Week 6: Synthetic Control Method

PUBL0050 Causal Inference

Tom Barton & Julia de Romémont

Term 2 2023-24

UCL Departement of Political Science

This Lecture

Motivation

Synthetic Control

Inference & Placebo Tests

Additional Applications

Conclusion

Motivation

Motivation

"If an instance in which the phenomena under investigation occurs and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former, the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon."

– J.S. Mill on the "Method of Difference"

Motivation

Comparative case studies have a long history in applied political science:

- **Qualitative**: "thick" description of the context/features of two or more instances of specific phenomena. Aim to describe contrasts or similarities across the cases and reason inductively about causality
- **Quantitative**: more explicitly causal, using aggregate data from one treated unit and a small set of control units. Often based on 'natural experiments' where a shock affects one unit, but not others.

Quantitative Comparative Case Studies

Goal

- Estimate effects of events or policy interventions that take place at an aggregate level
- Types of unit: cities, states, countries, etc
- Types of intervention: passage of laws, economic shocks, etc

Approach

- Compare the evolution of an aggregate outcome for the unit affected by the intervention to the evolution of the same outcome for some control group
- e.g. Card (1990), Card and Krueger (1994), Abadie and Gardeazabal (2003)

Quantitative Comparative Case Studies

Advantages

- Policy interventions often take place at an aggregate level
- Aggregate/macro data are often available

Problems

- Reasons for selection of control group are often ambiguous
- Standard errors do not reflect uncertainty about the ability of the control group to reproduce the counterfactual of interest

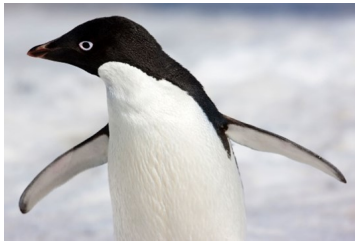
Solution

- If you don't have a good control group: synthesize one

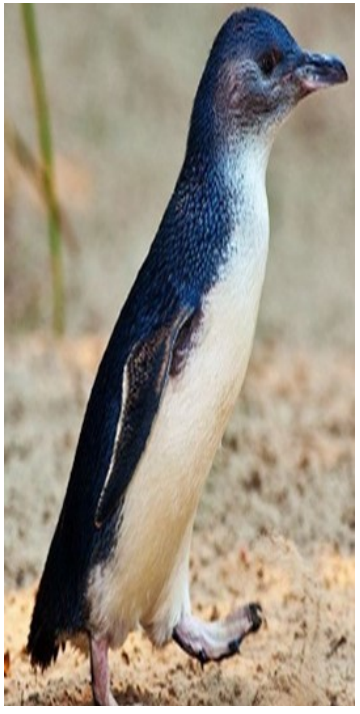
Synthetic Penguins

Synthetic Penguins





Synthetic Penguins



Synthetic Penguins

V

Running example

Reunification of West and East Germany

What were the economic effects of reunification on the West German economy? Many economic historians argue that reunification had large negative economic costs, but identification is difficult because there is no obvious country with which we can compare the growth trajectory of West Germany. Abadie et al (2015) estimate the effects of reunification by comparing the actual time series for West Germany with a **synthetic** control group which provides the counterfactual.

- Outcome: GDP per capita (inflation adjusted)
- Treatment: Reunification (1 for W. Germany after 1990, 0 otherwise)
- **Time**: Years (1960 to 2003)

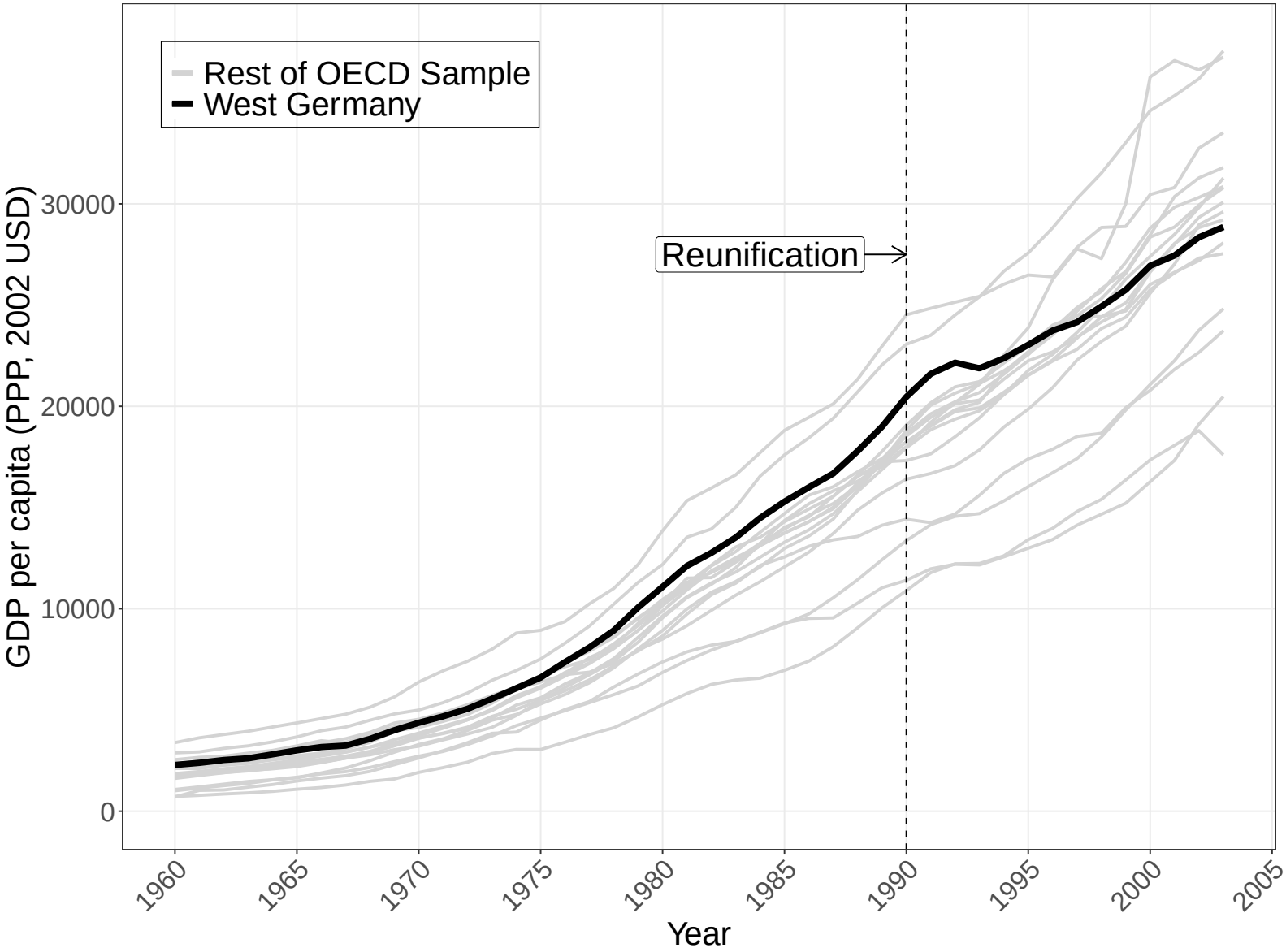
What should be the control group?

What is the most appropriate control group for evaluating the effects of reunification on West Germany in 1990?

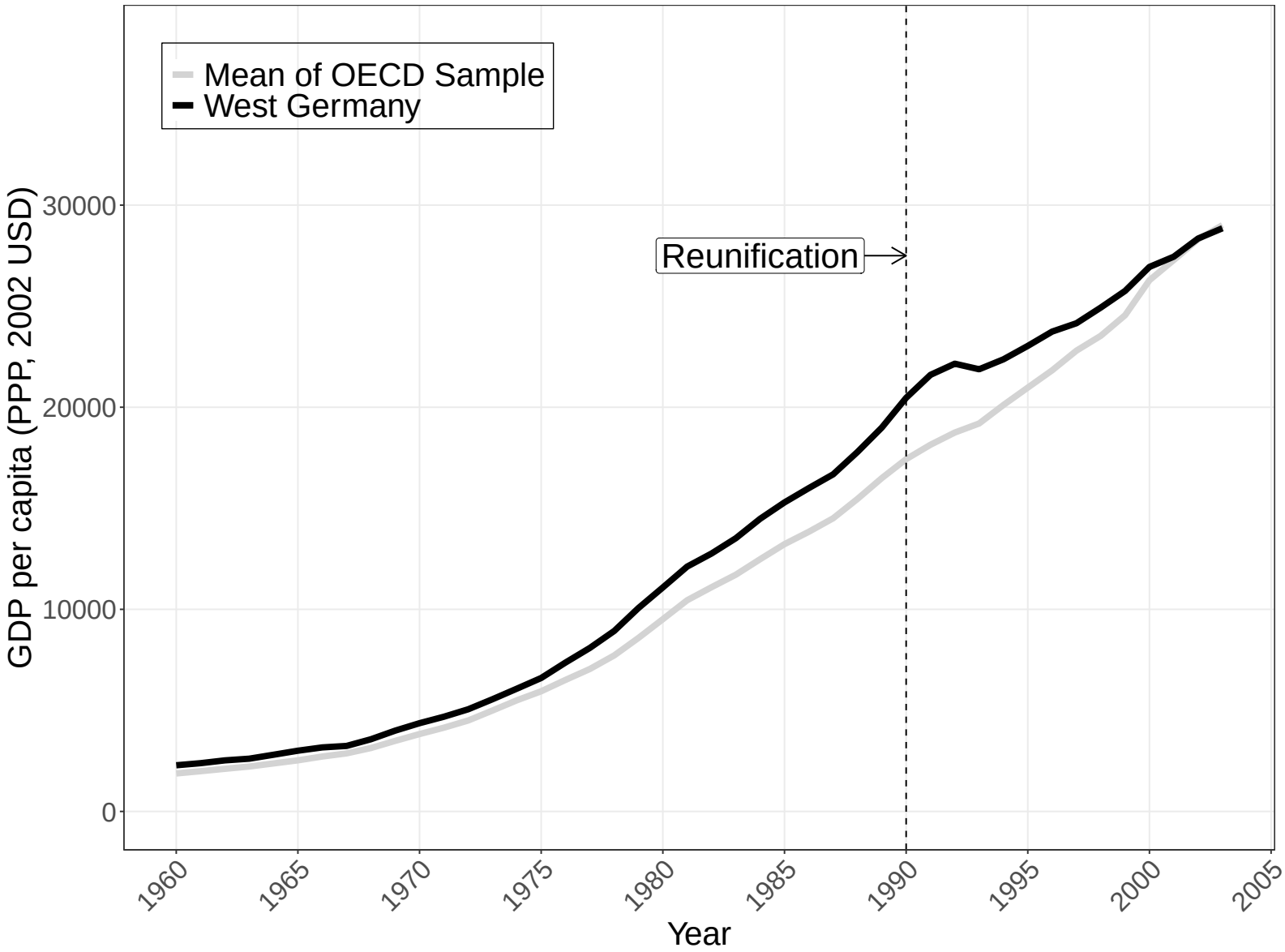
- **Geographical/cultural**: Austria?
- **Economic**: USA?
- **Average**: OECD countries?

The choice of the control group matters!

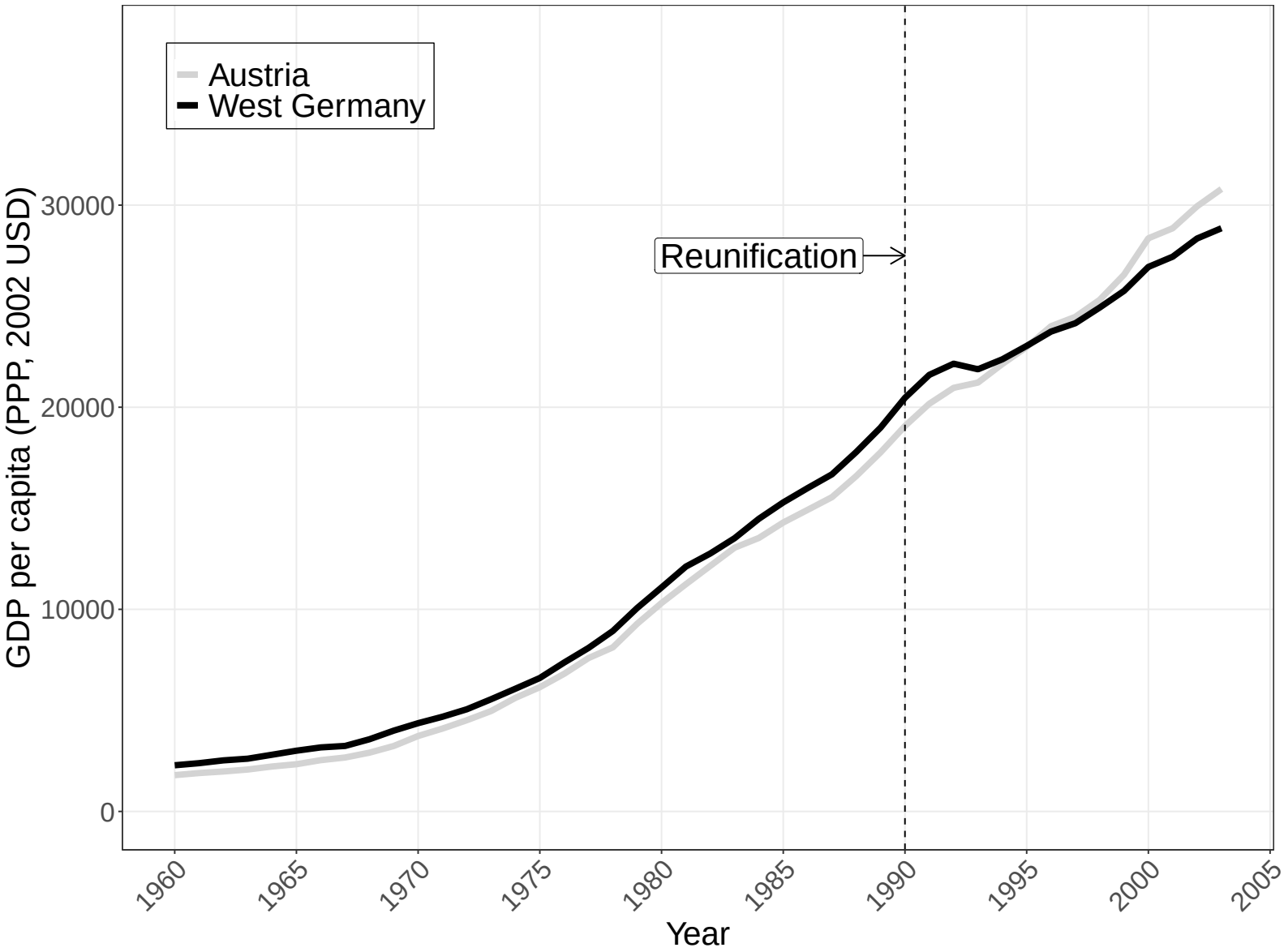
What should be the control group?



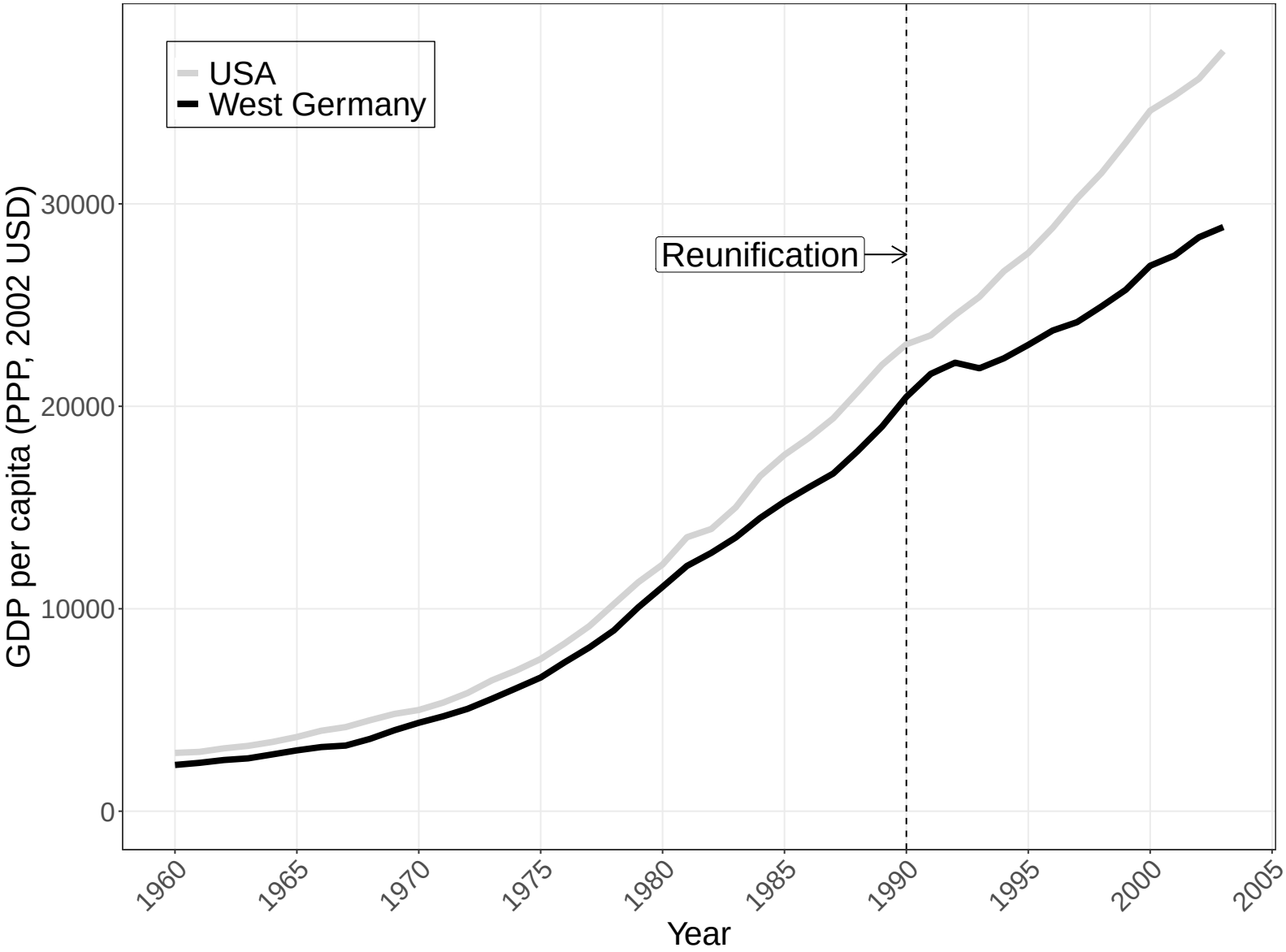
What should be the control group?



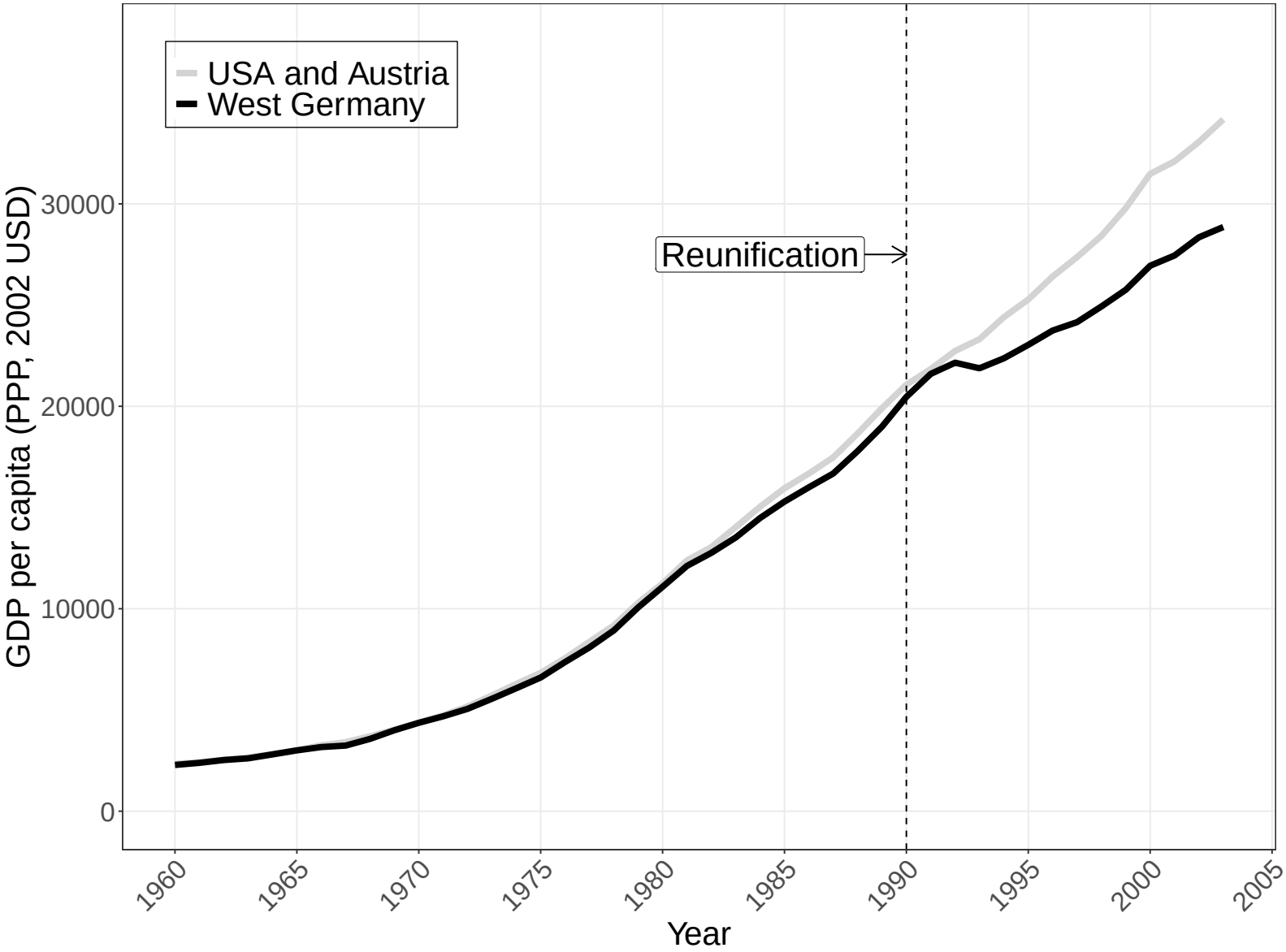
What should be the control group?



What should be the control group?



What should be the control group?



What should be the control group?

- Synthetic control moves away from using a single control unit or a simple average of control units.
- Instead we use a **weighted average** of the set of control or "donor" units.
- Rather than assuming that either the USA or Austria are similar to W. Germany, we calculate a weighted average (the synthetic control) which is more similar to West Germany than any individual country.

Intuition

When we only have a few aggregate units, a 'synthetic' combination of control units may do a better job of reproducing the characteristics of the treated unit than any one unit alone.

Synthetic Control

Notation

Definitions

For units $j \in 1, ..., J+1$:

- Unit 1 is the unit of interest (which receives the treatment)
- Units 2 to $J+1$ are the 'donor pool' or potential comparison units

Time periods $\in 1, ..., T$:

- Pre-treatment period: $t = 1, ..., T_0$
- Post-treatment period: $t = T_0 + 1, ..., T$

Notation

Definitions

Potential outcomes:

- Y_{it}^N = outcome for unit i at time t in the absence of the intervention
- Y_{it}^I = outcome for unit i at time t when exposed to the intervention

Time invariant characteristics of the units:

- Z_i = the characteristics of unit i
- Z_i can be/typically is a vector (of many covariates) but could also be fixed effects.

The target quantity

Estimand

$$\tau_{1t} = Y_{1t}^{I} - Y_{1t}^{N} = Y_{1t} - Y_{1t}^{N} \qquad \textit{for all} \qquad t > T_0$$

*i.e. the treatment effect on the **treated unit** in the post-treatment periods.*

Problem

We cannot observe Y_{1t}^{N}.
Why? $\rightarrow$ Fundamental problem of causal inference.

$\Rightarrow$ The critical question, as always, is how should we impute Y_{1t}^{N}?

Imputing Y_{1t}^N

1. **Matching**
   - For each time period t, find the M 'closest' units to unit 1 and average the observed outcomes:

$$Y_{1,t=1}^N = \frac{1}{M} \sum_{m=1}^{M} Y_{j_{m(1)},t=1}$$

2. **Diff-in-diff**
   - Add the average change in outcome for the control group to the treated unit's outcome in the pre-treatment period

$$Y_{1,t=1}^N = Y_{1,t=0} + \left(\bar{Y}_{0,t=1} - \bar{Y}_{0,t=0}\right)$$

3. **Synthetic control**
   - Take a weighted average of the outcomes of the donor units
   - Weights defined by closeness to the trend of the outcome for the treated unit in the pre-treatment period

$$Y_{1,t=1}^N = \sum_{j=2}^{J+1} w_j^* Y_{j,t=1}$$

Defining the synthetic control

Definition: Synthetic Control

A synthetic control is a vector of weights, W, associated with each of the available J donor units.

Going back to our three examples above: W is a vector with...

- ...equal weight for each unit (**OECD average**)
- ...0 weight for all units, except Austria where $w_j = 1$ (**Austria**)
- ...0 weight for all units, except USA where $w_j = 1$ (**USA**)
- ...0 weight for all units, except USA where $w_j = .5$ and Austria where $w_j = .5$ (**USA and Austria**)

There are many potential synthetic controls!

The goal is to select W such that the characteristics of the treated unit are best resembled by the characteristics of the synthetic control.

Estimating W

For each donor unit, define a weight $W = \{w_2, w_3, ..., w_{J+1}\}$, where:

$$\sum_{i=2}^{J+1} \hat{w}_j = 1 \qquad \text{and} \qquad 0 \leq w_j \leq 1 \quad \forall j \in 2, ..., J+1$$

Goal: Find values for w_j which make treatment and control units as similar as possible.

Estimating W

We want w_j such that treatment/control units are similar in terms of:

1. Pre-intervention outcome values

$$Y_{1,t} \approx \sum_{j=2}^{J+1} \hat{w}_j Y_{j,t} \qquad \text{for all} \qquad t \in 1, \ldots, T_0$$

2. Covariates that are predictive of post-intervention outcomes

$$Z_1 \approx \sum_{j=2}^{J+1} \hat{w}_j Z_j$$

The idea is to give **more weight** to units in the donor pool that **closely approximate the treated unit in the pre-intervention period**.

Crucial design decisions

1. Which variables should be included in Z_i?
   - Those that reflect the most important determinants of the outcome
   - Can use either time-varying or time-invariant covariates (`R` will average the time-varying values)
   - Remember in a time fixed-effect design we assume we account for time-varying covariates (becomes problematic if we think covariates would interact differently pre and post treatment)
2. Which units should be included in the donor pool?
   - Units whose outcome is determined in the same way as the treated unit
   - Control units should not become treated in any of the post-treatment period
   - Control units should not be subject to idiosyncratic shocks in the post treatment period

Estimating W

We find the values of W by minimizing the following expression:

$$\sum_{m=1}^{k} v_m (X_{1m} - X_{0m} W)^2$$

where

1. $X_1 = \{Z_1, Y_{1,1}, Y_{1,2}, ..., Y_{1,T_0}\}$ and X_0 is a matrix containing the same information for each of the control units
2. v_m is a weight that reflects the importance of the m^{th} variable that we use to measure the distance between treated and control units

Estimating v_m

We also need to establish which *variables* get the largest weights (v_m). To do so, we use cross-validation:

1. Split the pre-treatment period into a training period (1960-1980) and a validation period (1981-1990)
2. Using training period data, select v_m such that W minimizes the **root mean squared prediction error** for the validation period starting at t_{cv}

$$RMSPE_{cv} = \sqrt{\frac{1}{T_0 - t_{cv-1}} \sum_{t=t_{cv}}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} \hat{w}_j Y_{jt} \right)^2}$$

Estimating v_m

Implications

1. Selects v_m that minimizes out-of-sample prediction errors
2. v_m indicate which covariates are most predictive of the outcome
3. Most weight (w_j) is put on control units which are similar to the treated units on covariates $(Z_1,\ Z_0)$ that are predictive of the outcome $(Y_{1,t}, Y_{2,t}, \ldots, Y_{J+1,t})$ in the pre-intervention period $(t \leq T_0)$

Nothing new?

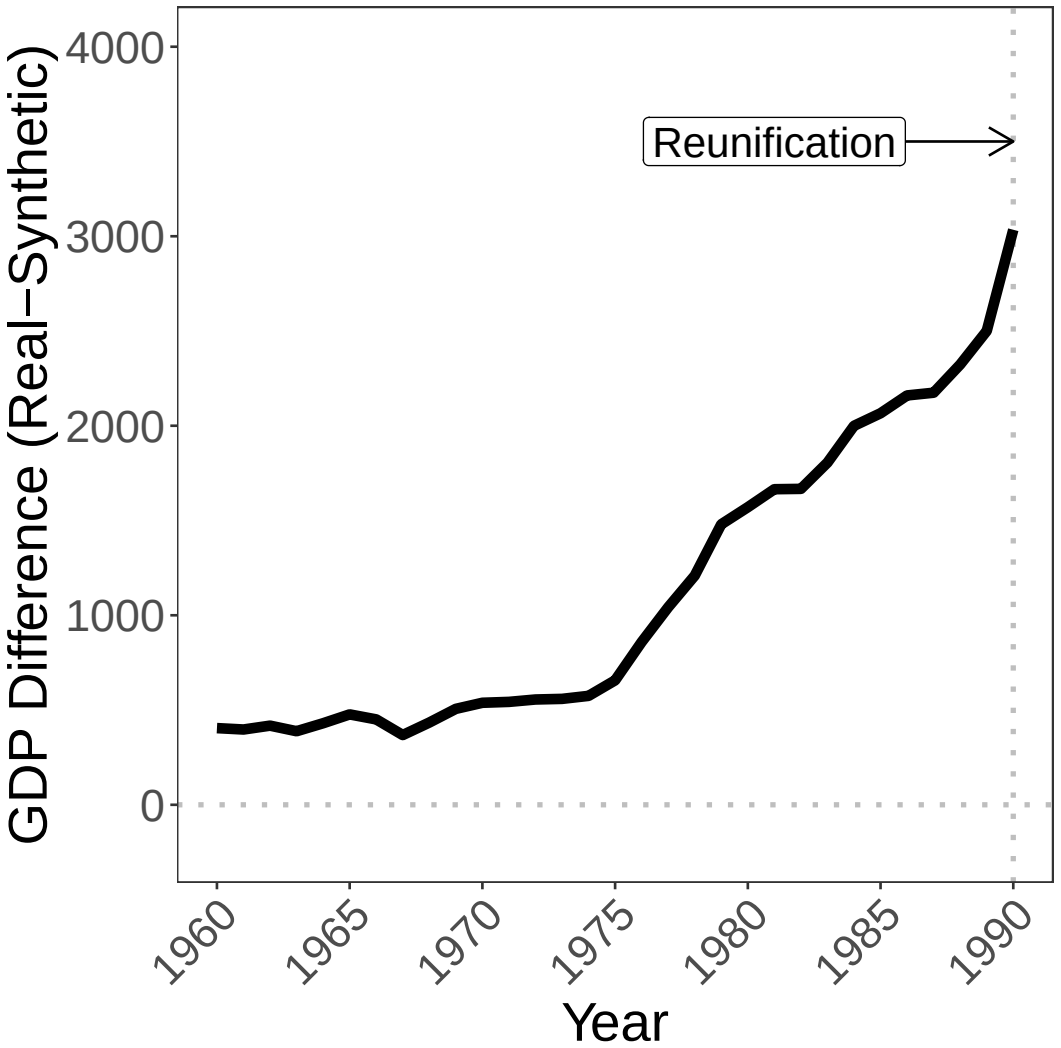
SC is, at heart, a sort of **difference-in-differences matching** estimator

- Diff-in-diff: establish a control group which follows a parallel trend in the absence of treatment (note we are still assuming that the parallel-ness would continue post-treatment!)
- Matching: w_j calculated using **observed** pre-treatment covariates

SC tries to find the weighted counterfactual that minimises the distance, in terms of time-invariant characteristics and pre-treatment outcomes, between the treated unit and the synthetic control.

Estimating W (Intuition)

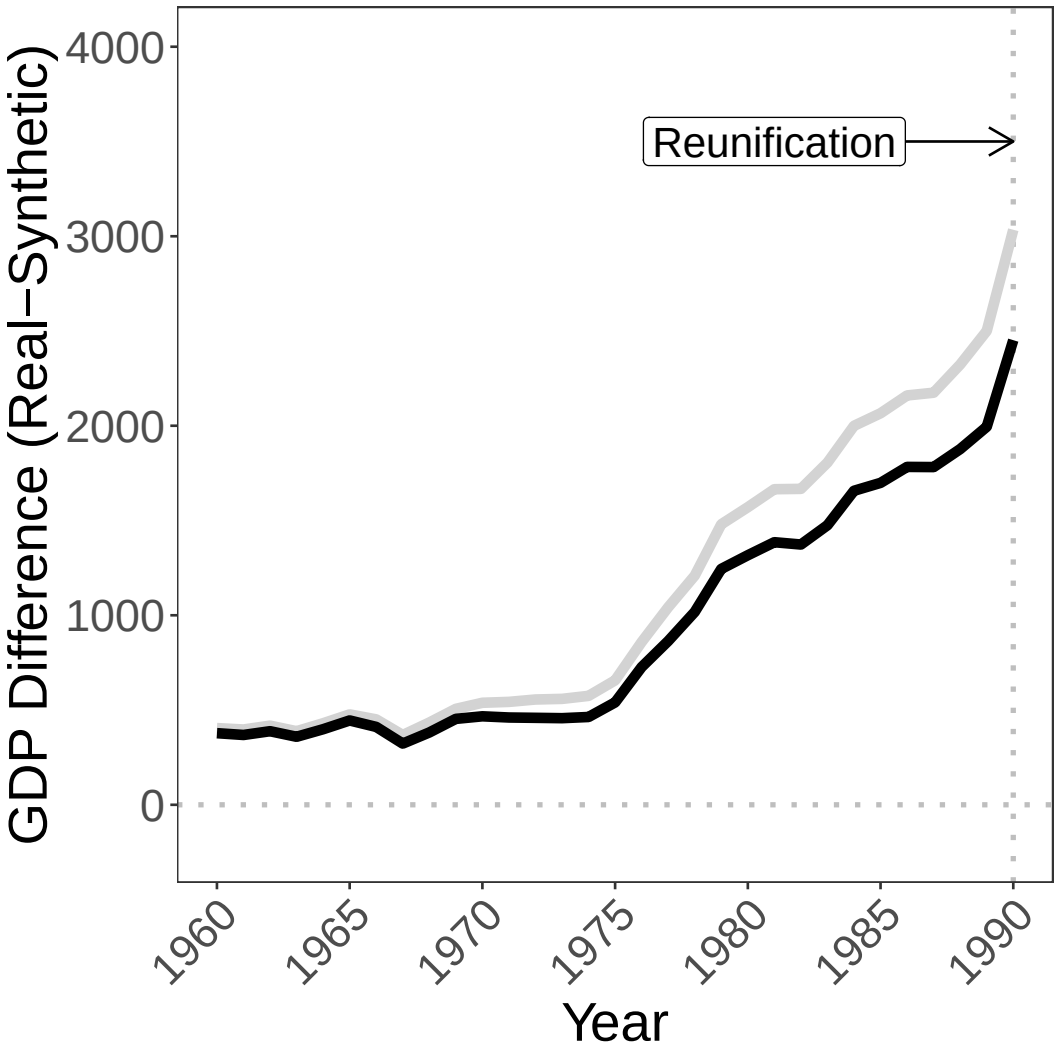
Goal: minimize difference in outcome trend in pre-treatment period.



country	weights
USA	0.06
UK	0.06
Austria	0.06
Belgium	0.06
Denmark	0.06
France	0.06
Italy	0.06
Netherlands	0.06
Norway	0.06
Switzerland	0.06
Japan	0.06
Greece	0.06
Portugal	0.06
Spain	0.06
Australia	0.06
New Zealand	0.06

Estimating W (Intuition)

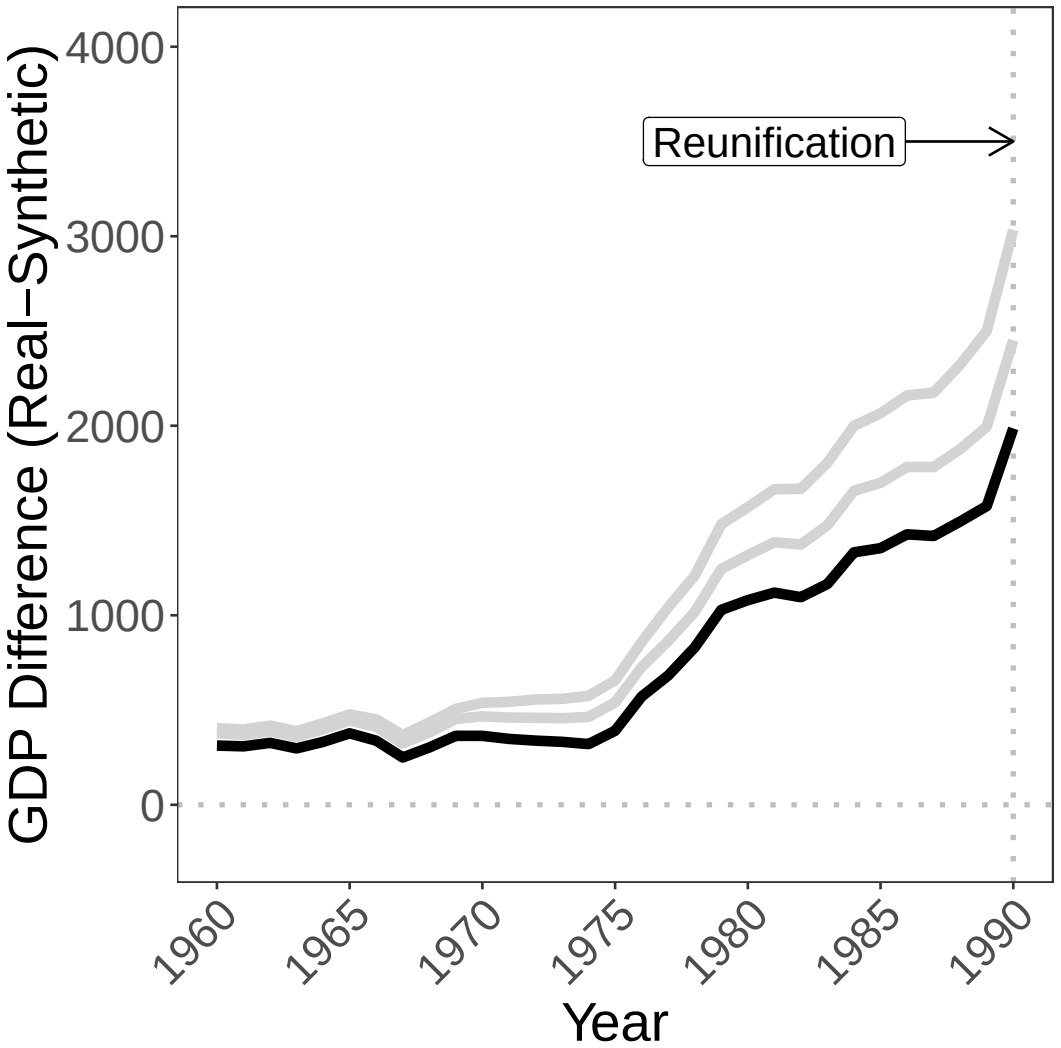
Goal: minimize difference in outcome trend in pre-treatment period.



country	weights
Austria	0.13
USA	0.10
Japan	0.08
Switzerland	0.07
Italy	0.06
Spain	0.06
Greece	0.06
Netherlands	0.05
UK	0.05
Belgium	0.05
Denmark	0.05
France	0.05
Portugal	0.05
Norway	0.04
Australia	0.04
New Zealand	0.04

Estimating W (Intuition)

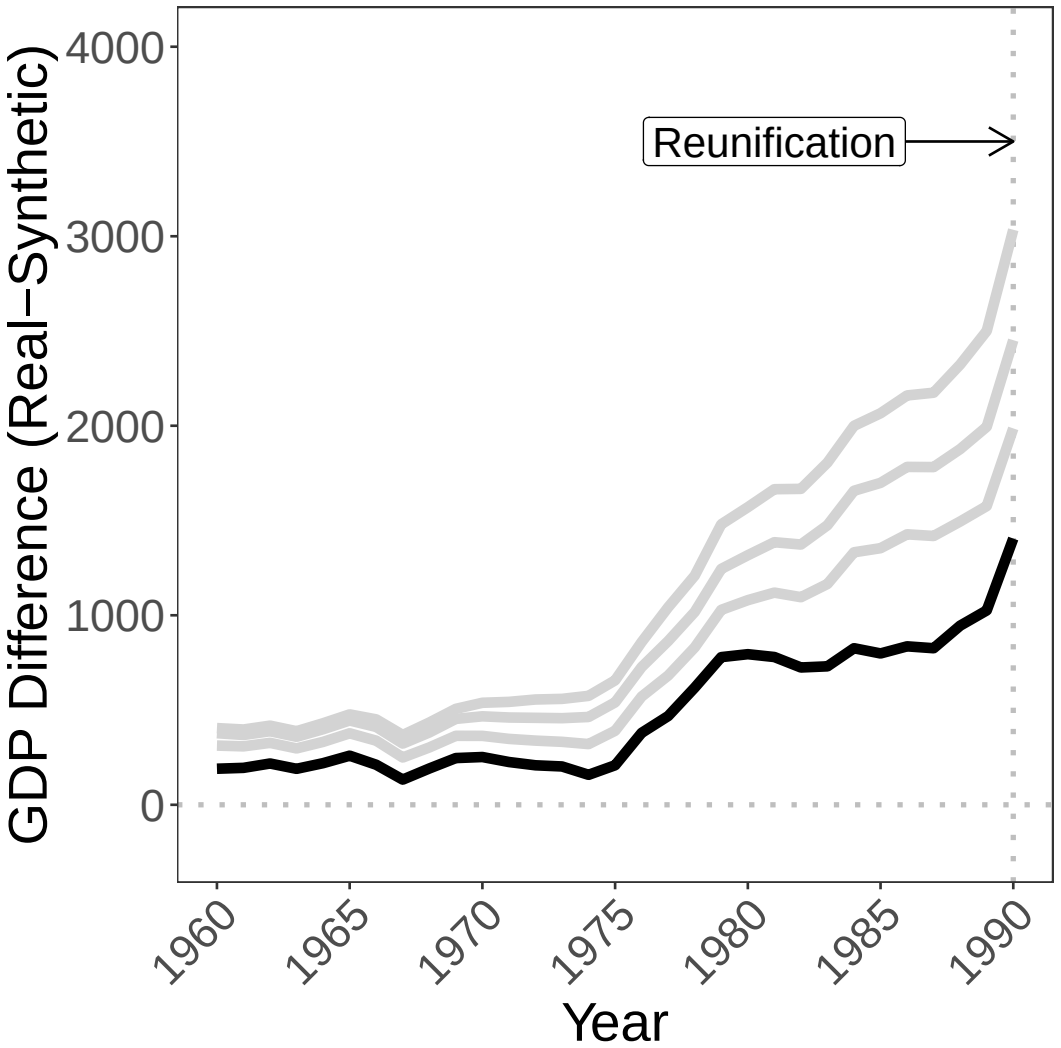
Goal: minimize difference in outcome trend in pre-treatment period.



country	weights
Austria	0.18
USA	0.11
Japan	0.09
Switzerland	0.08
Netherlands	0.07
UK	0.04
France	0.04
New Zealand	0.04
Belgium	0.04
Denmark	0.04
Italy	0.04
Spain	0.04
Australia	0.04
Greece	0.04
Portugal	0.04
Norway	0.04

Estimating W (Intuition)

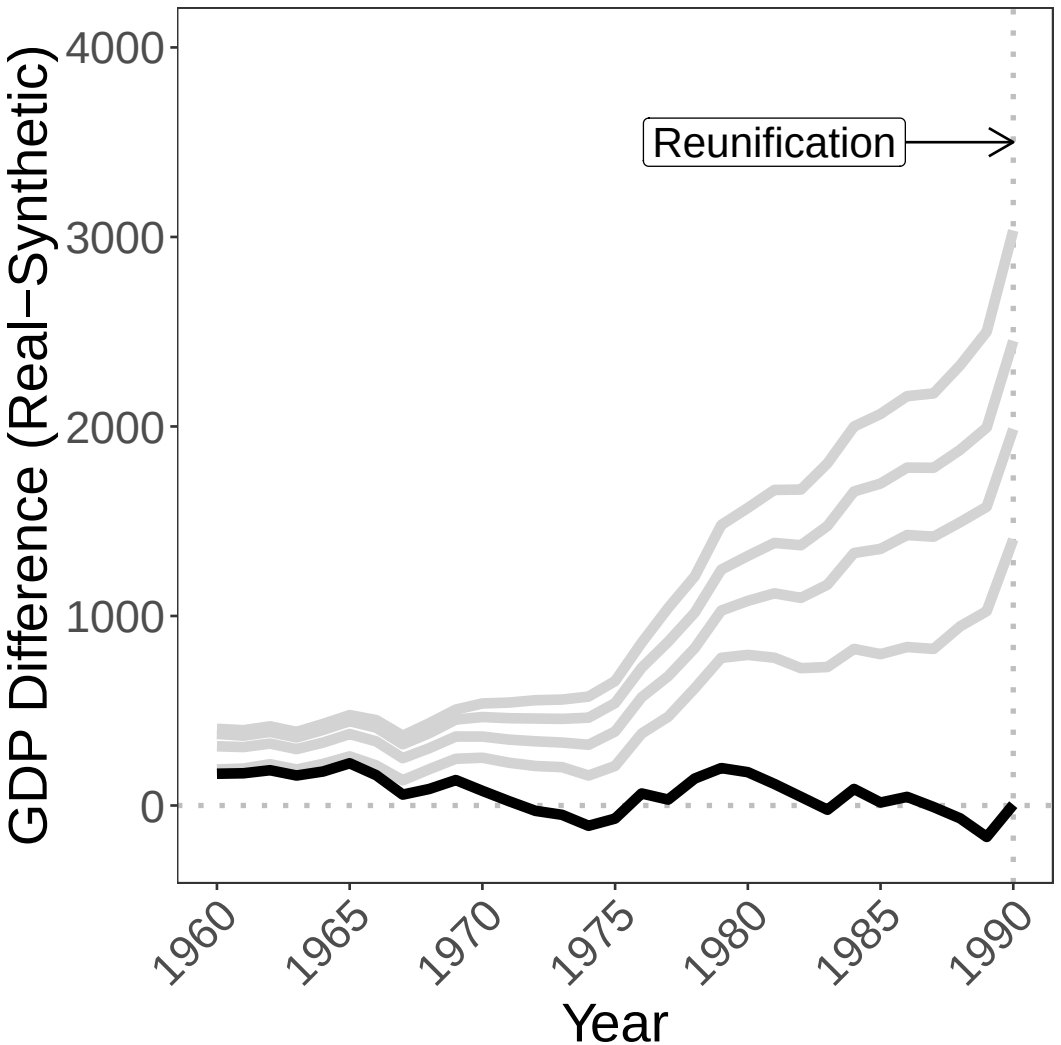
Goal: minimize difference in outcome trend in pre-treatment period.



country	weights
Austria	0.24
USA	0.14
Japan	0.11
Norway	0.09
New Zealand	0.09
Switzerland	0.09
Netherlands	0.08
UK	0.03
Belgium	0.03
Denmark	0.03
Portugal	0.03
Australia	0.03
France	0.00
Italy	0.00
Spain	0.00
Greece	0.00

Estimating W (Intuition)

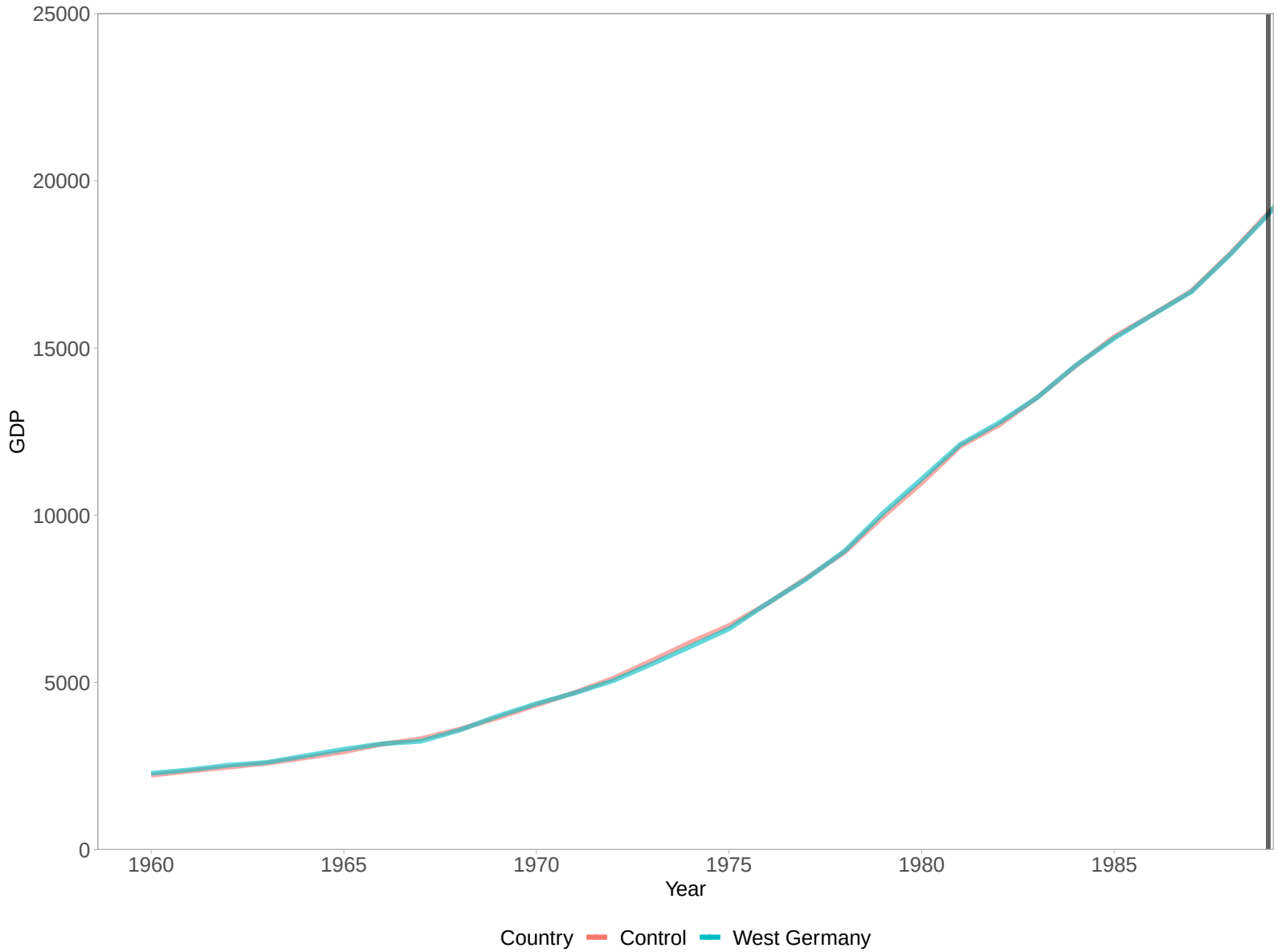
Goal: minimize difference in outcome trend in pre-treatment period.



country	weights
Austria	0.42
USA	0.22
Japan	0.16
Switzerland	0.11
Netherlands	0.09
UK	0.00
Belgium	0.00
Denmark	0.00
France	0.00
Italy	0.00
Norway	0.00
Spain	0.00
Greece	0.00
Portugal	0.00
Australia	0.00
New Zealand	0.00

We will use an `R` package which automates this optimization problem for us.

Estimating W (Intuition)



Interpreting the country weights W

country	SC weights	OLS weights
Austria	0.42	0.26
USA	0.22	0.13
Japan	0.16	0.19
Switzerland	0.11	0.05
Netherlands	0.09	0.14
UK	0.00	0.06
Belgium	0.00	0.00
Denmark	0.00	0.08
France	0.00	0.04
Italy	0.00	-0.05
Norway	0.00	0.04
Spain	0.00	-0.01
Greece	0.00	-0.09
Portugal	0.00	-0.08
Australia	0.00	0.12
New Zealand	0.00	0.12

- Regression weights can be greater than 1 or less than zero $\rightarrow$ extrapolation outside of the support of control units.
- Extrapolation is not possible in the SC case because the weights are bound between 0 and 1.
- Recall that the sum of all control weights will be equal to 1 and some units can have a weight of 0 (i.e. they do not feature in our new **synthetic** control unit.

Assessing balance

GDP predictor means:

	Treated	Synthetic	Rest of OECD Sample
GDP per-capita	15808.900	15802.240	8021.1
Trade openness	56.778	56.939	31.9
Inflation rate	2.595	3.495	7.4
Industry share	34.538	34.387	34.2
Schooling	55.500	55.180	44.1
Investment rate	27.018	27.034	25.9

Interpreting v

Which variables are most important for determining the synthetic control?

Predictor weights:

variable	v
GDP per-capita	0.442
Investment rate	0.245
Trade openness	0.134
Schooling	0.107
Inflation rate	0.072
Industry share	0.001

The weights $v_1, \ldots, v_k$ reflect the predictive value of the covariates.

Causal effects

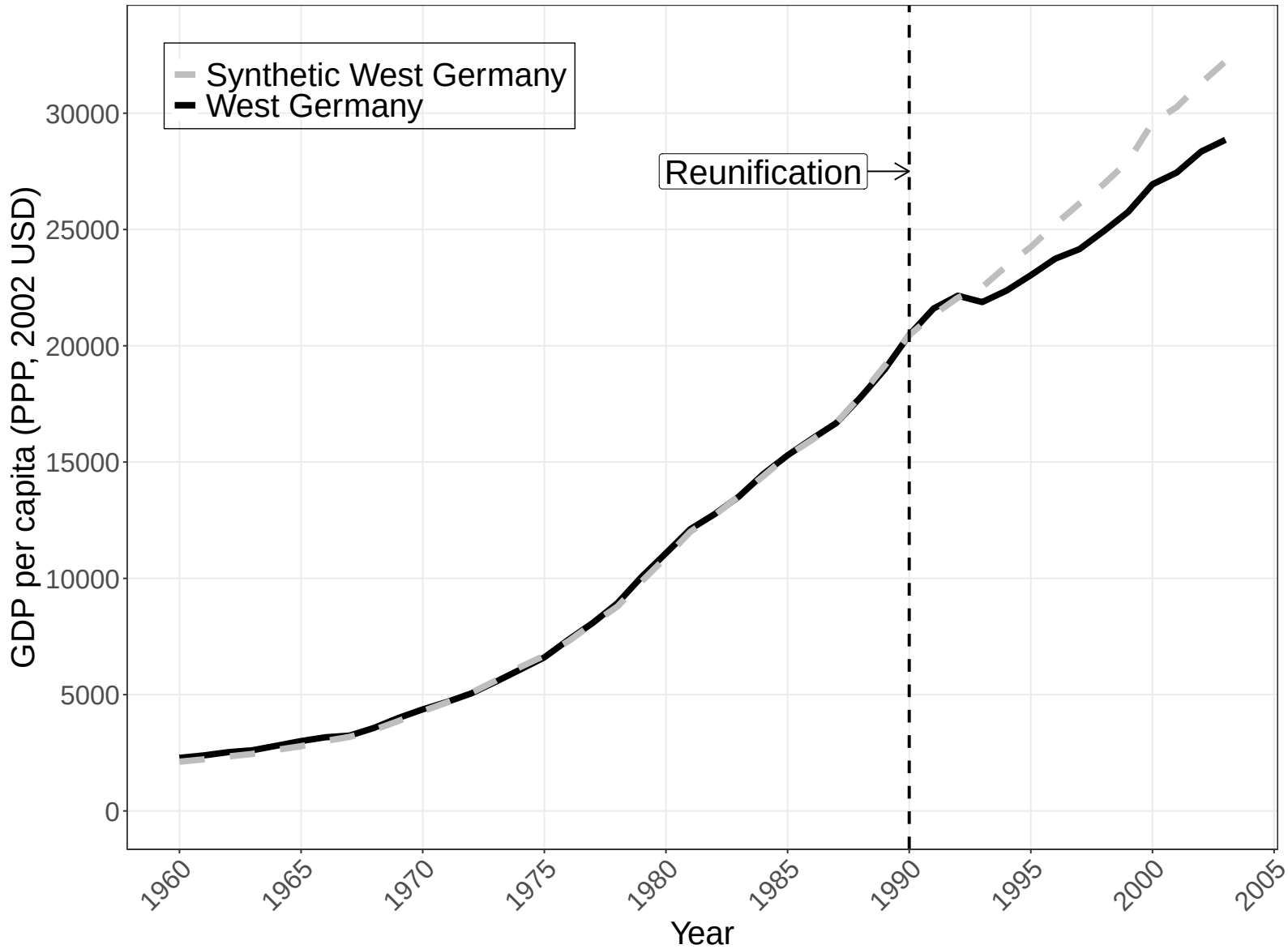
- Weighting donor units leads to a synthetic unit with a similar outcome trend **in the pre-intervention** period as the treated unit.
- Given $\hat{w}$, an unbiased estimator of τ_{1t} is:

$$\hat{\tau}_{1t} = Y_{1,t} - \sum_{j=2}^{J+1} w_j Y_{j,t} \qquad \text{for} \qquad t \in \{T_0 + 1, \ldots, T\}$$

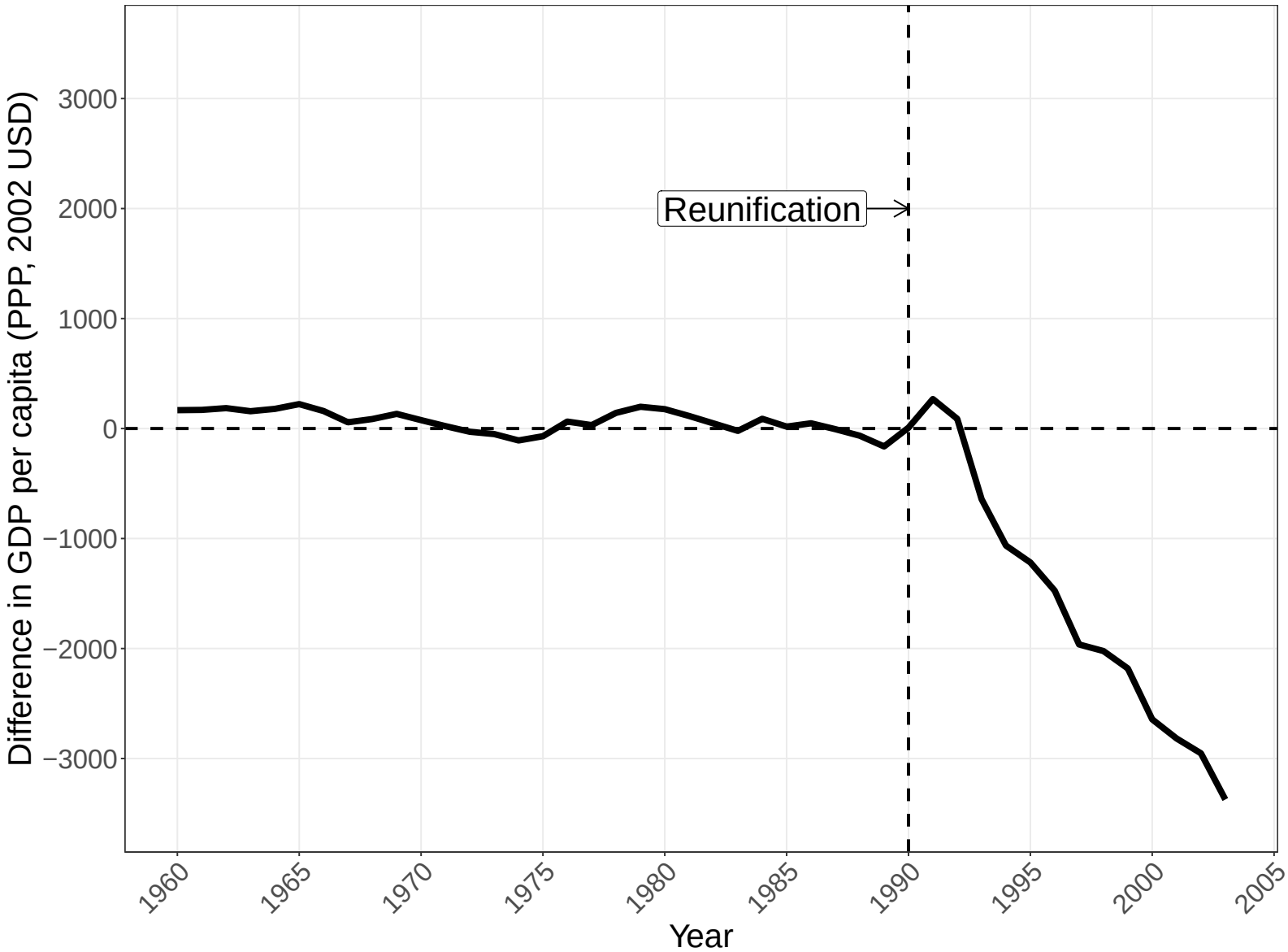
where

  - $Y_{1,t}$ is the outcome for the **treated unit** *in post-treatment period* t
  - $\sum_{j=2}^{J+1} w_j Y_{j,t}$ is the outcome for the **synthetic control unit** *in post-treatment period* t
  - $\hat{\tau}_{1,t}$ is the ATT for time period t

Causal effects graphically



Causal effects graphically



So what is the causal effect?

- We always speak in terms of the *average* treatment effects.
  - But here we only **one** treated unit
  - We do have more than one treated period
- If average this across all post-treatment periods, then we are estimating the **Average Treatment Effect on the Treated unit** $i = 1$ (ATT_1)
- If we look at each time period separately, we are estimating the **Treatment Effect on the Treated unit** i **for each time period** t (TT_{1t})

So what is the causal effect?

	Difference in GDP
TT for 1989	-163.060
TT for 1990	8.030
TT for 1991	268.210
TT for 1992	88.280
TT for 1993	-642.610
TT for 1994	-1064.120
TT for 1995	-1217.450
TT for 1996	-1474.310
TT for 1997	-1963.000
TT for 1998	-2022.840
TT for 1999	-2181.260
TT for 2000	-2644.020
TT for 2001	-2817.610
TT for 2002	-2952.750
TT for 2003	-3372.810
ATT	-1476.756

- The $ATT_{germany}$ across all post-treatment periods is -1477.
- Formally written: *The average treatment effect on GDP caused by German reunification was a loss of $1477 in Germany between 1989 to 2003*
- Or: *If there hadn't been German reunification, we would have expected average GDP to have been $1477 greater in Germany that it would otherwise had been.*

Identification assumption

- Under which conditions can we state that the numbers from the previous slides are *credible causal estimates*?
  - Remember, the Synthetic Control Method is a sort of **difference-in-differences matching** estimator.
- So the identification assumption is similar to the one for DiD
  - If the treated **unit** had not received the treatment, it would have followed the **exact same trend** as the **synthetic control unit** (cf. lecture 5, slide 21)

Identification Assumption

$$Y_{1t}^N = \sum_{j=2}^{J} w_j Y_{jt}^N \; \forall \; t > T_0 \quad \textit{(equal trend)}$$

i.e. the potential untreated outcomes of the treated unit are the same as the weighted average of the untreated potential outcomes of the untreated units for all post-treatment periods.

Inference & Placebo Tests

Asymptotic inference in synthetic control

- Standard errors from regression/t-tests are typically used to characterise uncertainty about aggregate data:
  - i.e. use a sample of restaurants in NJ and PA to estimate employment trends in each state
  - standard errors reflect unavailability of aggregate data on employment
- So, if we use aggregate data, is there zero uncertainty? No!
  - We do not have perfect information about **potential** outcomes, even when we use aggregate data
  - We have uncertainty about the potential outcome under control for the treated unit
- But, because the number of units is small in most SC applications, large sample inferential techniques are not appropriate.

Permutation inference

Instead, we turn to an alternative inference technique: **permutation inference.**

1. Calculate the the test-statistic under the actual treatment assignment
2. Calculate the distribution of the test-statistic under alternative treatment assignments assuming treatment effects of zero
3. Assess whether the 'true' test-statistic is unlikely under the **null distribution** of treatment effects

$\Rightarrow$ Here, this implies constructing a synthetic control for **every** country in our sample, summarising the treatment effect, and comparing it to the treatment effect in West Germany.

Remembering hypothesis testing

What are we doing when testing statistical significance?

- We want to get a sense of how clear/systematic the pattern we observe in a sample is
- From this we can then infer in what range the true value in the population might or might not be
- When doing hypothesis testing, we are evaluating **how likely**[1] it is that we would see the value we see in the sample **if** the **true value** of the thing we are interested in[2] **is in fact zero**[3]
- The basic logic is the same for asymptotic inference as well as permutation inference

[1] This is expressed in the p-value
[2] Mean, correlation, treatment effect etc.
[3] This is the null hypothesis.

Permutation inference

- When we have (very) small samples, we cannot rely on asymptotic theory to tell us how the sampling distribution looks like
  - Mainly because we do not have enough data to estimate a standard error!
- So instead, we have to **build** the null-distribution ourselves, rather than assuming its shape
- This is what permutation inference does!
  - Placebo test: look at what the *estimated* treatment effect is for units where there **should not be a treatment effect**
  - If you get several other treatment effect estimates that are as high or higher than the one unit you **know** is treated, this means it is likely that you would get the value you observe, **if in fact the true effect was zero**

RMSPEs for all units

For **each unit** calculate:

$$\text{RMSPE}_{j,T_0} = \sqrt{\frac{1}{T_0}\sum_{t=1}^{T_0}\left(Y_{1,t} - \sum_{j=2}^{J+1} w_j Y_{i,t}\right)^2}$$

$$\text{RMSPE}_{j,T_1} = \sqrt{\frac{1}{T_1}\sum_{t=T_0+1}^{T}\left(Y_{1,t} - \sum_{j=2}^{J+1} w_j Y_{i,t}\right)^2}$$

Where

- RMSPE_{j,T_0} $\rightarrow$ pre-treatment difference between unit and SC
- RMSPE_{j,T_1} $\rightarrow$ post-treatment difference between unit and SC

Test statistic

Given these, the test-statistic is:

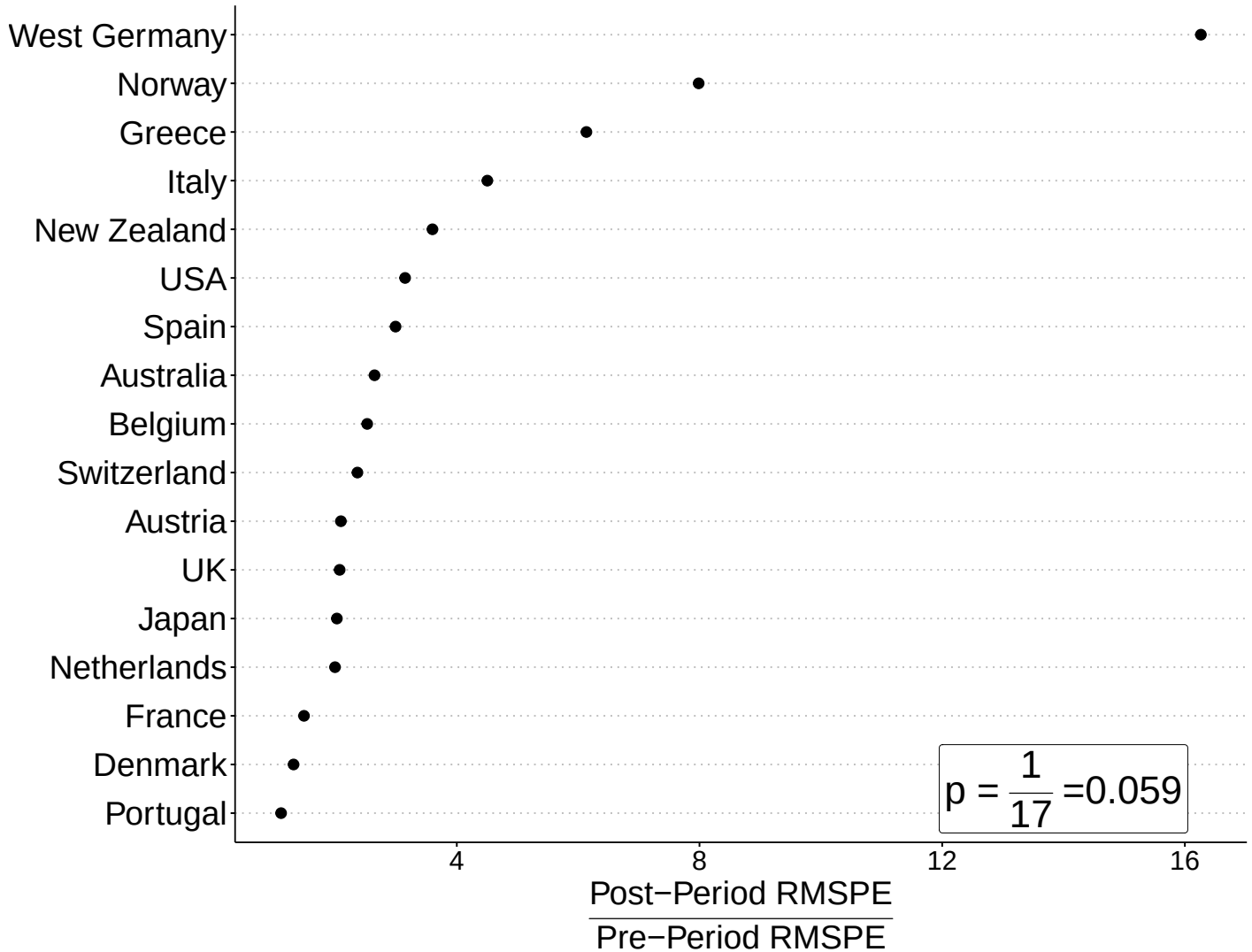
$$t_j = \frac{\text{RMSPE}_{j,T_1}}{\text{RMSPE}_{j,T_0}} = \frac{\text{Post-intervention 'fit'}}{\text{Pre-intervention 'fit'}}$$

Intuition:

- More confident that the effect is different from zero when the estimated treatment effect is larger (RMSPE_{j,T_1})
- Less confident that the effect is different from zero when the pre-treatment fit with the SC is larger (RMSPE_{j,T_0})

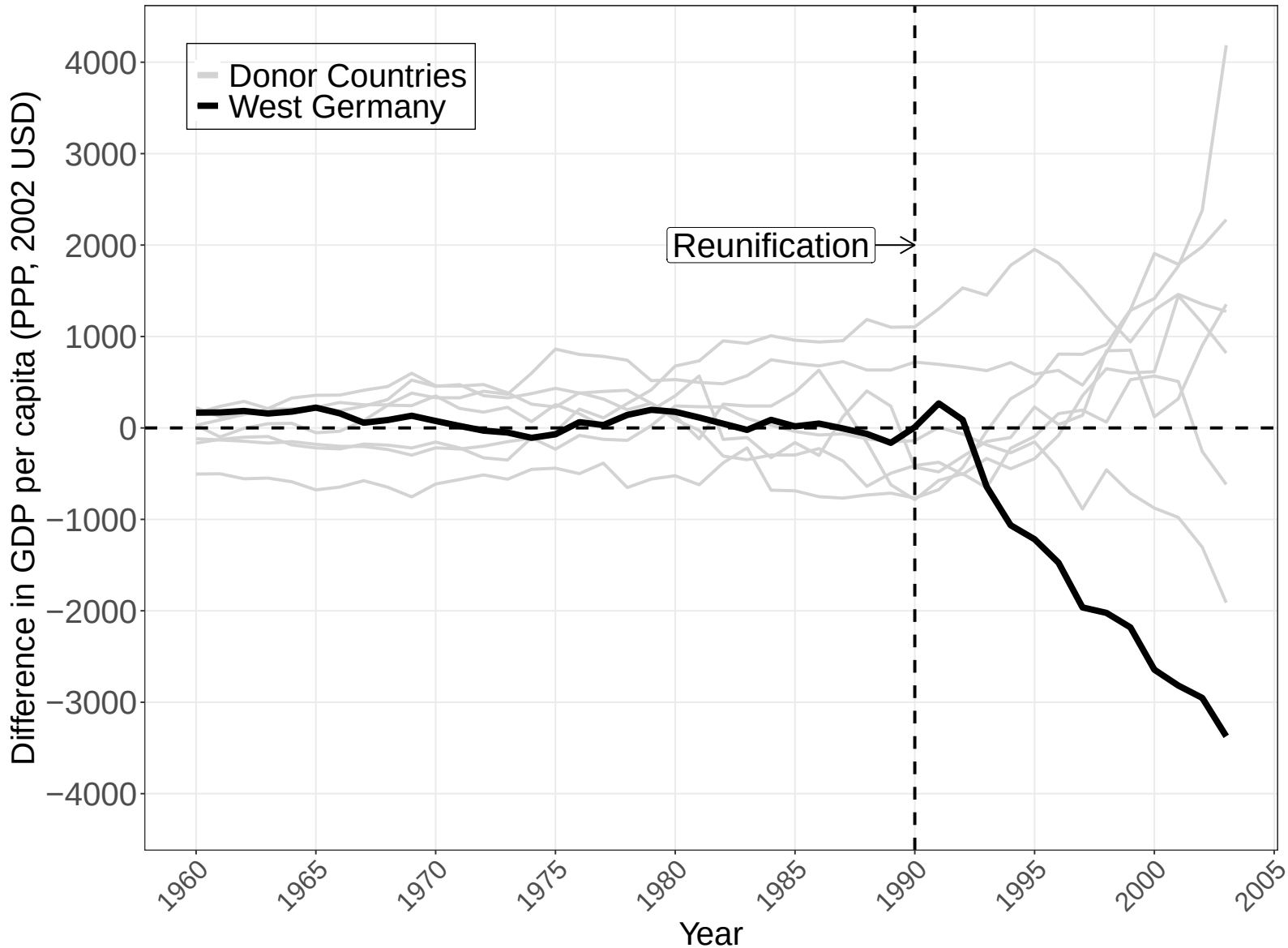
P-value *How likely would it be to observe a ratio as large as the one we actually observe if the treatment effects were zero and we picked a country at random?*

Permutation inference in practice

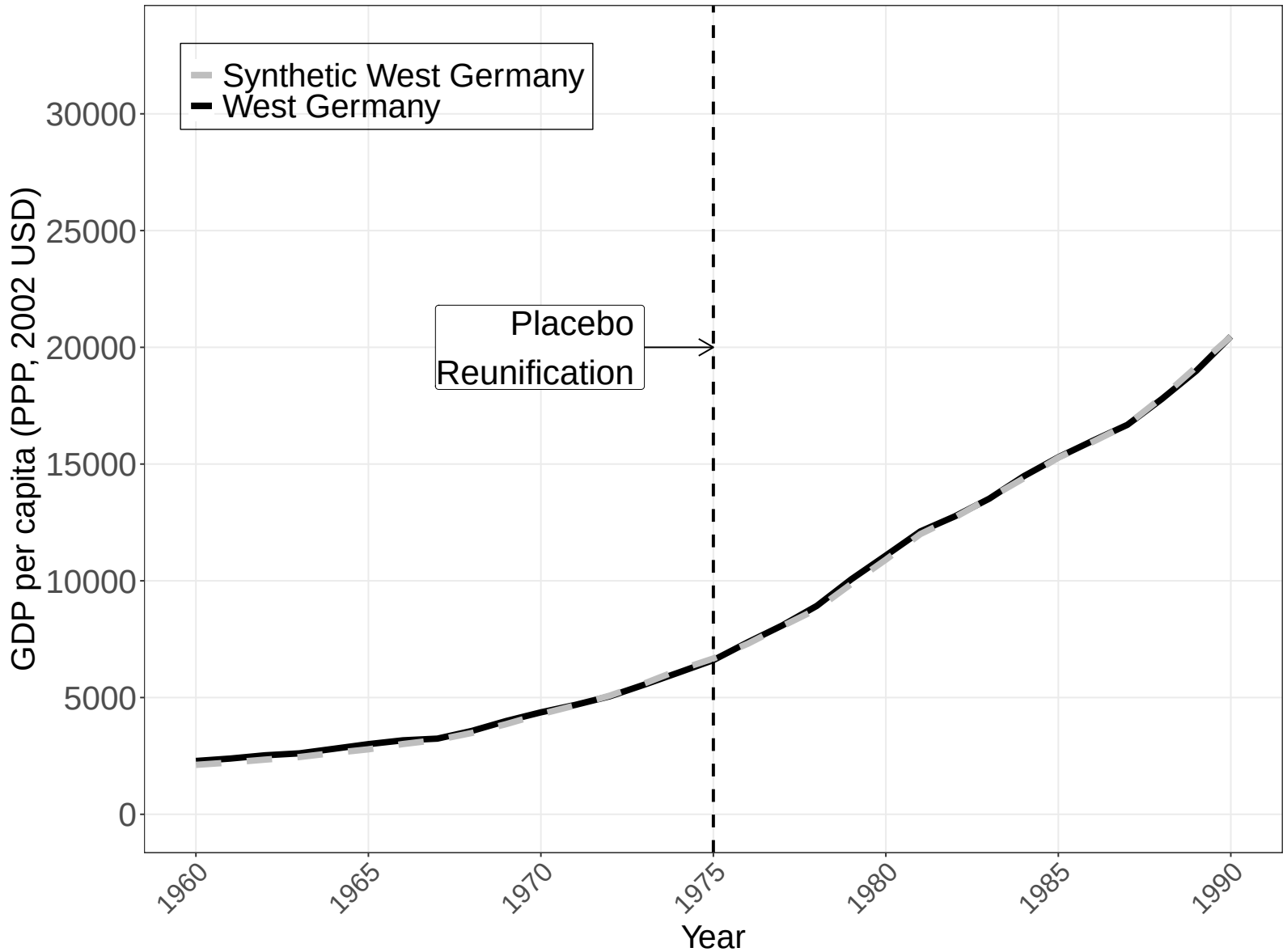


Placebos in Space

GDP gap is countries with pre-treatment $RMSPE < 5 \times RMSPE_{WGer}$



Placebos in Time



Additional Applications

California's Proposition 99

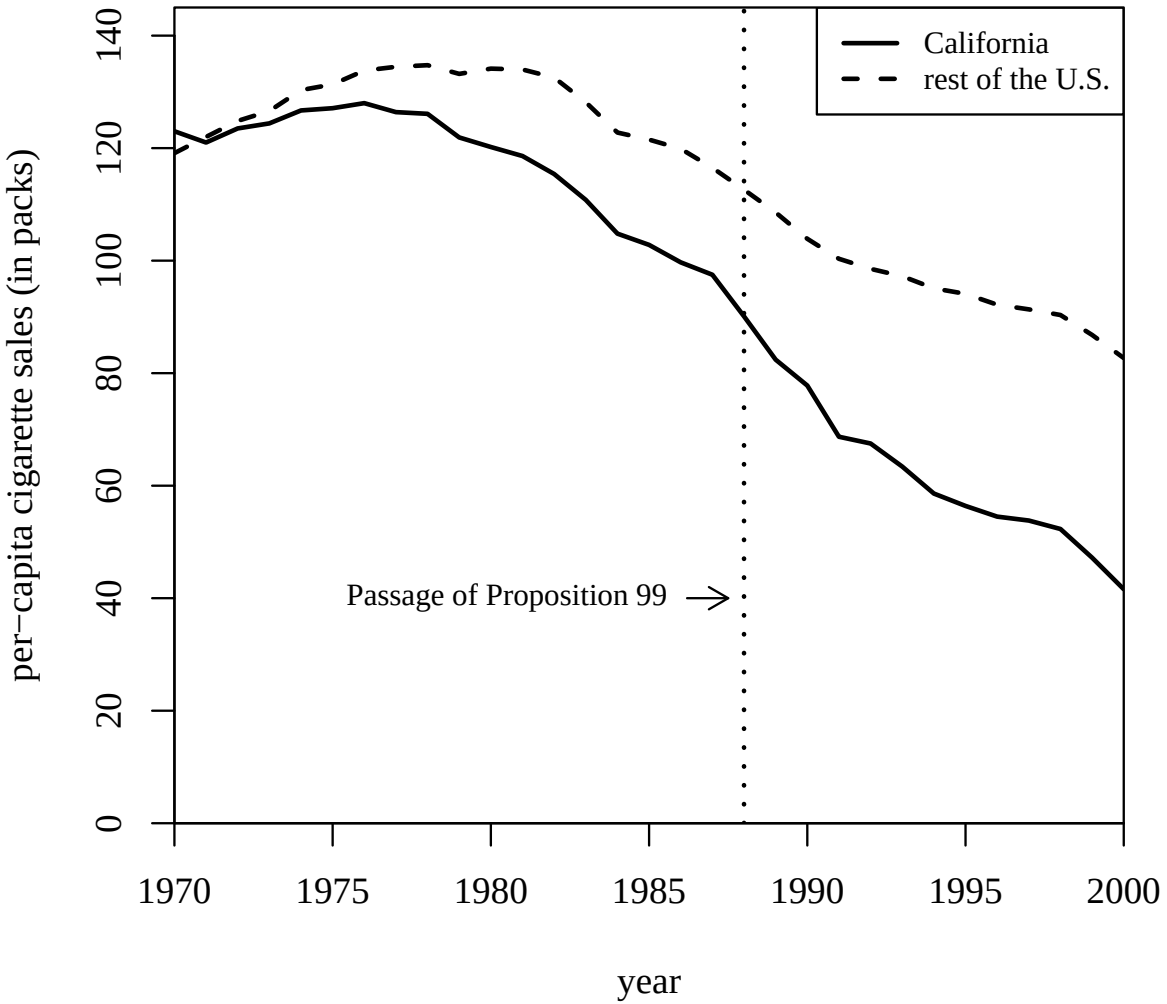
Anti-smoking legislation and cigarette consumption

In 1988, California passed comprehensive tobacco control legislation. This was a package of measures that included a tax increase, more earmarked spending to anti-smoking health initiatives, and anti-smoking media campaigns. We will investigate the effect of this legislation on cigarette consumption in California using synthetic control methods.

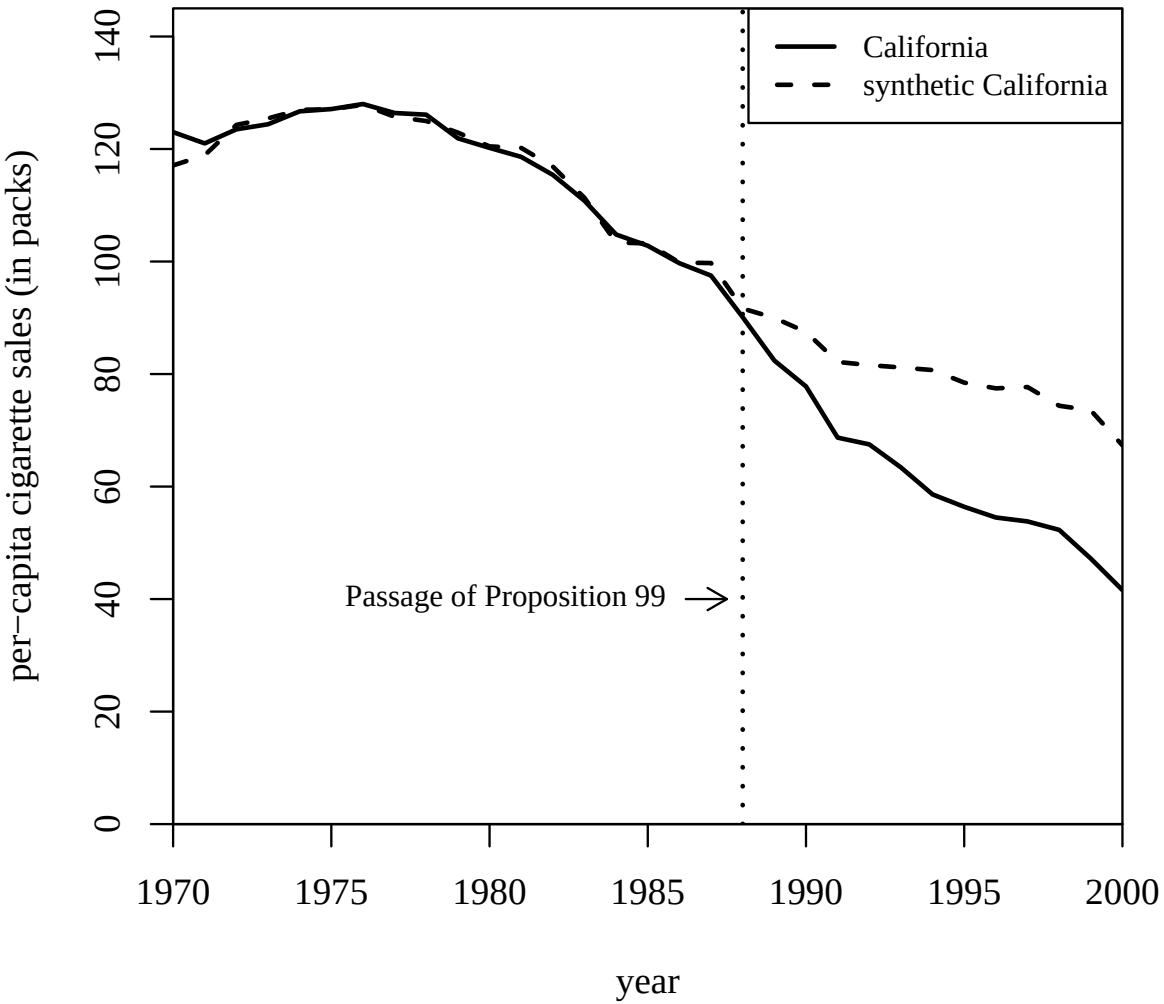
- Outcome variable (Y): Per capita cigarette sales (packs)
- Treatment (D): 1 for CA after 1988, 0 for all other periods/states[1]
- Time (T): 1970 to 2000

[1] All states which passed similar legislation are excluded from the donor pool.

California's Proposition 99



State Weights in Synthetic California

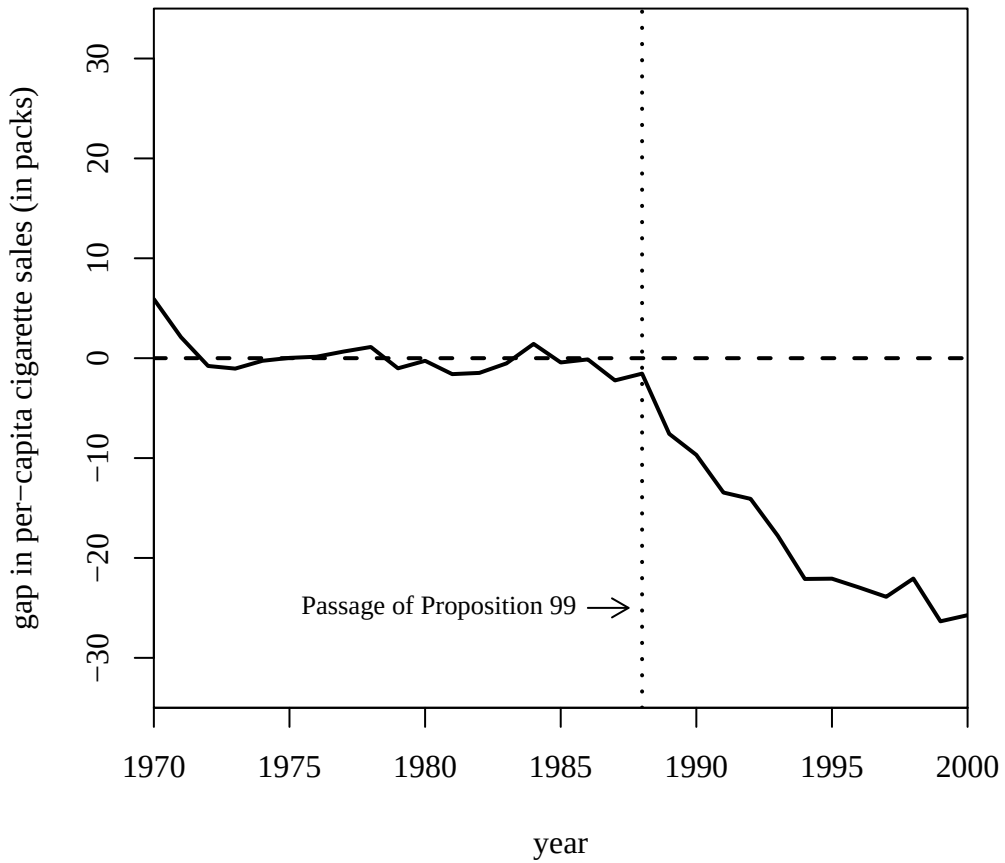


State	Weight
Utah	0.334
Nevada	0.234
Montana	0.199
Colorado	0.164
Connecticut	0.069

Real vs. Synthetic California

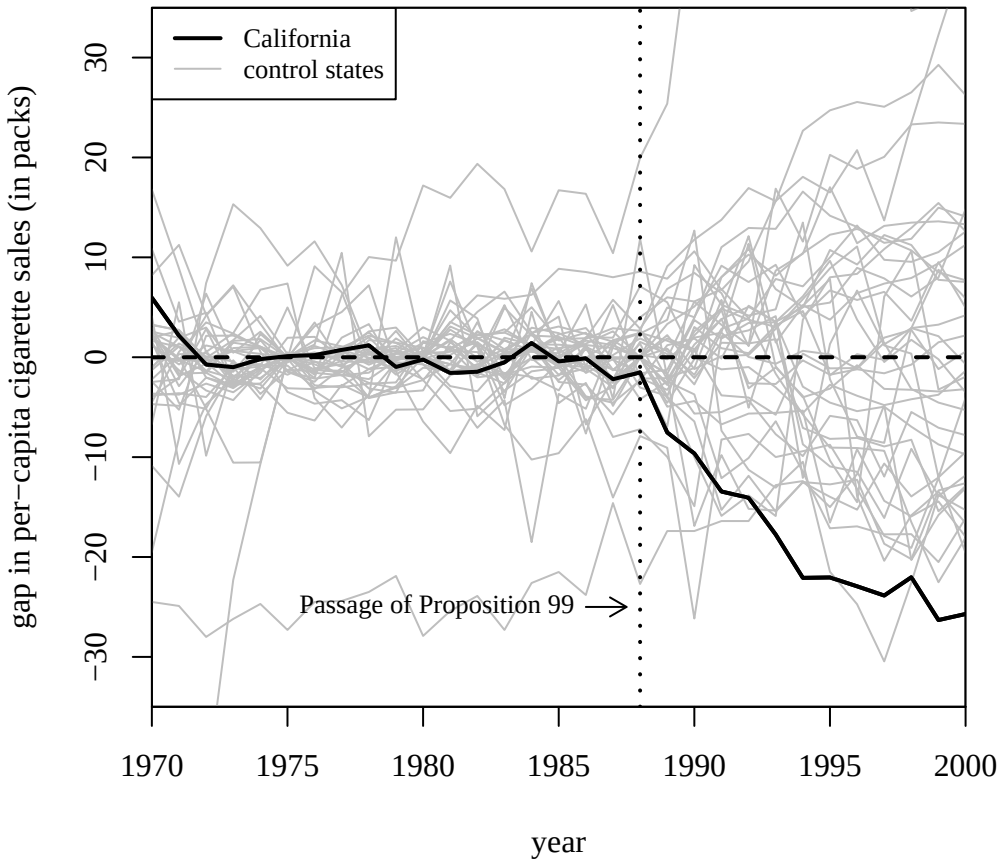
Variables	California		Average of
	Real	Synthetic	38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15-24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

Note: All variables except lagged cigarette sales are averaged for the 1980-1988 period (beer consumption is averaged 1984-1988).



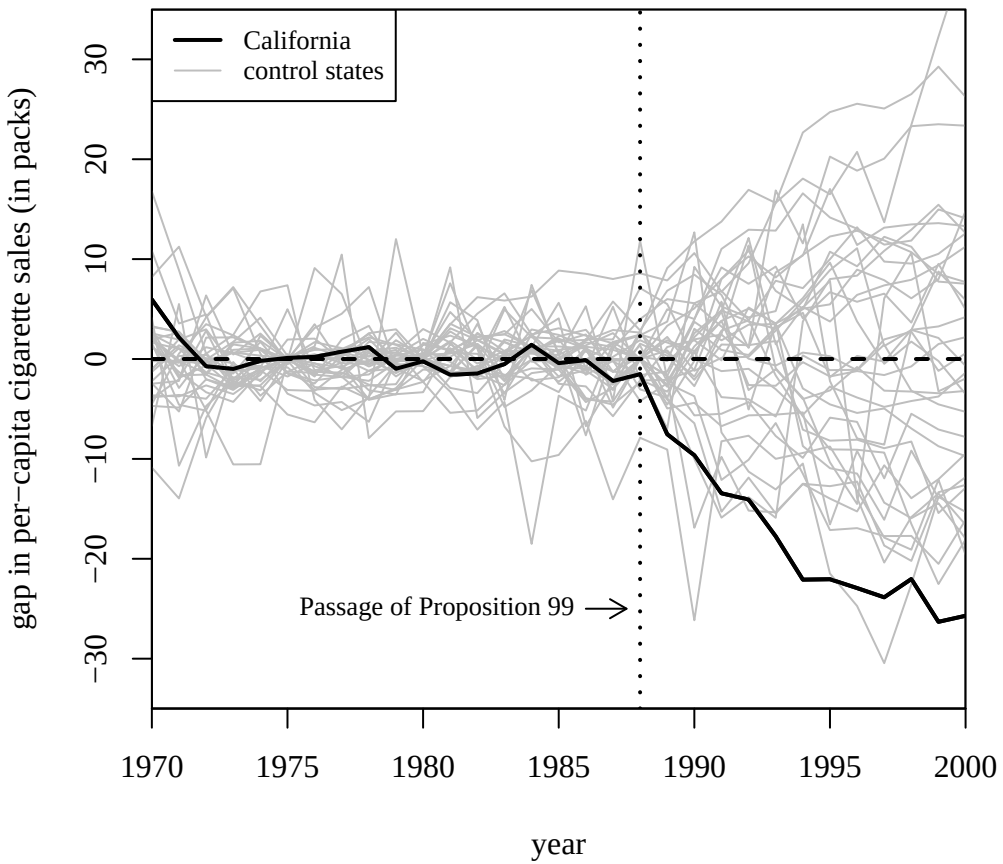
Cigarette sales gap in CA (versus synthetic CA).

Placebos in Space



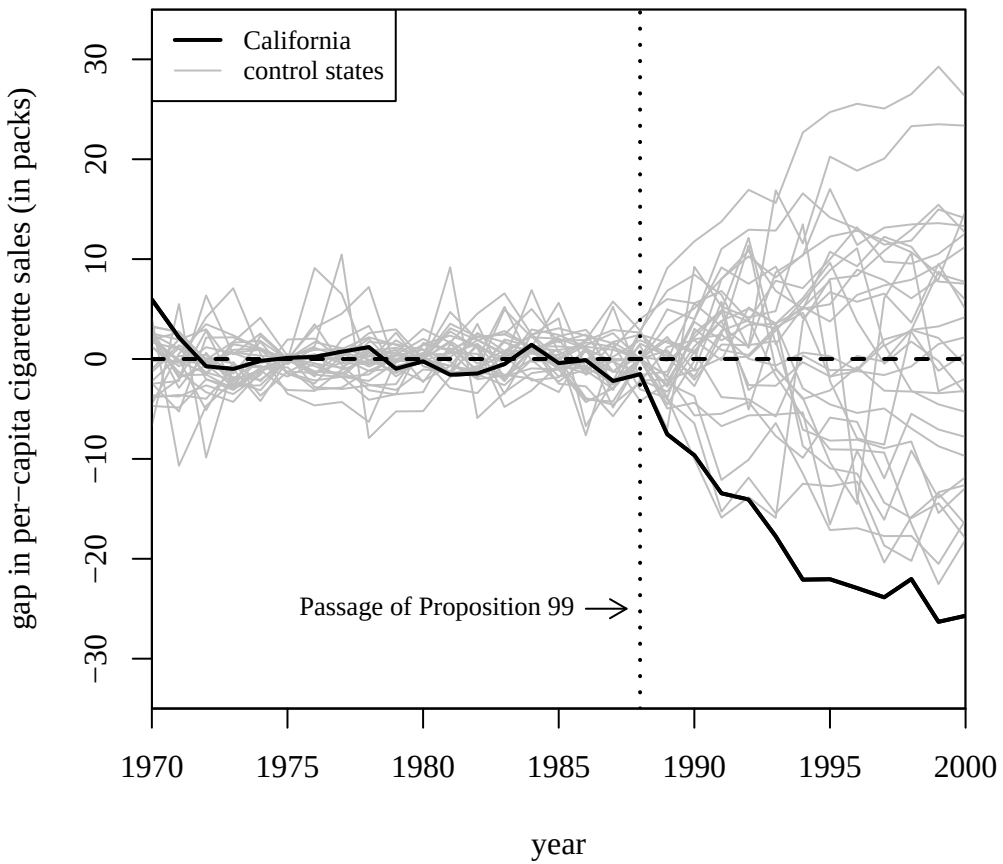
Cigarette sales gap in all 38 states.

Placebos in Space



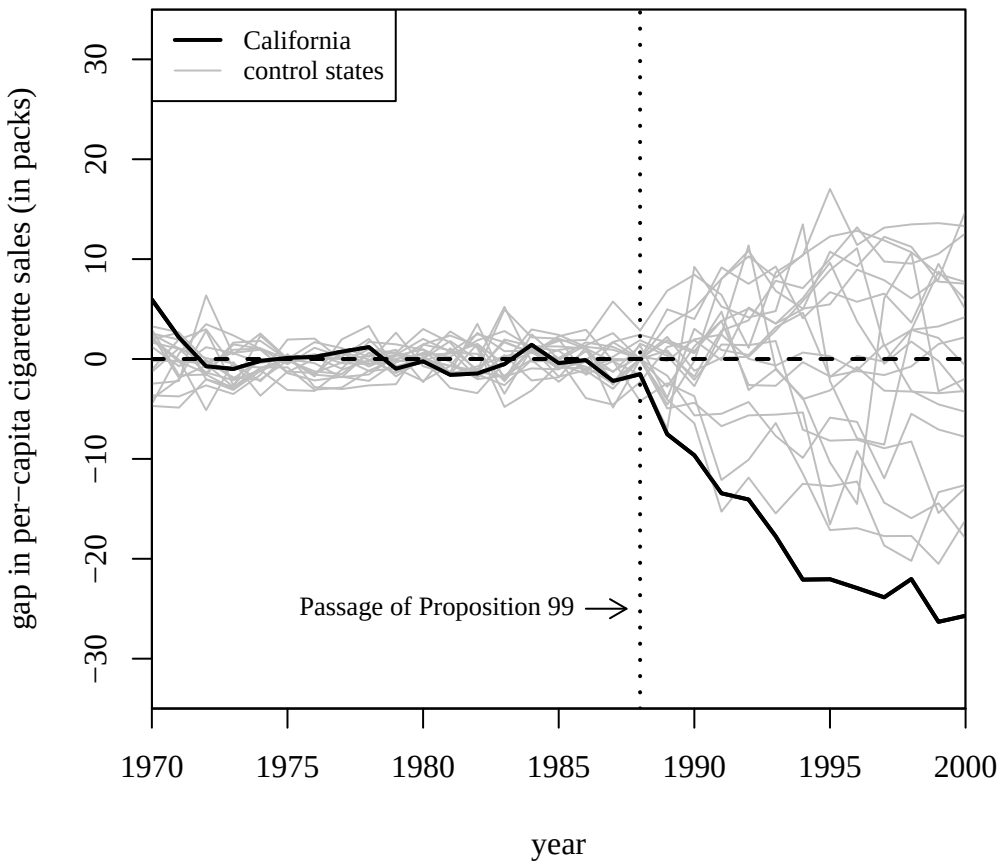
Cigarette sales gap in states with pre-intervention MSPE $< 20 \cdot \text{MSPE}_{\text{CA}}$.

Placebos in Space



Cigarette sales gap in states with pre-intervention MSPE $< 5 \cdot \text{MSPE}_{\text{CA}}$.

Placebos in Space



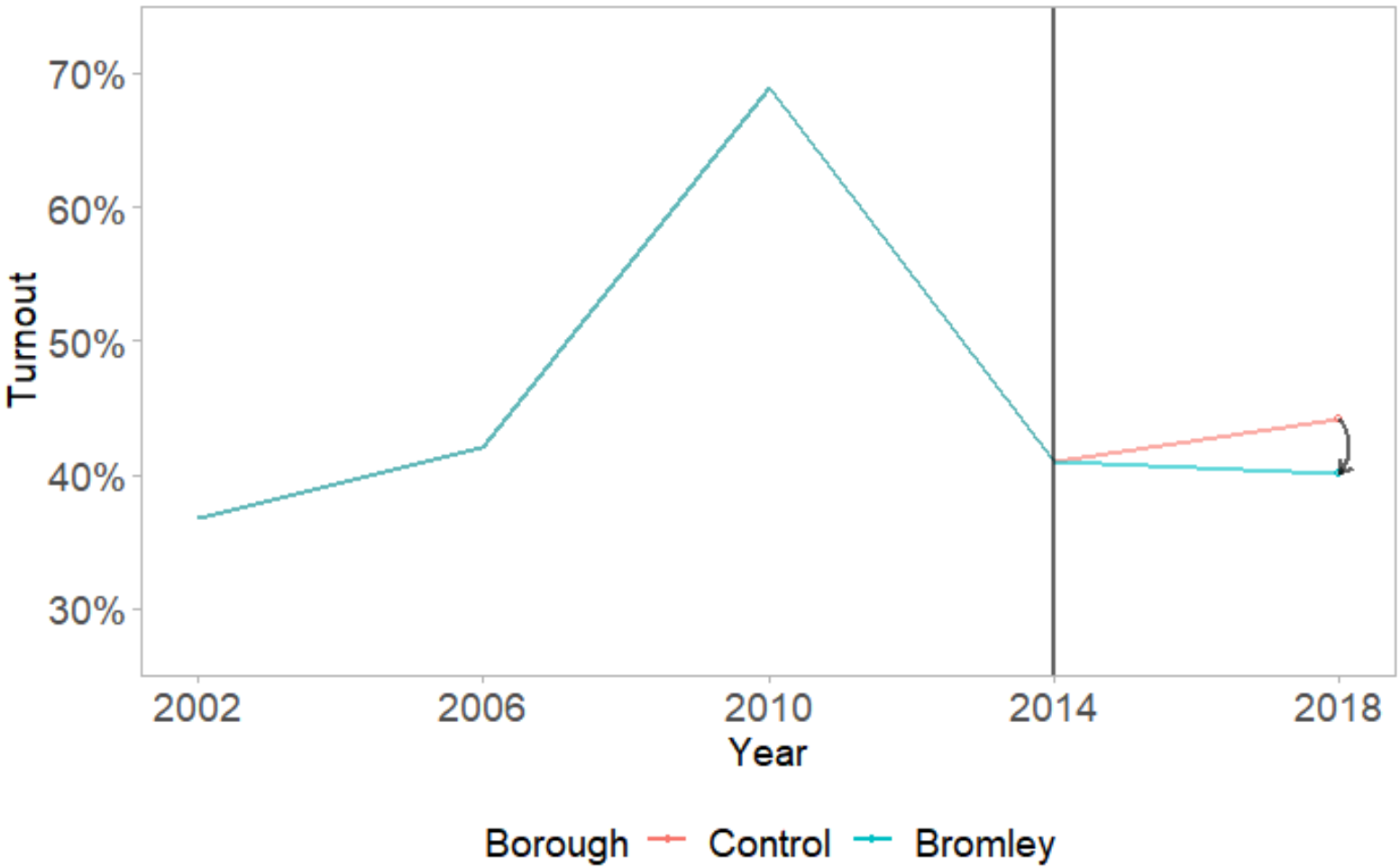
Cigarette sales gap in states with pre-intervention MSPE $< 2 \cdot \text{MSPE}_{\text{CA}}$.

The Voter ID Pilot Scheme in 2018

Do more restrictive voter ID laws reduce turnout?

In 2018, the UK government piloted a more restrictive voter ID law. Usually voters had to only give their name and address, the pilot scheme changed this to a range of different requirements. Although there were multiple treament units they all had slightly different treatments, thus a regular diff-in-diff would be problematic. So lets pick 1 and use synthetic control.

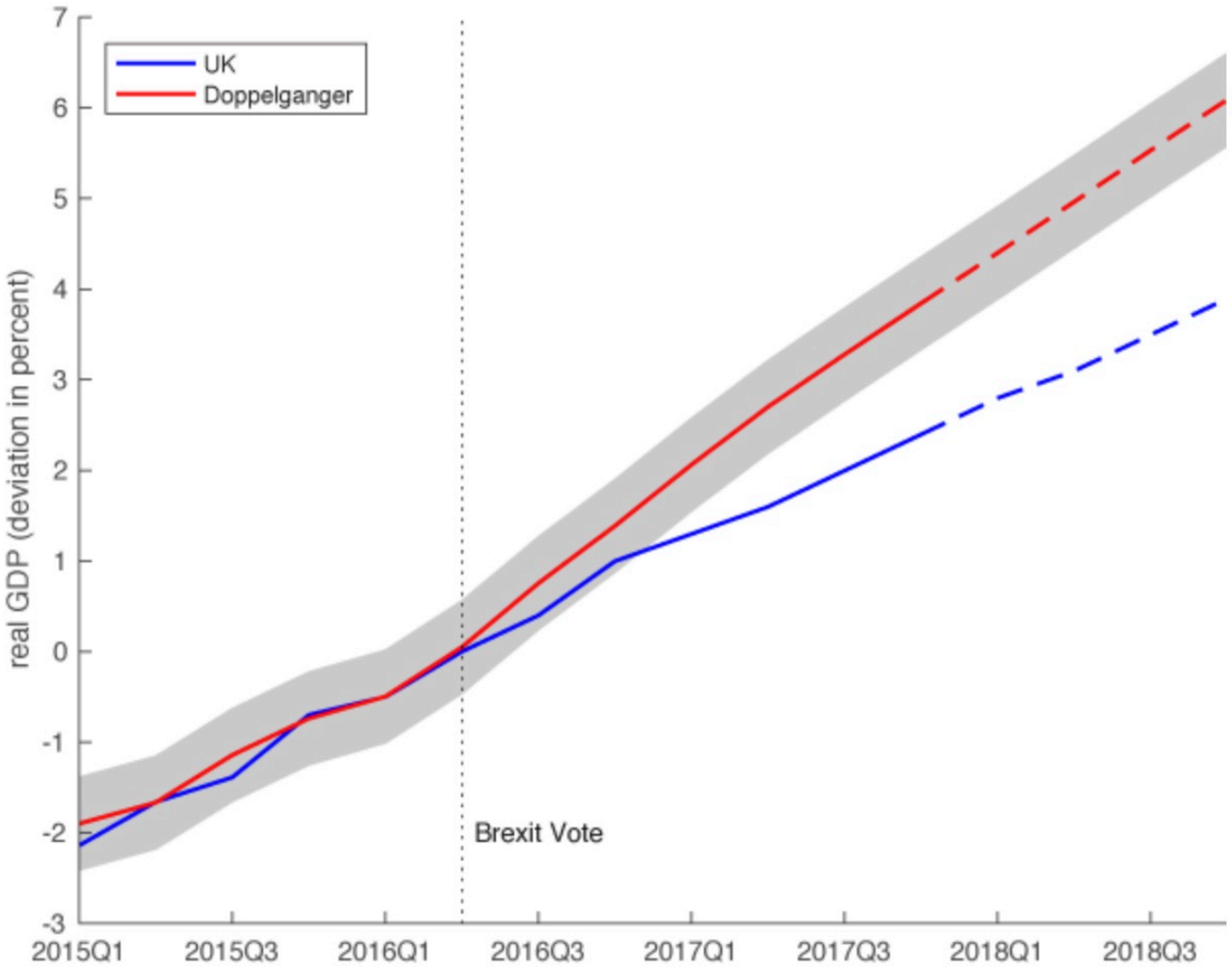
The Voter ID Pilot Scheme in 2018



"the voter ID pilot scheme caused a 4.1% point drop in turnout in Bromley in 2018"

The Economic Cost of the Brexit Vote

Figure 2 UK (blue line) versus doppelganger (red line), zoomed in



Note: Dashed lines are forecasts. Shaded area corresponds to one standard deviation of the pre-treatment difference between UK and doppelganger.

Conclusion

Data requirements

Synthetic control has relatively low data requirements:

- Can use aggregate data (often administrative)
  - e.g. economic indicators such as GDP, current-account balance, etc; political indicators such as turnout, vote share, etc
- Causal factors can be big and important
  - e.g. legislation changes, macro-shocks, etc
- Units of analysis can be large
  - Countries, states, regions, etc
- Does not even require full panel data for the pre-treatment period
  - Can use averages of covariates rather than full panel data (useful when covariates do not vary yearly)

Extensions

- We can also weight by time: **synthetic difference-in-difference**
  - Weakens our reliance on parallel trends (The newest part of the literature).
- Two-way fixed effects
  - Like last week it's useful when we have a few valid controls.
- Clustering
  - As we have seen throughout the course treatment is rarely assigned randomly.

Conclusion

The synthetic control approach...is arguably the most important innovation in the policy evaluation literature in the last 15 years.

–Athey and Imbens, 2017

Advantages

- Builds on D&D and Matching by essentially **forcing** the data to exhibit **equal** trends in the pre-treatment period
- Amenable to small-ish N comparisons (often easier to get data)
- Clear, transparent, and easily communicable comparisons (e.g. Germany is part Austria, part USA, etc)

Disadvantages

- Provides inferences limited to single cases, not "average" treatment effects but ATT
- Often easy to think of "compound" treatments, or multiple changes affecting the treated unit at the same time as the treatment
- Pre-intervention period must be relatively large for us to trust parallel trends holds in the post-intervention period (Synthetic Diff-in-Diff can *somewhat* help us get past this)
- Inference is not straightforward! Asymptotic inference does not work with the SC method
- Coding is highly the dependent on the package you use! They all have their own quirks. In seminar we use *tidysynth*