14.385 Nonlinear Econometric Analysis

Synthetic controls

Maximilian Kasy

Department of Economics, MIT

Fall 2022

Outline

- Synthetic controls and comparative case studies.
- Construction of synthetic controls.
- Factor model and bias bounds.
- Permutation inference.
- Applications.
- Requirements for validity.

Takeaways for this part of class

- Synthetic controls are a convex combination of similar units.
- They match pre-treatment outcomes and covariates.
- Factor models can rationalize synthetic controls.
- Permutation inference provides an analog to randomization inference for experiments.

Introduction

- Synthetic control methods were originally proposed in Abadie and Gardeazabal (2003) and Abadie et al. (2010) with the aim to estimate the effects of aggregate interventions.
- Many events or interventions of interest naturally happen at an aggregate level affecting a small number of large units (such as cities, regions, or countries).
- Even in experimental settings micro-interventions may not be feasible (e.g., fairness) or effective (e.g., interference).

Introduction

- When the units of analysis are a few aggregate entities, a combination of comparison units (a "synthetic control") often does a better job reproducing the characteristics of a treated unit than any single comparison unit alone.
- The comparison unit in the synthetic control method is selected as the weighted average of all potential comparison units that best resembles the characteristics of the treated unit(s).

Setup

- Suppose that we observe $J+1$ units in periods $1, 2, \ldots, T$.
- Unit "one" is exposed to the intervention of interest (that is, "treated") during periods $T_0+1, \ldots, T$.
- The remaining J are an untreated reservoir of potential controls (a "donor pool").
- Let Y_{it}^N be the outcome that would be observed for unit i at time t in the absence of the intervention.
- Let Y_{it}^I be the outcome that would be observed for unit i at time t if unit i is exposed to the intervention in periods T_0+1 to T.
- We aim to estimate the effect of the intervention on the treated unit,

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - Y_{1t}^N$$

for $t > T_0$, and Y_{1t} is the outcome for unit one at time t.

Setup continued

- Let $\boldsymbol{W} = (w_2, \dots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \dots, J+1$ and $w_2 + \dots + w_{J+1} = 1$. Each value of $\boldsymbol{W}$ represents a potential synthetic control.
- Let $\boldsymbol{X}_1$ be a $(k \times 1)$ vector of pre-intervention characteristics for the treated unit. Similarly, let $\boldsymbol{X}_0$ be a $(k \times J)$ matrix which contains the same variables for the unaffected units.
- The vector $\boldsymbol{W}^* = (w_2^*, \dots, w_{J+1}^*)'$ is chosen to minimize $\|\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W}\|$, subject to our weight constraints.
- Let Y_{jt} be the value of the outcome for unit j at time t. For a post-intervention period t (with $t \geq T_0$) the synthetic control estimator is:

$$\widehat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}.$$

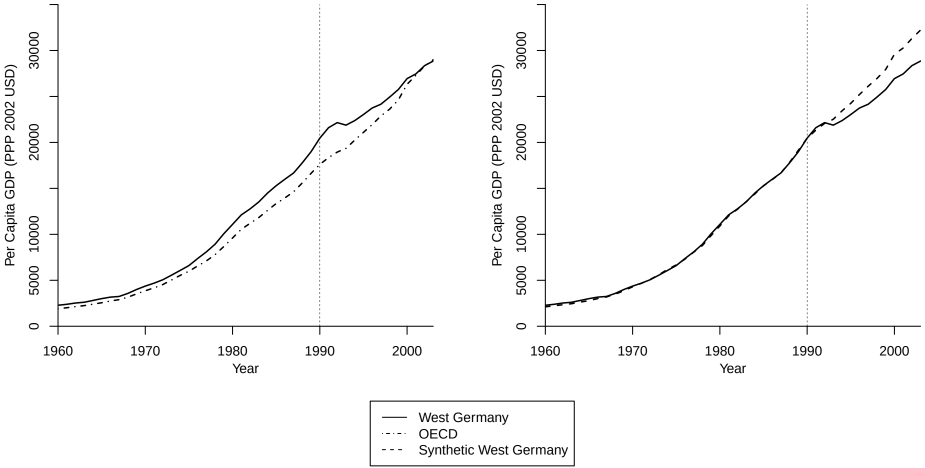
Weighted squared error

- Typically,

$$\|\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W}\| = \left(\sum_{h=1}^{k} v_h \left(X_{h1} - w_2X_{h2} - \cdots - w_{J+1}X_{hJ+1}\right)^2\right)^{1/2}$$

- The positive constants $v_1, \ldots, v_k$ reflect the predictive power of each of the k predictors on Y^N_{1t}.
- $v_1,\ \ldots,\ v_k$ can be chosen via out-of-sample validation.

Application: German reunification



Application: German reunification

	West Germany (1)	Synthetic West Germany (2)	OECD Sample (3)
GDP per-capita	15808.9	15802.24	13669.4
Trade openness	56.8	56.9	59.8
Inflation rate	2.6	3.5	7.6
Industry share	34.5	34.5	34.0
Schooling	55.5	55.2	38.7
Investment rate	27.0	27.0	25.9

Note: First column reports $\boldsymbol{X}_1$, second column reports $\boldsymbol{X}_0\boldsymbol{W}^*$, and last column reports a simple average for the 16 OECD countries in the donor pool. GDP per capita, inflation rate, and trade openness are averages for 1981–1990. Industry share (of value added) is the average for 1981–1989. Schooling is the average for 1980 and 1985. Investment rate is averaged over 1980–1984.

Application: German reunification

country j	W_j^*	country j	W_j^*
Australia	0	Netherlands	0.10
Austria	0.42	New Zealand	0
Belgium	0	Norway	0
Denmark	0	Portugal	0
France	0	Spain	0
Greece	0	Switzerland	0.11
Italy	0	United Kingdom	0
Japan	0.16	United States	0.22

Introduction

Bias bound and inference

Applications

Why use synthetic controls?

Requirements

References

Factor model and bias bound

- Abadie et al. (2010) establish a bias bound under the factor model

$$Y_{it}^N = \delta_t + \theta_t \boldsymbol{Z}_i + \lambda_t \mu_i + \varepsilon_{it},$$

where $\boldsymbol{Z}_i$ are observed features, μ_i are unobserved features, and ε_{it} is a unit-level transitory shock, modeled as random noise.

- Suppose that we can choose $\boldsymbol{W}^*$ such that:

$$\sum_{j=2}^{J+1} w_j^* \boldsymbol{Z}_j = \boldsymbol{Z}_1, \sum_{j=2}^{J+1} w_j^* Y_{j1} = Y_{11}, \cdots , \sum_{j=2}^{J+1} w_j^* Y_{jT_0} = Y_{1T_0}.$$

In practice, these may hold only approximately.

Practice problem

Suppose that the factor model holds,
and that the pre-treatment fit of the synthetic control is exact.
Derive an expression for the estimation error for treatment effects.

Solution

If $\sum_{t=1}^{T_0} \lambda_t' \lambda_t$ is nonsingular, then

$$Y_{1t}^N - \sum_{j=2}^{J+1} Y_{jt} =$$

$$\sum_{j=2}^{J+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left(\sum_{n=1}^{T_0} \lambda_n' \lambda_n \right)^{-1} \lambda_s' (\varepsilon_{js} - \varepsilon_{1s})$$

$$- \sum_{j=2}^{J+1} w_j^* (\varepsilon_{jt} - \varepsilon_{1t}).$$

Fit and validity

- The bias bound is predicated on close fit, and controlled by the ratio between the scale of ε_{it} and T_0.
- In particular, the credibility of a synthetic control depends on the extent to which it is able to fit the trajectory of Y_{1t} for an extended pre-intervention period.

Fit and validity

- There are no ex-ante guarantees on the fit. If the fit is poor, Abadie et al. (2010) recommend against the use of synthetic controls.
- In particular, settings with small T_0, large J, and large noise create substantial risk of overfitting.
- To reduce interpolation biases and risk of overfitting, restrict the donor pool to units that are similar to the treated unit.

Permutation inference

- Abadie et al. (2010) propose a mode of inference for the synthetic control framework that is based on permutation methods.
- A permutation distribution can be obtained by iteratively reassigning the treatment to the units in the donor pool and estimating “placebo effects” in each iteration.
- The effect of the treatment on the unit affected by the intervention is deemed to be significant when its magnitude is extreme relative to the permutation distribution.

Permutation inference

- Permutation inference is complicated by the fact that the pre-intervention fit on the outcome variable may be of different quality for different sample units.

- This can be addressed by using the ratio between post-treatment and pre-treatment RMSE as a test statistic. Let

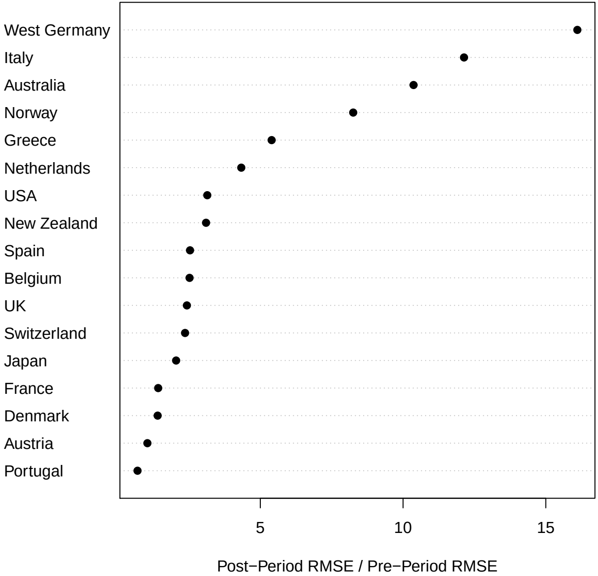
$$R_j(t_1, t_2) = \left(\frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} (Y_{jt} - \widehat{Y}_{jt}^N)^2 \right)^{1/2},$$

  where $\widehat{Y}_{jt}^N$ is the outcome on period t produced by a synthetic control when unit j is coded as treated and using all other J units to construct the donor pool.

- Abadie et al. (2010) use the permutation distribution of

$$r_j = \frac{R_j(T_0 + 1, T)}{R_j(1, T_0)}.$$

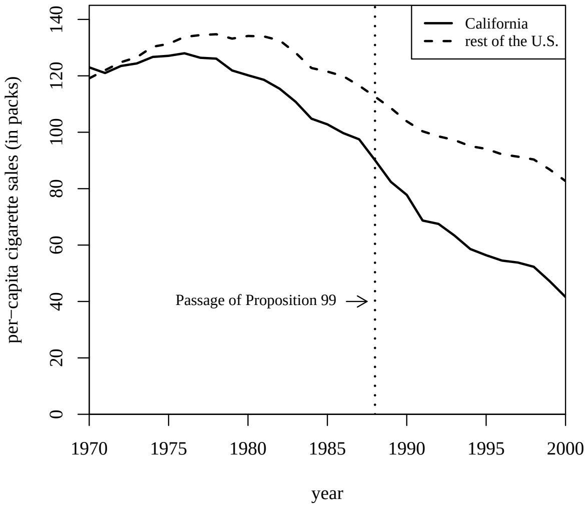
Application: German reunification



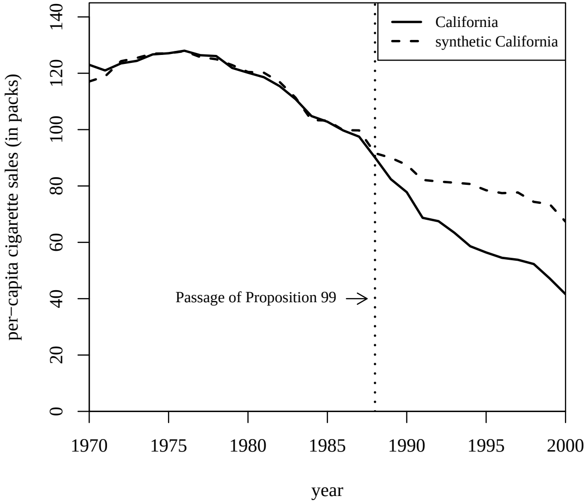
Application: German reunification

- The permutation distribution is more informative than mechanically looking at p-values alone.
- Depending on the number of units in the donor pool, conventional significance levels may be unrealistic or impossible.
- Often, one sided inference is most relevant.

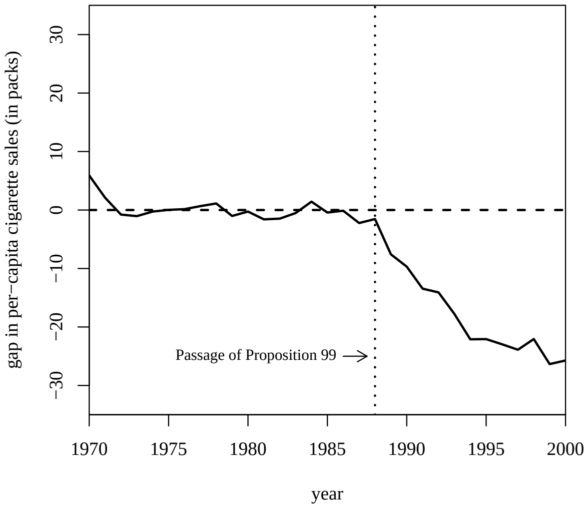
Application: California tobacco control program



Application: California tobacco control program

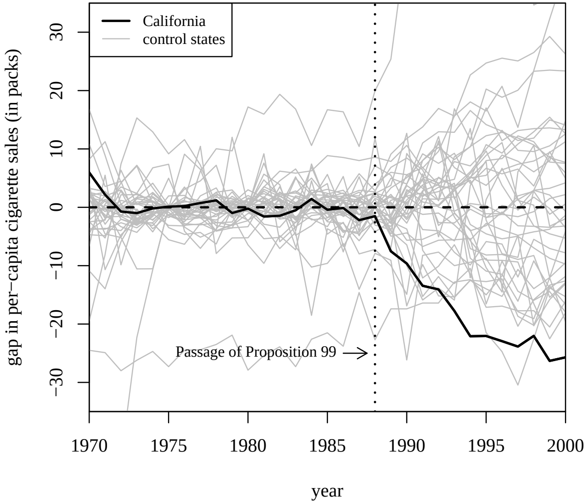


Application: California tobacco control program



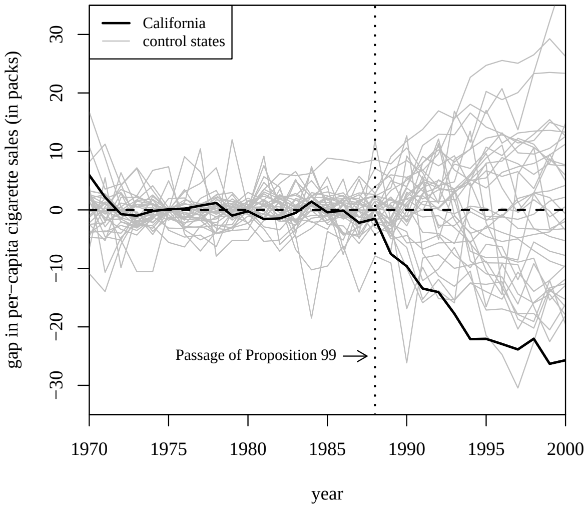
Application: California tobacco control program

(All States in Donor Pool)



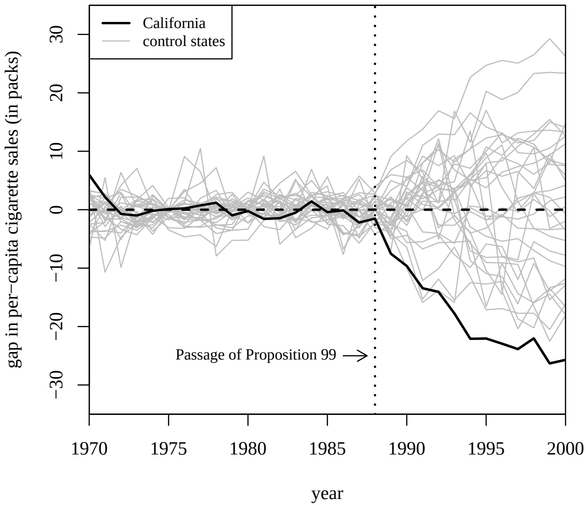
Application: California tobacco control program

(Pre-Prop. 99 MSPE $\leq$ 20 Times Pre-Prop. 99 MSPE for CA)



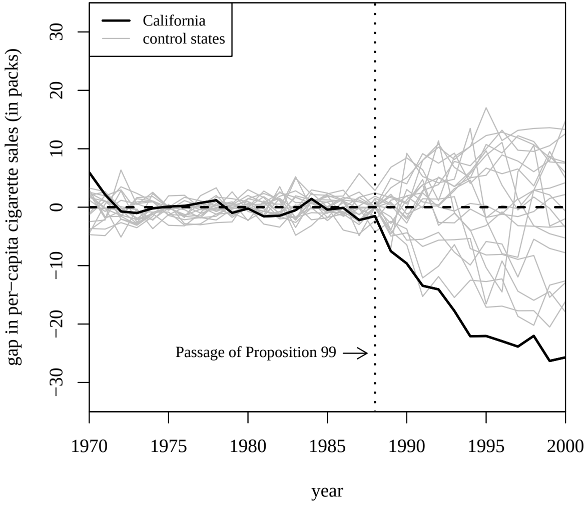
Application: California tobacco control program

(Pre-Prop. 99 MSPE $\leq$ 5 Times Pre-Prop. 99 MSPE for CA)



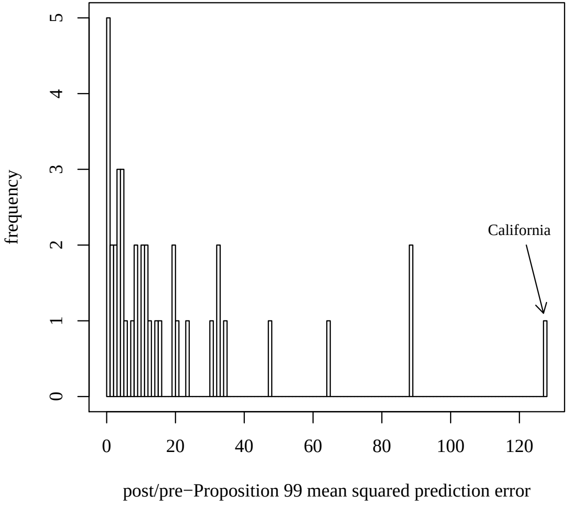
Application: California tobacco control program

(Pre-Prop. 99 MSPE $\leq$ 2 Times Pre-Prop. 99 MSPE for CA)



Application: California tobacco control program

(All 38 States in Donor Pool)



Permutation inference

- The availability of a well-defined procedure to select the comparison unit makes the estimation of the effects of placebo interventions feasible.
- The permutation method we just described does not attempt to approximate the sampling distributions of test statistics.
- Sampling-based inference is complicated in a comparative case study setting, sometimes because of the absence of a well-defined sampling mechanism and sometimes because the sample is the same as the population.

Permutation inference

- This mode of inference reduces to classical randomization inference (Fisher, 1935) when the intervention is randomly assigned, a rather improbable setting.
- More generally, this mode of inference evaluates significance relative to a benchmark distribution for the assignment process, one that is implemented directly in the data.

Why use synthetic controls?

- Compare to linear regression. Let:
  - $\boldsymbol{Y}_0$ be the $(T - T_0) \times J$ matrix of post-intervention outcomes for the units in the donor pool.
  - $\overline{\boldsymbol{X}}_1$ and $\overline{\boldsymbol{X}}_0$ be the result of augmenting $\boldsymbol{X}_1$ and $\boldsymbol{X}_0$ with a row of ones.
  - $\widehat{\boldsymbol{B}} = (\overline{\boldsymbol{X}}_0 \overline{\boldsymbol{X}}_0')^{-1} \overline{\boldsymbol{X}}_0 \boldsymbol{Y}_0'$ collects the coefficients of the regression of $\boldsymbol{Y}_0$ on $\overline{\boldsymbol{X}}_0$.
- $\widehat{\boldsymbol{B}}' \overline{\boldsymbol{X}}_1$ is a regression-based estimator of the counterfactual outcome for the treated unit without the treatment.
- Notice that $\widehat{\boldsymbol{B}}' \overline{\boldsymbol{X}}_1 = \boldsymbol{Y}_0 \boldsymbol{W}^{reg}$, with

$$\boldsymbol{W}^{reg} = \overline{\boldsymbol{X}}_0' (\overline{\boldsymbol{X}}_0 \overline{\boldsymbol{X}}_0')^{-1} \overline{\boldsymbol{X}}_1.$$

- The components of $\boldsymbol{W}^{reg}$ sum to one, but may be outside $[0, 1]$, allowing extrapolation.

Application: German reunification

country j	W_j^{reg}	country j	W_j^{reg}
Australia	0.12	Netherlands	0.14
Austria	0.26	New Zealand	0.12
Belgium	0.00	Norway	0.04
Denmark	0.08	Portugal	-0.08
France	0.04	Spain	-0.01
Greece	-0.09	Switzerland	0.05
Italy	-0.05	United Kingdom	0.06
Japan	0.19	United States	0.13

Why use synthetic controls?

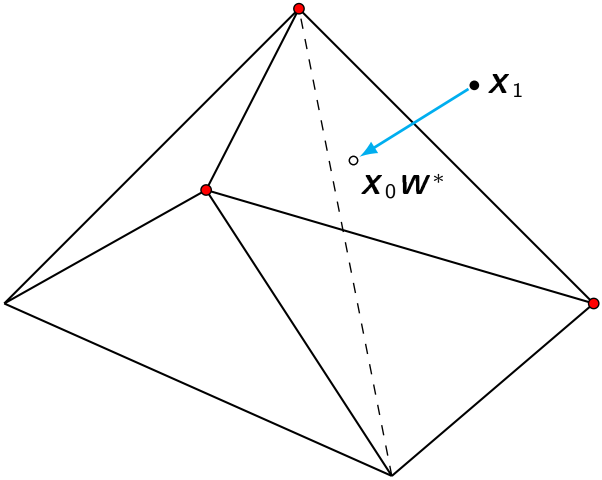
- **No extrapolation.** Synthetic control estimators preclude extrapolation outside the support of the data.
- **Transparency of the fit.** Linear regression uses extrapolation to guarantee a perfect fit of the characteristics of the treated unit, $\boldsymbol{X}_0\boldsymbol{W}^{reg} = \boldsymbol{X}_1$, even when the untreated units are completely dissimilar in their characteristics to the treated unit. In contrast, synthetic controls make transparent the actual discrepancy between the treated unit and the convex hull of the units in the donor pool, $\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W}^*$.
- **Safeguard against specification searches.** Synthetic controls do not require access to post-treatment outcomes in the design phase of the study, when synthetic control weights are calculated. Therefore, all design decisions can be made without knowing how they affect the conclusions of the study.

Why use synthetic controls?

- **Safeguard against specification searches (cont.)** Synthetic control weights can be calculated and pre-registered before the post-treatment outcomes are realized, or before the actual intervention takes place, providing a safeguard against specification searches and p-hacking.
- **Transparency of the counterfactual.** Synthetic controls make explicit the contribution of each comparison unit to the counterfactual of interest.
- **Sparsity.** Because the synthetic control coefficients are proper weights and are sparse, they allow a precise interpretation of the nature of the estimate of the counterfactual of interest (and of potential biases).

Sparsity: Geometric interpretation

Sparsity comes from projecting $\boldsymbol{X}_1$ on the convex hull of $\mathbf{X_0}$



Why use synthetic controls?

- In some cases, especially in applications with many treated units, the values of the predictors for some of the treated units may fall in the convex hull of the columns of $\boldsymbol{X}_0$.
- Then, synthetic controls are not unique or necessarily sparse.
- A modification of the synthetic control estimator that is always unique and sparse is developed in Abadie and L'Hour (2019).

Contextual requirements

- **Size of the effect and volatility of the outcome.** Small effects will be indistinguishable from other shocks to the outcome of the affected unit, especially if the outcome variable of interest is highly volatile.
- **Availability of a comparison group.** Untreated units that
  - Do not adopt interventions similar to the one under investigation during the period of the study.
  - Do not suffer large idiosyncratic shocks to the outcome of interest during the study period.
  - Have characteristics similar to the characteristics of the affected unit.
- **No anticipation.** Can be addressed by backdating.

Contextual requirements

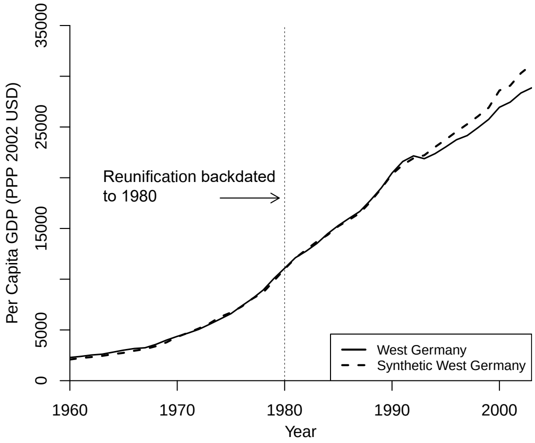
- **No interference.** Sparsity makes it possible to address interference issues.
- **Convex hull condition.** Synthetic control estimates are predicated on the idea that a combination of unaffected units can approximate the pre-intervention characteristics of the affected unit.
- **Time horizon.** The effect of some interventions may take time to emerge or to be of enough magnitude to be quantitatively detected in the data.

Data requirements

- **Aggregate data on predictors and outcomes.** Sometimes, when aggregate data do not exist aggregates of micro-data are employed in comparative case studies.
- **Sufficient pre-intervention information.** The credibility of a synthetic control estimator depends in great part on its ability to steadily track the trajectory of the outcome variable for the affected unit before the intervention. (Recall bias bound.)
- **Sufficient post-intervention information.** This may be problematic if the effect of an intervention is expected to arise gradually over time and if no forward looking measures of the outcome are available.

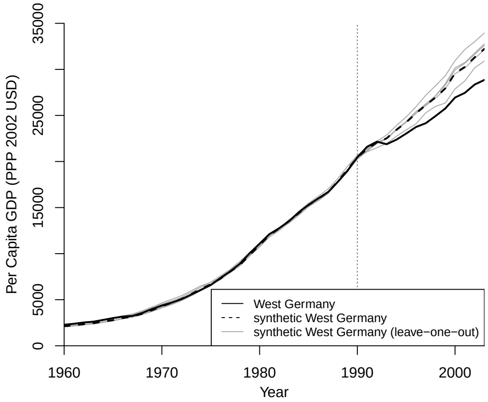
Robustness and Diagnosis Checks

- **Backdating.** Backdating was discussed before as a way to address anticipation effects on the outcome variable before an intervention occurs. In the absence of anticipation effects, he same idea can be applied to assess the credibility of a synthetic control in concrete empirical applications.



Robustness and Diagnosis Checks

- **Robustness tests.** With respect to changes in the study design. In the context of synthetic controls:
  - Units in the donor pool
  - Predictors of the outcome variable.



References

Abadie, A. and 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., and 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*

These slides are based on the slides by **Alberto Abadie** for previous iterations of 14.385.