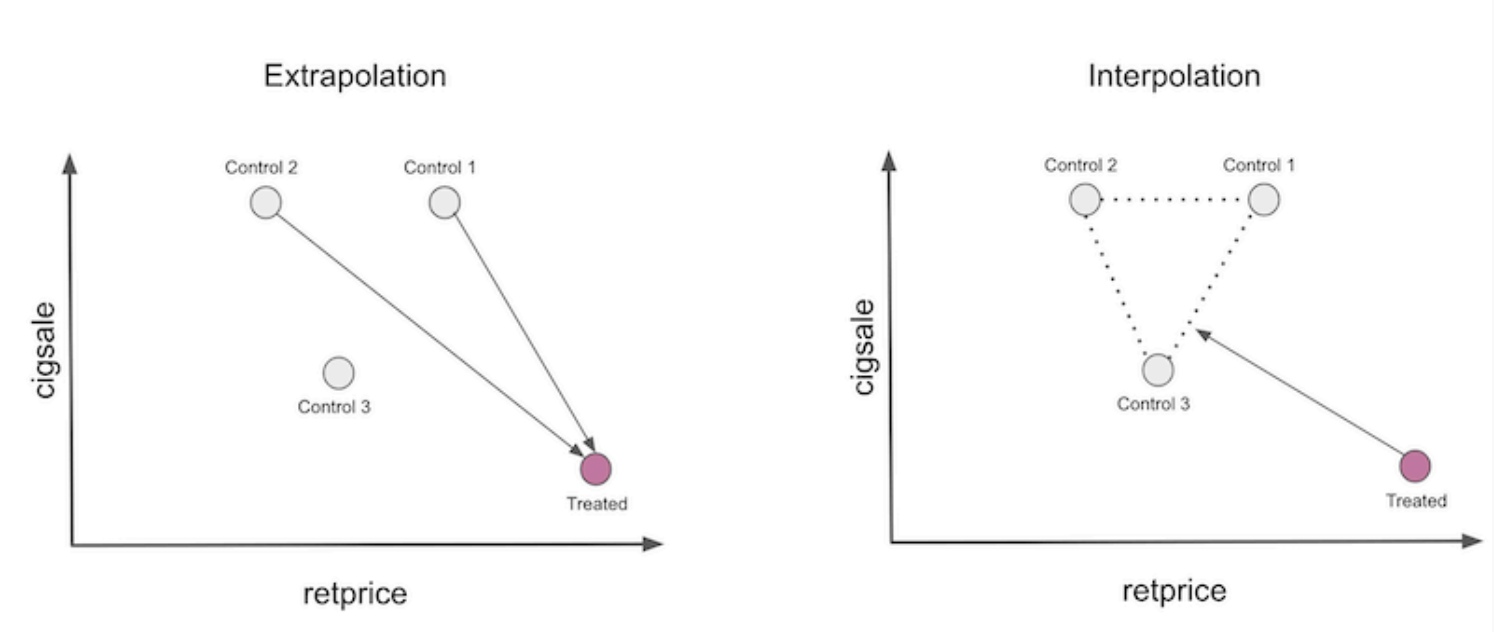
# Synthetic controls: Why choose positive weights summing to one?

## This is called interpolation (vs extrapolation)



# Synthetic controls: Why choose positive weights summing to one?

## This is called interpolation (vs extrapolation)




## Interpolation enforces regularization, thus limits overfitting

Same kind of regularization than L1 norm in Lasso: forces some coefficient to be zero.

# Synthetic controls: Extrapolation failure with unconstrained weight

$p = 2T_0$ covariates:

$$X_j = \begin{pmatrix} Y_{j,1} \\ .. \\ Y_{j,T_0} \\ Z_{j,1} \\ .. \\ Z_{j,T_0} \end{pmatrix}^T \in R^{2T_0}$$

Y cigarette sales, Z cigarette prices.

# Synthetic controls: Extrapolation failure with unconstrained weight

$p = 2T_0$ covariates:

$$X_j = \begin{pmatrix} Y_{j,1} \\ .. \\ Y_{j,T_0} \\ Z_{j,1} \\ .. \\ Z_{j,T_0} \end{pmatrix}^T \in R^{2T_0}$$

Y cigarette sales, Z cigarette prices.

Model: $\underbrace{X_{\text{treat}}}_{p\times 1} \sim \underbrace{X_{\text{control}}}_{p\times n_0}\underbrace{w}_{n_0}$

-> simple linear regression estimated by OLS

# Synthetic controls: Extrapolation failure with unconstrained weight

$p = 2T_0$ covariates:

$$X_j = \begin{pmatrix} Y_{j,1} \\ .. \\ Y_{j,T_0} \\ Z_{j,1} \\ .. \\ Z_{j,T_0} \end{pmatrix}^T \in R^{2T_0}$$

Y cigarette sales, Z cigarette prices.

Model: $\underbrace{X_{\text{treat}}}_{p \times 1} \sim \underbrace{X_{\text{control}}}_{p \times n_0} \underbrace{w}_{n_0}$

Prediction: $\hat{Y}_{\text{synth}} = (Y_{t,j})_{\substack{t=1..T \\ j=2..n_0+1}} w$

# Synthetic controls: Extrapolation failure with unconstrained weight

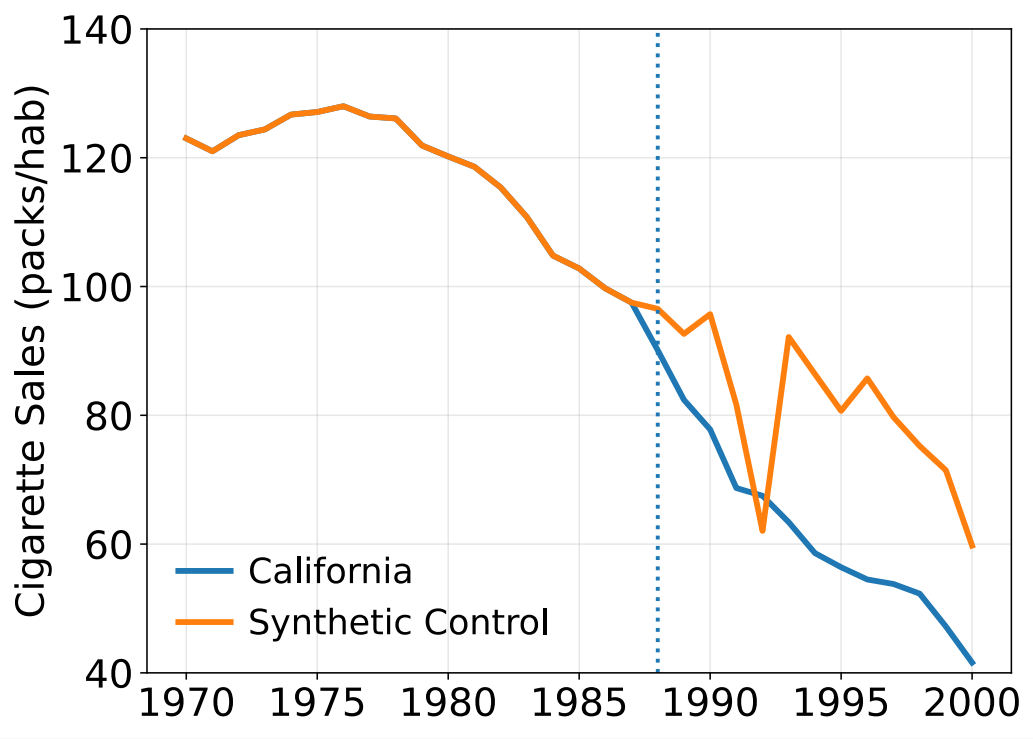
$p = 2T_0$ covariates:

$$X_j = \begin{pmatrix} Y_{j,1} \\ .. \\ Y_{j,T_0} \\ Z_{j,1} \\ .. \\ Z_{j,T_0} \end{pmatrix}^T \in R^{2T_0}$$

Y cigarette sales, Z cigarette prices.

Model: $\underbrace{X_{\text{treat}}}_{p \times 1} \sim \underbrace{X_{\text{control}}}_{p \times n_0} \underbrace{w}_{n_0}$

Prediction: $\hat{Y}_{\text{synth}} = (Y_{t,j})_{\substack{t=1..T \\ j=2..n_0+1}} w$



# Synthetic controls: Extrapolation failure with unconstrained weight

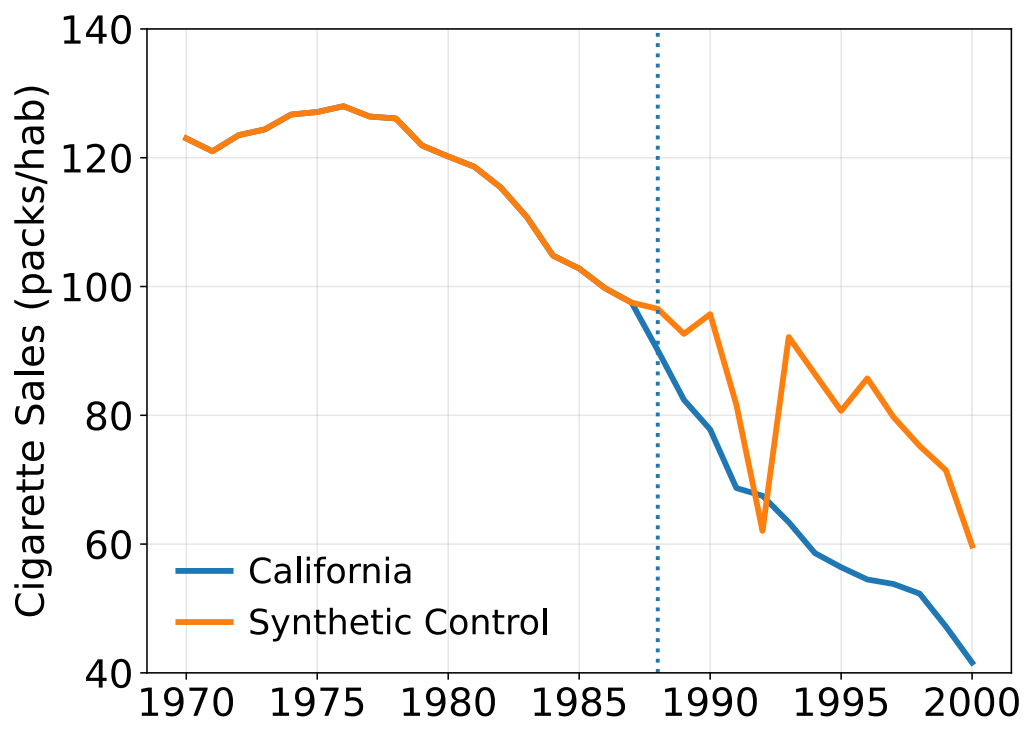
$p = 2T_0$ covariates:

$$X_j = \begin{pmatrix} Y_{j,1} \\ .. \\ Y_{j,T_0} \\ Z_{j,1} \\ .. \\ Z_{j,T_0} \end{pmatrix}^T \in R^{2T_0}$$

Y cigarette sales, Z cigarette prices.

Model: $\underbrace{X_{\text{treat}}}_{p \times 1} \sim \underbrace{X_{\text{control}}}_{p \times n_0} \underbrace{w}_{n_0}$

Prediction: $\hat{Y}_{\text{synth}} = (Y_{t,j})_{\substack{t=1..T \\ j=2..n_0+1}} w$




😭 **Overfitting**

# Synthetic controls: How to choose the predictor weights $V$?

1. Don't choose: set $V = I_p$, ie. $\|X\|_V = \|X\|_2$.
2. Rescale by the variance of the predictors:
   $V = \text{diag}\left(\text{var}\left(Y_{j,1}\right)^{-1}, .., \text{var}\left(Y_{j,T_0}\right)^{-1}, \text{var}\left(Z_{j,1}\right)^{-1}, .., \text{var}\left(Z_{j,T_0}\right)^{-1}\right)$.
3. Minimize the pre-treatment mean squared prediction error (MSPE) of the treated unit:

# Synthetic controls: How to choose the predictor weights $V$?

1. Don't choose: set $V = I_p$, ie. $\|X\|_V = \|X\|_2$.
2. Rescale by the variance of the predictors:
   $V = \text{diag}\left(\text{var}\left(Y_{j,1}\right)^{-1}, .., \text{var}\left(Y_{j,T_0}\right)^{-1}, \text{var}\left(Z_{j,1}\right)^{-1}, .., \text{var}\left(Z_{j,T_0}\right)^{-1}\right)$.
3. Minimize the pre-treatment mean squared prediction error (MSPE) of the treated unit:

# Synthetic controls: How to choose the predictor weights $V$?

1. Don't choose: set $V = I_p$, ie. $\|X\|_V = \|X\|_2$.
2. Rescale by the variance of the predictors:
   $V = \text{diag}\left(\text{var}\left(Y_{j,1}\right)^{-1}, .., \text{var}\left(Y_{j,T_0}\right)^{-1}, \text{var}\left(Z_{j,1}\right)^{-1}, .., \text{var}\left(Z_{j,T_0}\right)^{-1}\right)$.
3. Minimize the pre-treatment mean squared prediction error (MSPE) of the treated unit:

$$\begin{aligned}\text{MSPE}(V) &= \sum_{t=1}^{T_0}\left[Y_{1,t} - \sum_{j=2}^{n_0+1} w_j^*(V) Y_{j,t}\right]^2 \\ &= \| \left(Y_{1,t}\right)_{t=1..T_0} - \left(Y_{j,t}\right)_{\substack{j=2..n_0+1 \\ t=1..T_0}}^{T} \hat{w} \|_2^2\end{aligned}$$

This solution is solved by running two optimization problems:

- Inner loop solving $w^*(V) = \text{argmin}_w \|X_{\text{treat}} - X_{\text{control}} w\|_V^2$
- Outer loop solving $V^* = \text{argmin}_V \text{MSPE}(V)$

# Synthetic controls: estimation without the outer optimization problem

Same coviarates: $X_j = \begin{pmatrix} Y_{j,1} \\ .. \\ Y_{j,T_0} \\ Z_{j,1} \\ .. \\ Z_{j,T_0} \end{pmatrix}^T$

Y cigarette sales, Z cigarette prices.

SCM minization with $V = I_p$, hence, $\|X\|_V = \|X\|_2$.

$$w^* = \text{argmin}_w \ \|X_{\text{treat}} - X_{\text{control}} w\|_2^2$$
$$s.t.\ w_j \geq 0,$$
$$\sum_{j=2}^{n_0+1} w_j = 1$$

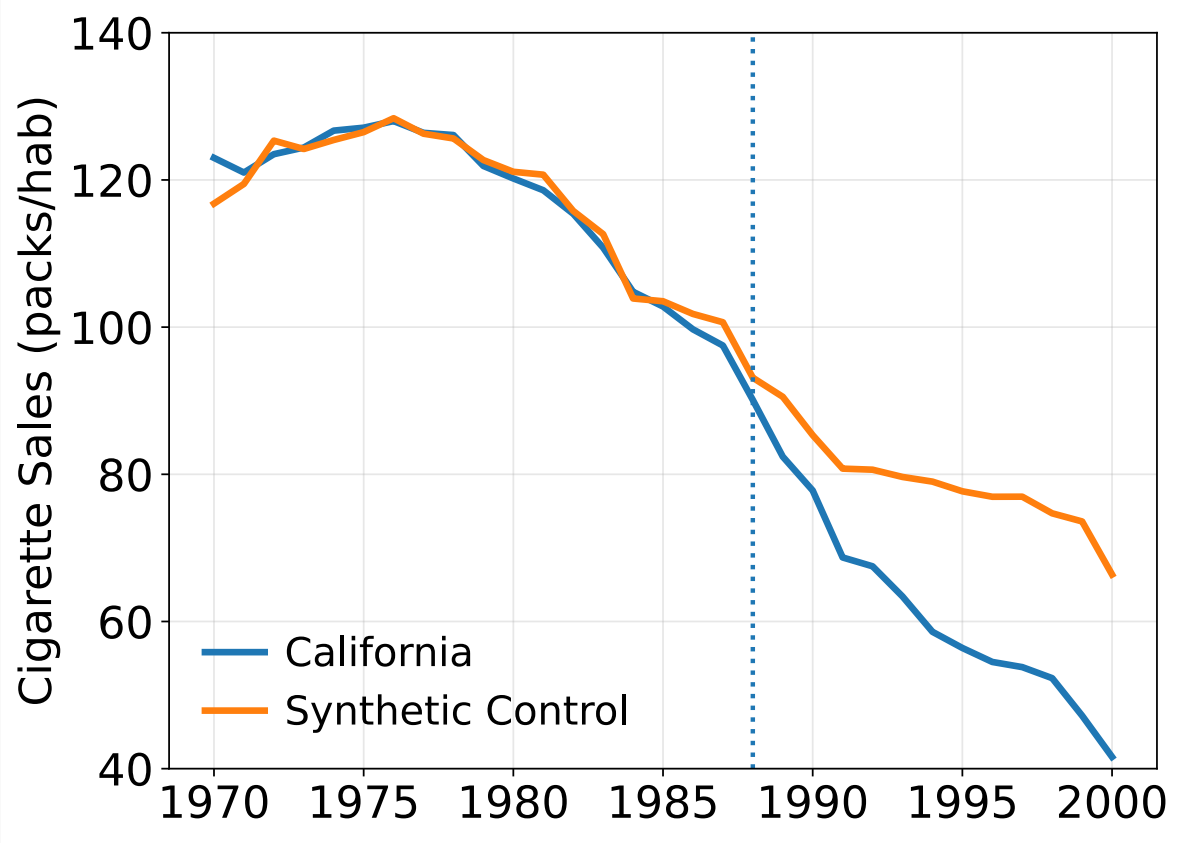
# Synthetic controls: estimation without the outer optimization problem

Same coviarates: $X_j = \begin{pmatrix} Y_{j,1} \\ .. \\ Y_{j,T_0} \\ Z_{j,1} \\ .. \\ Z_{j,T_0} \end{pmatrix}^T$

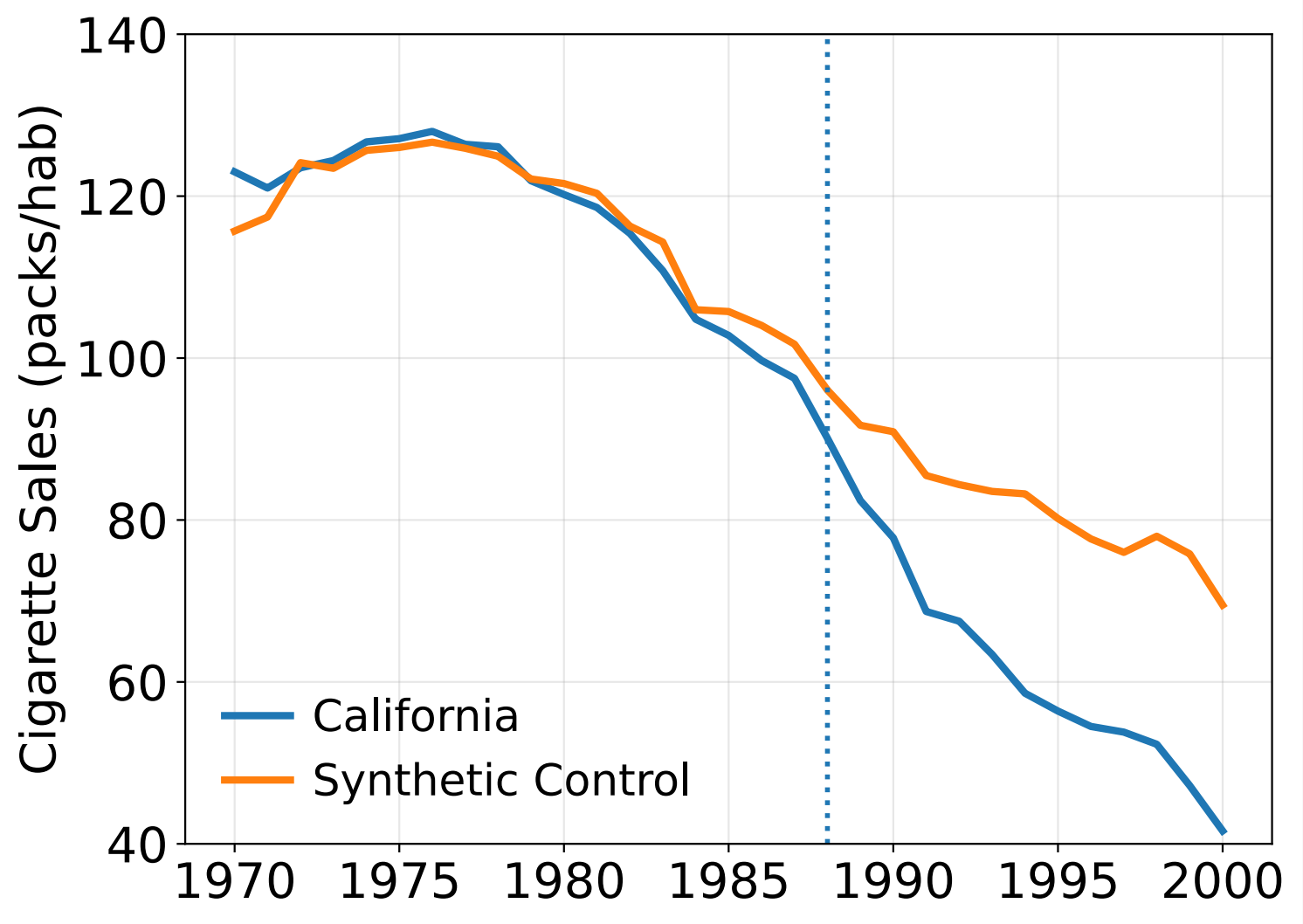
Y cigarette sales, Z cigarette prices.

SCM minization with $V = I_p$, hence, $\|X\|_V = \|X\|_2$.

$$w^* = \text{argmin}_w \ \|X_{\text{treat}} - X_{\text{control}} w\|_2^2$$
$$s.t.\ w_j \geq 0,$$
$$\sum_{j=2}^{n_0+1} w_j = 1$$



# Synthetic controls: estimation with the outer optimization problem



# Synthetic controls: inference

**Variability does not come from the variability of the outcomes**

Indeed, aggregates are often not very noisy (once deseasonalized)…

# Synthetic controls: inference

**Variability does not come from the variability of the outcomes**

Indeed, aggregates are often not very noisy (once deseasonalized)…

**… but from the variability of the chosen control units**

Treatment assignment introduces more noise than outcome variability.

# Synthetic controls: inference

**Variability does not come from the variability of the outcomes**

Indeed, aggregates are often not very noisy (once deseasonalized)...

**... but from the variability of the chosen control units**

Treatment assignment introduces more noise than outcome variability.

(Abadie et al., 2010) introduced the placebo test to assess the variability of the synthetic control.

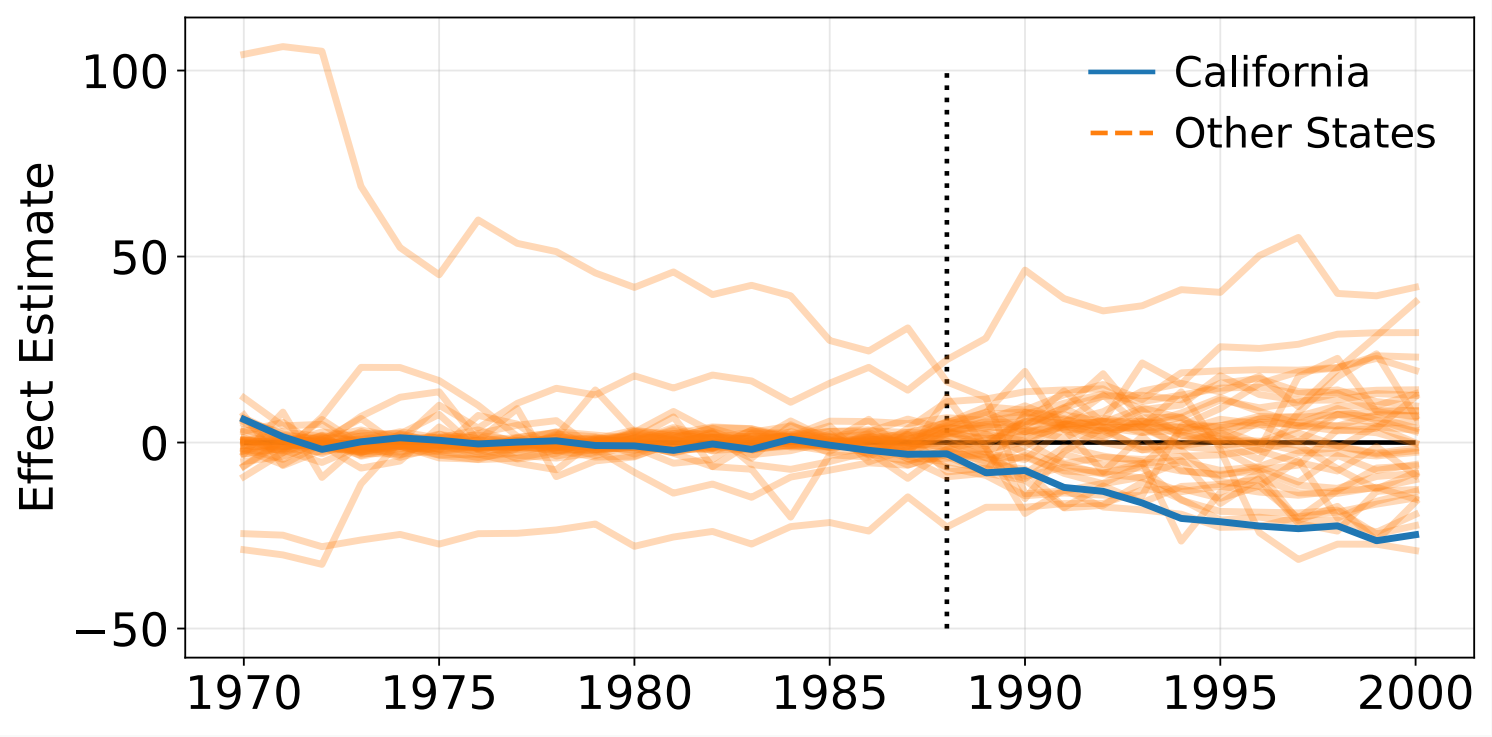
# Synthetic controls: inference with Placebo tests

**Idea of Fisher's Exact tests**

- Permute the treated and control exhaustively.
- For each unit, we pretend it is the treated while the others are the control: we call it a placebo
- Compute the synthetic control for each placebo: it should be close to zero.

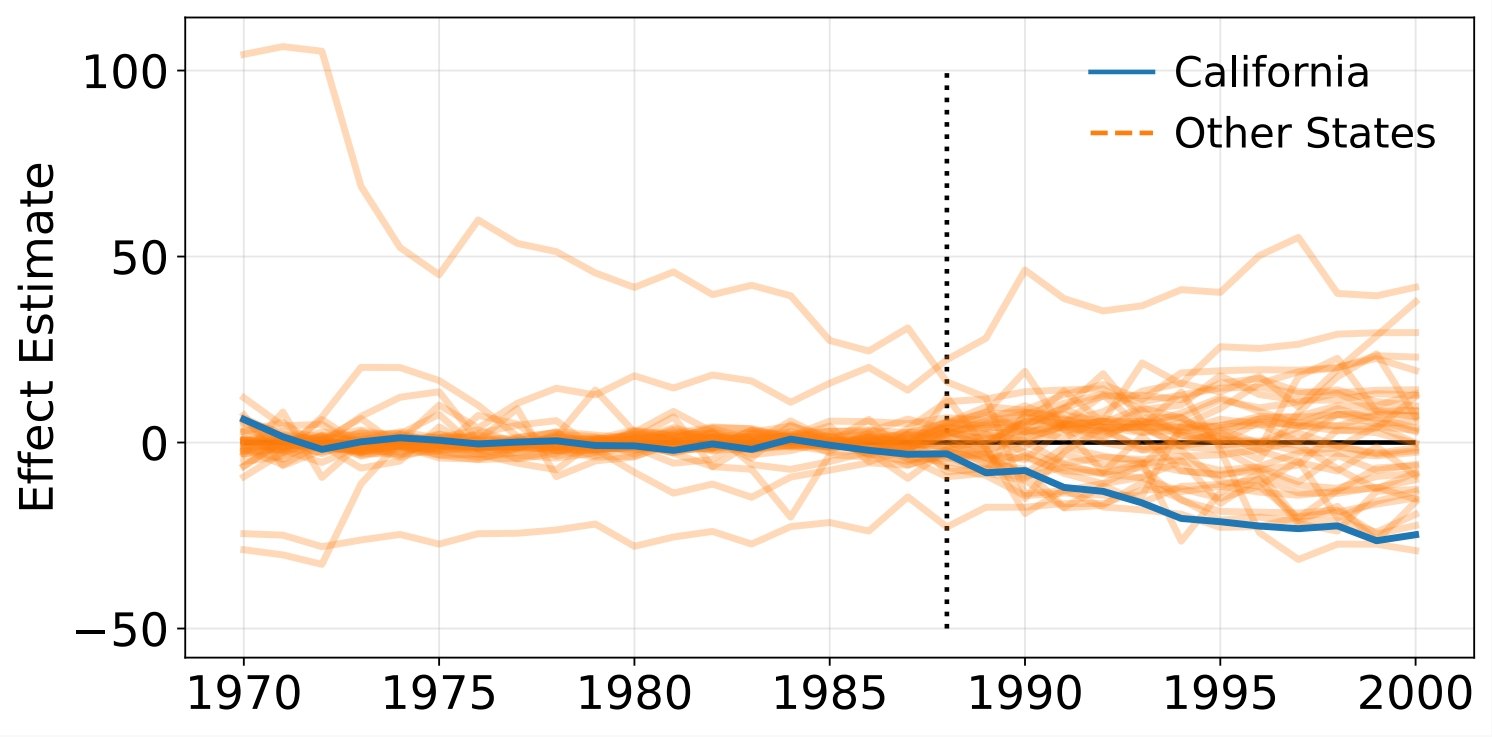
# Synthetic controls: inference with Placebo tests, example

## Placebo estimation for all 38 control states



# Synthetic controls: inference with Placebo tests, example

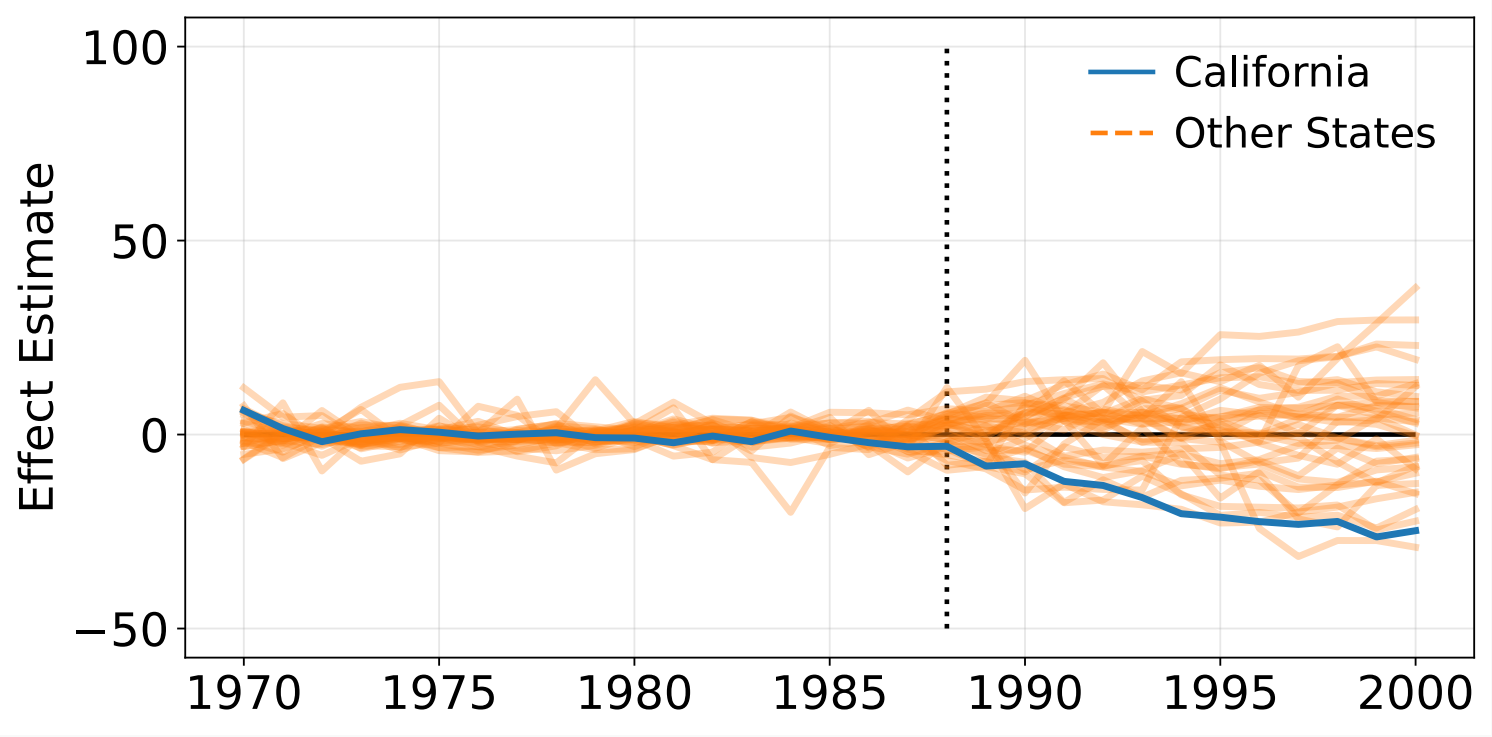
## Placebo estimation for all 38 control states




- More variance after the treatment for California than before.
- Some states have pre-treatment trends which are hard to predict.

# Synthetic controls: inference with Placebo tests, example

**Placebo estimation for 34 control states with "good" pre-treatment fit**

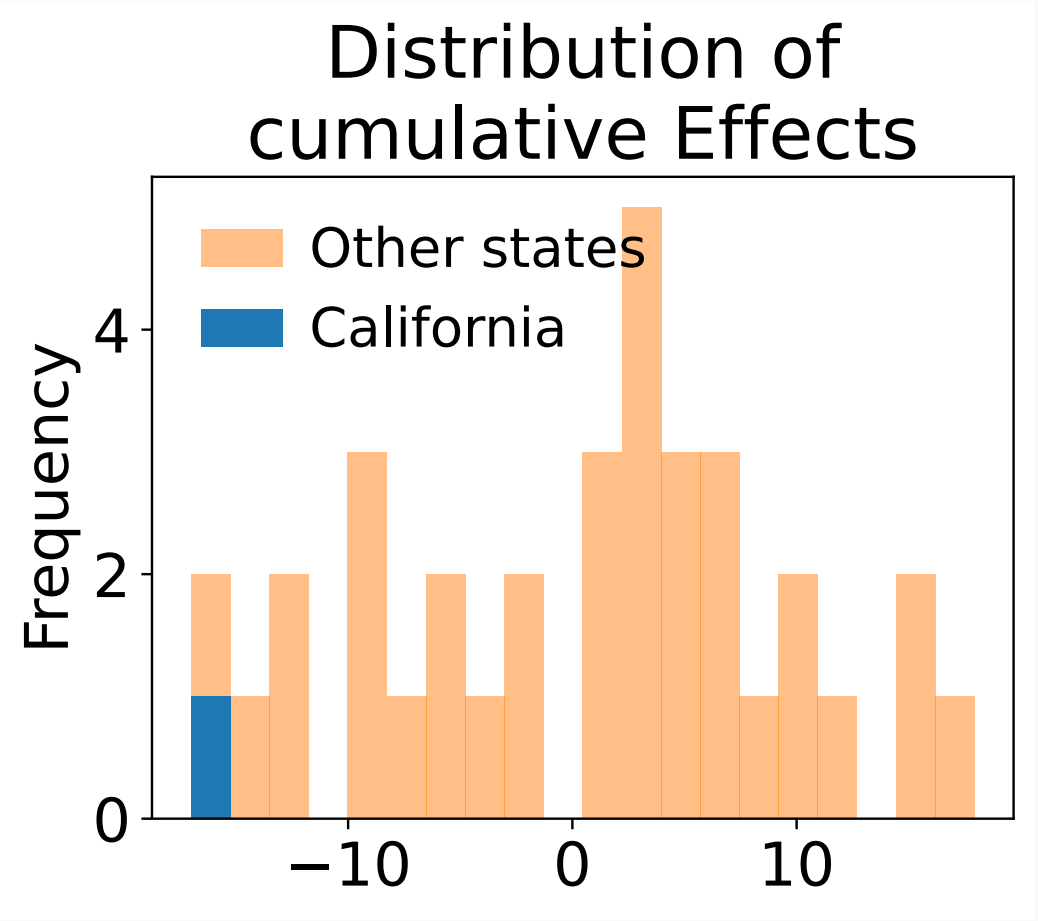



I removed the states above the 90 percentiles of the distribution of the pre-treatment fit.

# Synthetic controls: inference with Placebo tests, example

**California absolute cumulative effect**

$\hat{\tau}_{\text{scm, california}} = -17.00$



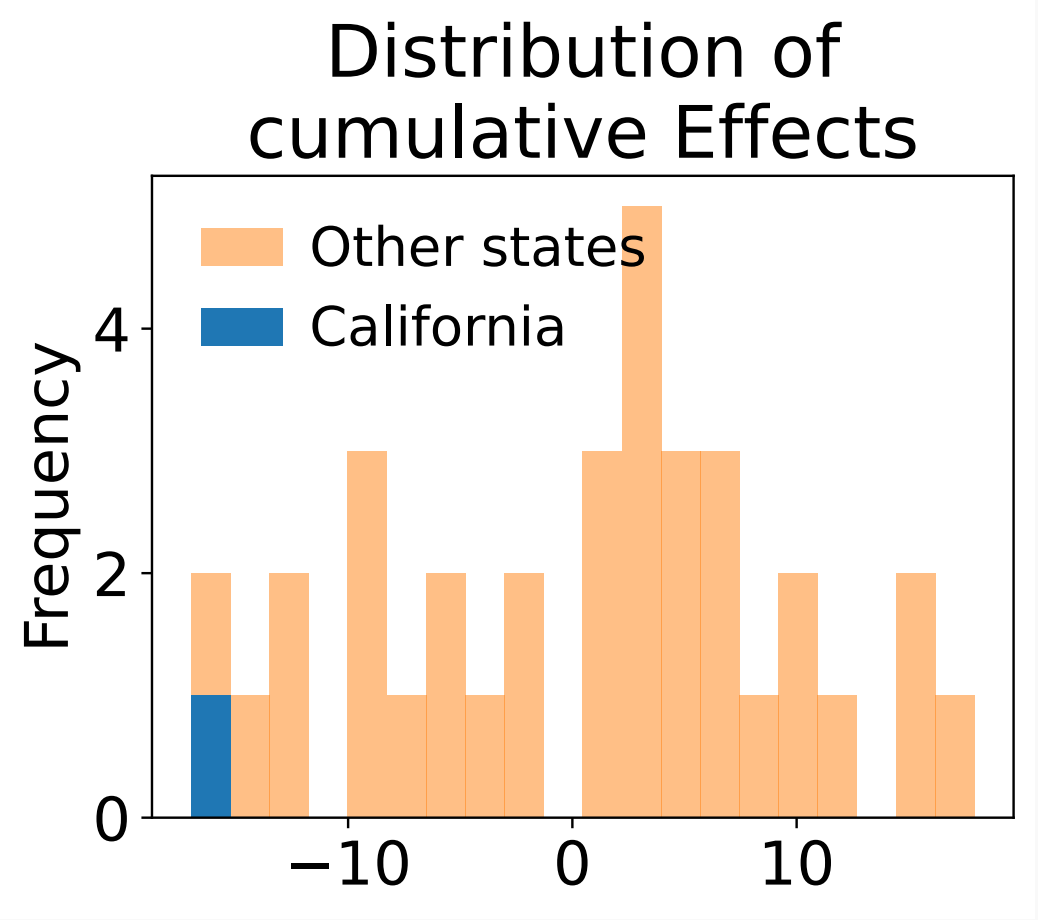
# Synthetic controls: inference with Placebo tests, example

**California absolute cumulative effect**

$$\hat{\tau}_{\text{scm, california}} = -17.00$$

**Get a p-value**

$$\text{PV} = \frac{1}{n_0} \sum_{j=2}^{n_0} \mathbb{1}\left(|\hat{\tau}_{\text{scm, california}}| > |\hat{\tau}_{\text{scm},j}|\right)$$
$$= 0.029$$



# Synthetic controls: inference with conformal prediction

## Failure of conditional ignorability

TODO

# Synthetic controls: Take-away

## Pros

- More convincing for parallel trends assumption.
- Simple for multiple time periods.
- Gives confidence intervals.

# Synthetic controls: Take-away

## Pros

- More convincing for parallel trends assumption.
- Simple for multiple time periods.
- Gives confidence intervals.

## Cons

- Requires many control units to yield good pre-treatment fits.
- Might be prone to overfitting during the pre-treatment period.
- Still requires a strong assumption: the weights should also balance the post-treatment unexposed outcomes ie. conditional ignorability. See (Arkhangelsky et al., 2021) for discussions.
- Still requires the no-anticipation assumption.

# Conditional difference-in-differences

# Time-series modelisation: methods without a control group

# Interrupted Time Series: intuition

## Setup

- One treated unit, no control unit.
- Multiple time periods.
- Sometimes, predictors are availables: there are called exogeneous covariates.

# Interrupted Time Series: intuition

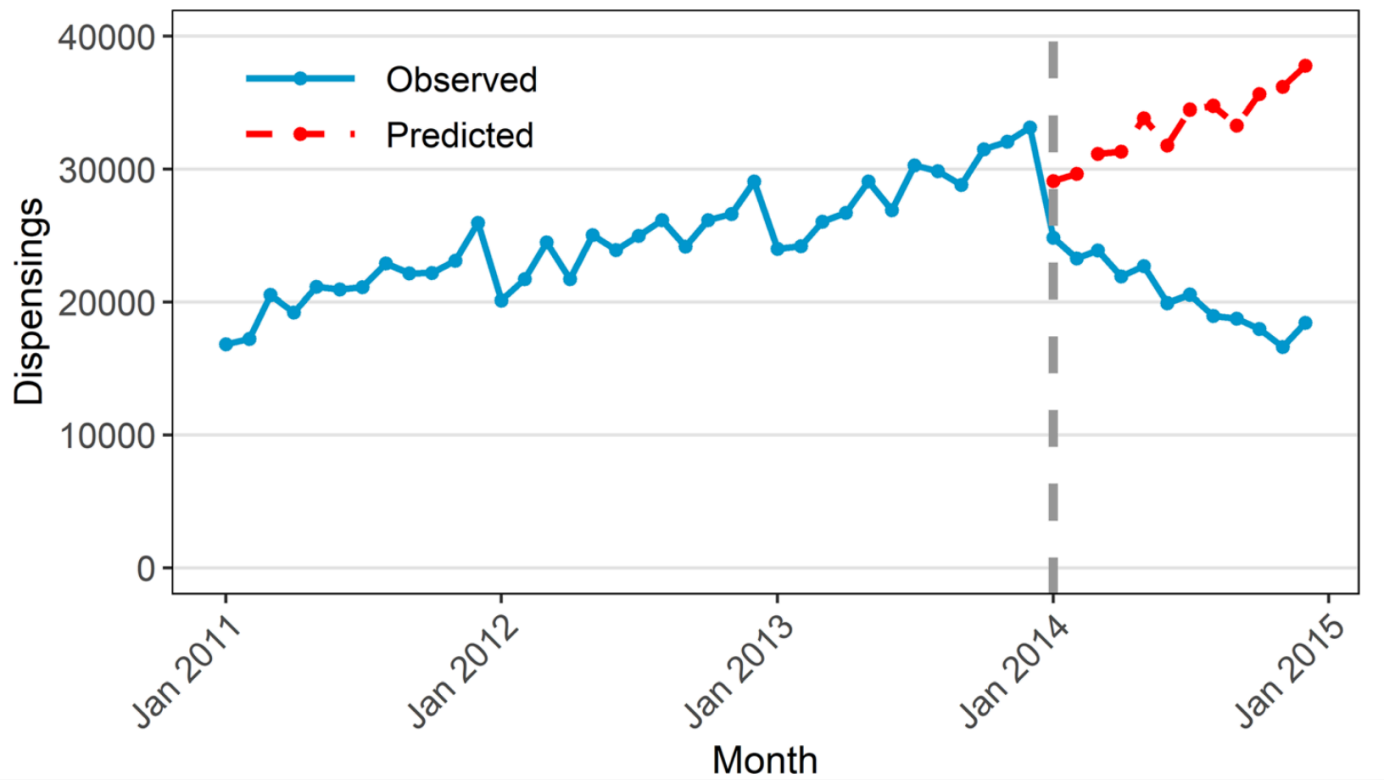
## Setup

- One treated unit, no control unit.
- Multiple time periods.
- Sometimes, predictors are availables: there are called exogeneous covariates.

## Intuition

- Model the pre-treatment trend: $Y_{t(1)}$ for $t < T_0$
- Predict post-treatment trend as the control: $\hat{Y}_t(0)$ for $t > T_0$
- Obtain treatment effect by taking the difference between observed and predicted post-treatment observations: $Y_t(1) - \hat{Y}_t(0)$

# Interrupted Time Series: illustration from (Schaffer et al., 2021)




$Y_t$: Dispensations of quetiapine, an anti-psychotic medicine.

Treatment: Restriction of the conditions under which quetiapine could be subsidised.

## Tools

- ARIMA models: AutoRegressive Integrated Moving Average

**Motivation of ARIMA**

Take into account:

- the structure of autodependance between observation,
- linear trends,
- seasonality.

A good reference for ARIMA: Forecasting: Principles and Practice, chapter 8

# ARIMA are State Space Models (SSM) says the machine learning community

## Why showing this formulation ?

- I better understand ARIMA formulated as state space models.
- SSM are more general than ARIMA models.
- ARIMA are (almost always) fitted with SSM optimization algorithms.

## What is a state space model?

# State space models: Graphical representation with DAG

# State space models: a general formulation

## Idea

- Decompose the time series into two components: the state and the observation.
- The state is a latent variable that evolves over time.
- The observation is a noisy version of the state.

## Formalization

- State equation: $x_t = Fx_{t-1} + v_t$
- Observation equation: $y_t = Hx_t + w_t$
- $v_t$ and $w_t$ are noise terms.

## Example

- ARIMA(1,1,1) model: $x_t = x_{t-1} + v_t$, $y_t = x_t + w_t$

# State space models: a general formulation

- $F = 1, H = 1, v_t = 0, w_t = 0.$

# Fitting state space model

# Modern state space models

- Long Short Term Memory (LSTM) networks (Graves & Graves, 2012): a type of Recurrent Neural Network (RNN) that can learn long-term dependencies. Was state of the art for language tasks before transformers.
- Mamba (Gu & Dao, 2023): A recent proposition to mitigate the main limitations of transfomers which is high complexity relative to the length of the sequence. Good blog-style introduction in (Ayonrinde, 2024).

# A word on model families for ITS

We saw ARIMA models and the more general class of state space models.

However, we could any model that we want to fit the pre-treatment trend !

# A word on model families for ITS

We saw ARIMA models and the more general class of state space models.

However, we could any model that we want to fit the pre-treatment trend !

- Facebook prophet model (Taylor & Letham, 2018) uses Generalized Additive Models (GAM).
- Any sklearn estimator could do the trick: Linear regression, Random Forest, Gradient Boosting...

# A word on model families for ITS

We saw ARIMA models and the more general class of state space models.

However, we could any model that we want to fit the pre-treatment trend !

- Facebook prophet model (Taylor & Letham, 2018) uses Generalized Additive Models (GAM).
- Any sklearn estimator could do the trick: Linear regression, Random Forest, Gradient Boosting...

# A word on model families for ITS

We saw ARIMA models and the more general class of state space models.

However, we could any model that we want to fit the pre-treatment trend !

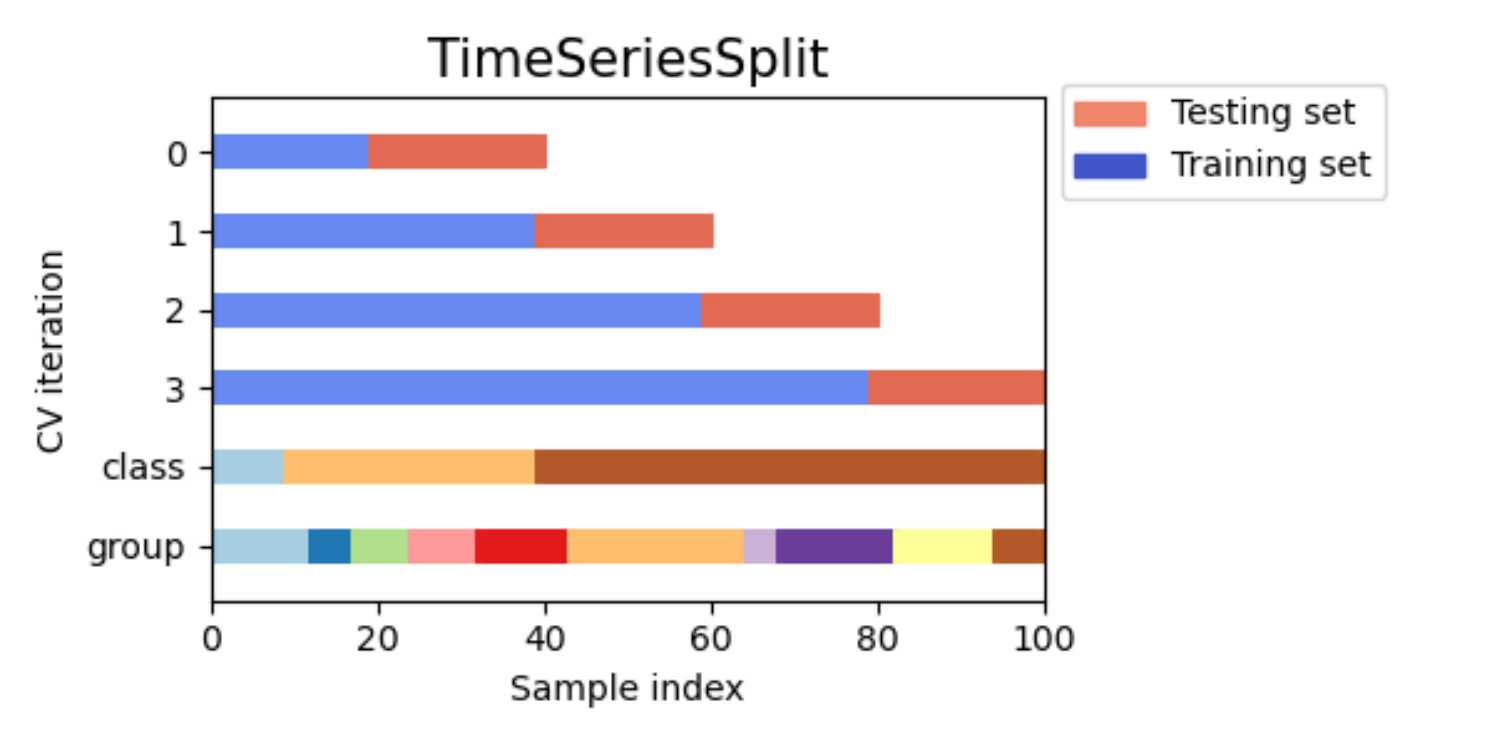
- Facebook prophet model (Taylor & Letham, 2018) uses Generalized Additive Models (GAM).
- Any sklearn estimator could do the trick: Linear regression, Random Forest, Gradient Boosting...

⚠️ You should pay attention to appropriate train/test split when cross-validating a time-series model not to use the future to predict the past.

Relevant remark for all time series models (even ARIMA or state space models).

# Cross-validation for time-series models

```python
from sklearn.model_selection import TimeSeriesSplit
```




This avoids to use the future to predict the past.

# Main threat to validity for an ITS: historical bias

⚠️ If there is a co-intervention, it will impact the outcome trend and bias the treatment effect estimation.

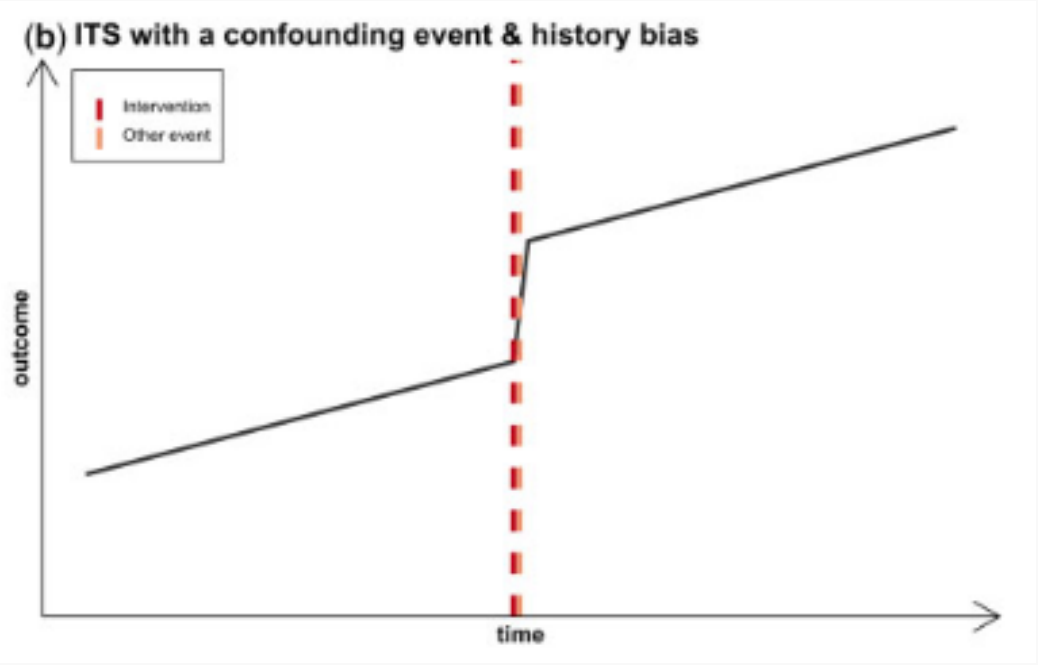



Illustration from (Degli Esposti et al., 2020, Fig. 1)

# Main threat to validity for an ITS: historical bias

⚠️ If there is a co-intervention, it will impact the outcome trend and bias the treatment effect estimation.

💡 Adding a control series of predictors can help to mitigate this bias.

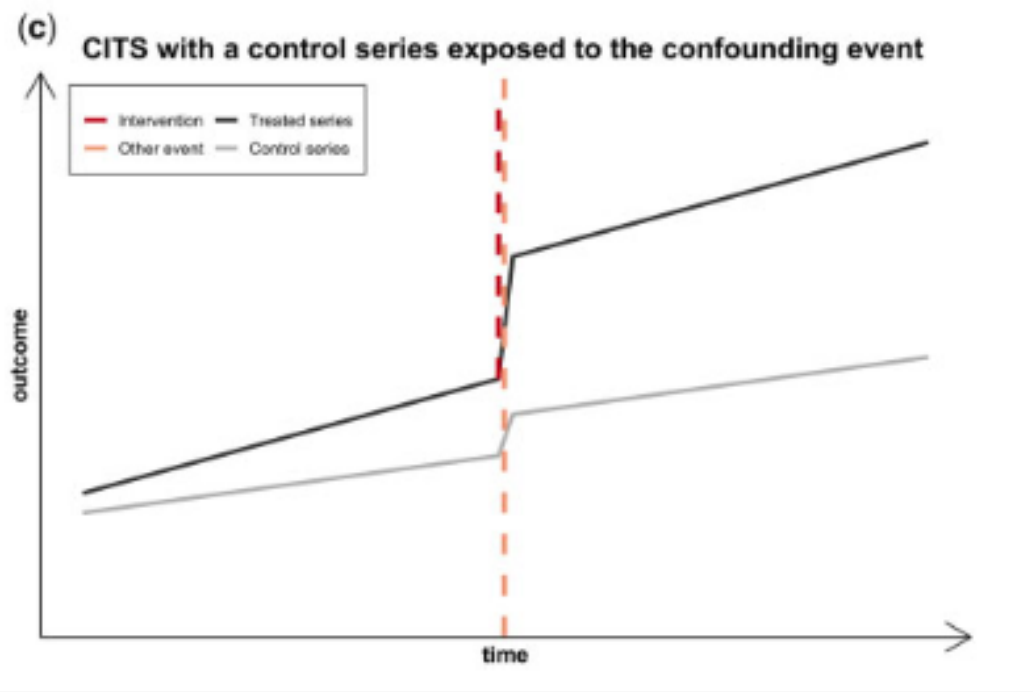



Illustration from (Degli Esposti et al., 2020, Fig. 1)

# Take-away on ITS

## Pros

- Suitable when no control unit is available. The pre-treatment trend is the control.
- Handles multiple time periods.
- A lot of software available (eg. ARIMA models).
- Simple: few parameters to tune.

## Cons

- Prone to bias by other events happening around the treatment time and impacting the outcome trend.
- Prone to overfitting of the pre-treatment trend.

# An attempt to map event study methods

| **Methods** | Characteristics | Hypotheses | Community | Introduction | Good reference |
|---|---|---|---|---|---|
| DID/TWFE | Treated/control units, few time periods, no predictors | Parallel trends, no anticipation, prone to overfitting | Economics | Causal Inference for the Brave and True, chapter 13 | (Arkhangelsky & Imbens, 2024) |
| ARIMA, ITS | No controls, no/few predictors, seasonality | Stationnarity , no anticipation, prone to overfitting | Epidemiology, Economics | Forecasting: Principles and Practice | (Schaffer et al., 2021) |
| State space models | Multiple time periods, control units or predictors, generalization of ARIMA | Contional ignorability on predictors, goodness of fit pretreatment | Machine learning, bayesian methods | (Brockwell & Davis, 2016, chapter 9) | (Murphy, 2022, chapter 18) |
| Synthetic control | Treated/control units, multiple time periods | Conditional parallel trend on controls, goodness of fit pretreatment | Economics | Causal Inference for the Brave and True | (Abadie, 2021) |

# A summary on R packages for event studies

| **Package name** | Methods | Predictors | Control units | Multiple time periods |
|---|---|---|---|---|
| did | Difference-in-differences | ❌ | ❌ | ❌ |
| forecast | ARIMA, ITS | ✅ | ❌ | ✅ |
| Synth | Synthetic control | ❌ | ✅ | ✅ |
| Causal impact | Bayesian state space models | ✅ | ❌ | ✅ |

# A summary on Python packages for event studies

| **Package name** | Methods | Predictors | Control units | Multiple time periods |
|---|---|---|---|---|
| statsmodels.OLS | Difference-in-differences, TWFE | ❌ | ❌ | ❌ |
| statsmodels | ARIMA(X), ITS, bayesian state space models | ✅ | ❌ | ✅ |
| pmdarima | ARIMA(X), ITS | ✅ | ❌ | ✅ |
| SyntheticControlMethods | Synthetic control | ❌ | ✅ | ✅ |
| pysyncon | Synthetic control | ❌ | ✅ | ✅ |
| causalimpact (pymc implementation) | Bayesian state space models | ✅ | ❌ | ✅ |
| causal-impact (statsmodels implementation) | Bayesian state space models | ✅ | ❌ | ✅ |

# State Space Models

# State space models: Take-away

# Final word -- What methods to chose: some guides

## DID-family methods

- Control units available (at least one)
- Few time periods
- Parallel trend is credible (if necessary by adjusting the model on predictors).

## Synthetic Control Methods

- Mutiple and different controls as well as multiple time periods
- Pre-treatment outcomes of the control units predict well the treated unit outcome.
- No-spill over from the treatment to the control units.

## ITS: SARIMA or state space models

- No evident control units
- Pre-treatment outcome of the treated unit seems a good control
- Control predictors not impacted by the treatment availables
- No co-intervention that could impact the treated outcome.

# Python hands-on

# To your notebooks 👩‍💻 !

- url: https://github.com/strayMat/causal-ml-course/tree/main/notebooks

# Bibliography

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

*Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. American Economic Review, 93(1), 113–132.*

*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.*

*Arkhangelsky, D., & Imbens, G. (2024). Causal models for longitudinal and panel data: A survey. The Econometrics Journal, 27(3), C1–C61.*

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

# Bibliography


*Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. The Review of Economics and Statistics, 47–57.*

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

*Ayonrinde, K. (2024). Mamba Explained. The Gradient.*

*Bonander, C., Humphreys, D., & Degli Esposti, M. (2021). Synthetic control methods for the evaluation of single-unit interventions in epidemiology: a tutorial. American Journal of Epidemiology, 190(12), 2700–2711.*

*Brockwell, P. J., & Davis, R. A. (2016). Introduction to time series and forecasting, Third edition. Springer. http://repository.cinec.edu/bitstream/cinec20/1109/1/2016_Book_IntroductionToTimeSeriesAndFor.pdf*

*De Chaisemartin, C., & d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. American Economic Review, 110(9), 2964–2996.*

# Bibliography


*Degli Esposti, M., Spreckelsen, T., Gasparrini, A., Wiebe, D. J., Bonander, C., Yakubovich, A. R., & Humphreys, D. K. (2020). Can synthetic controls improve causal inference in interrupted time series evaluations of public health interventions?. International Journal of Epidemiology, 49(6), 2010–2020.*

*Graves, A., & Graves, A. (2012). Long short-term memory. Supervised Sequence Labelling with Recurrent Neural Networks, 37–45.*

*Grupp, T., Mishra, P., Reynaert, M., & Benthem, A. A. van. (2023). An evaluation of protected area policies in the european union.*

*Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. Arxiv Preprint Arxiv:2312.00752.*

*Miller, S., Johnson, N., & Wherry, L. R. (2019). Medicaid and mortality: new evidence from linked survey and administrative data.*

*Murphy, K. P. (2022). Probabilistic machine learning: an introduction. MIT press.*

# Bibliography

*Puig-Codina, L., Pinilla, J., & Puig-Junoy, J. (2021). The impact of taxing sugar-sweetened beverages on cola purchasing in Catalonia: an approach to causal inference with time series cross-sectional data. The European Journal of Health Economics, 22(1), 155–168.*

*Schaffer, A. L., Dobbins, T. A., & Pearson, S.-A. (2021). Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions. BMC Medical Research Methodology, 21, 1–12.*

*Snow, J. (1855). On the mode of communication of cholera. John Churchill.*

*Stechemesser, A., Koch, N., Mark, E., Dilger, E., Klösel, P., Menicacci, L., Nachtigall, D., Pretis, F., Ritter, N., Schwarz, M., & others. (2024). Climate policies that achieved major emission reductions: Global evidence from two decades. Science, 385(6711), 884–892.*

*Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37–45.*

*Wager, S. (2024, ). Causal inference: A statistical learning approach. preparation.*