Differences in Differences

Compass-Lexecon

Raquel Carrasco

Universidad Carlos III de Madrid

May-June 2024

Outline Session 2

- Synthetic Control Methods
  - Basic Setup
  - Bias
  - Weights
  - Inference
  - Comparison to Regression
  - Comparison to Matching
  - Recent Developments: Synthetic DD
- DD for Nonlinear Models
  - Nonlinear (binary) specifications
  - Challenges

Synthetic Control methods for Comparative Case Studies

- Synthetic control has been called the most important innovation in causal inference of the last two decades (Athey and Imbens, 2017).
- It is the most popular method for evaluating comparative case studies:
  - A distinctive feature of comparative case studies is that units of analysis are usually aggregate entities for which suitable single comparisons often do not exist.
- Data feature: (1) only one or a few treated units, and many more control units; (2) long time series both before and after

Synthetic control methods for Comparative Case Studies

- The SCM is based on the observation that, 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.
- Motivated by this consideration, 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.

Synthetic control methods for Comparative Case Studies

- Synthetic control methods were originally proposed in Abadie and Gardeazabal (2003) when estimating the effects of the terrorist conflict in the Basque Country using other Spanish regions as a comparison groups.
    - They do not use a standard DD method because none of the other Spanish regions followed the same time trend as the Basque Country.
- Abadie, Diamond and Hainmuller (2010, JASA) study the effect of California's 1988 tobacco control program.
    - They generate a "synthetic California" that mirrors the values of the predictors of cigarette consumption in California before the passage of the law, build on a combination of Colorado, Connecticut, Montana, Nevada, and Utah.

Synthetic control methods for Comparative Case Studies

- In the standard SC approach the synthetic control unit is created out of a collection of J control units (donor pool).
- The comparability of the synthetic control unit to the treated unit in the preintervention period is determined by a set of predictors of the outcome from several pretreatment periods.
- The SC creates a synthetic control unit by matching predictors from a set of donor units.
  - The predictors can be outcomes from some or all of the observations in the pretreatment period, or they can include non-outcome predictors from the pretreatment period (covariates).

Synthetic control methods: Pros and cons

- Advantages:
  - Policy interventions often take place at an aggregate level, and affect aggregate entities, such as schools, or geographic or administrative areas.
  - Aggregate/macro data are often available
  - Precludes extrapolation
  - Does not require access to post-treatment outcomes in the "design" phase of the study, when synthetic controls are calculated
  - Makes explicit the contribution of each comparison unit to the counterfactual of interest
- Disadvantages:
  - Selection of control group is ambiguous
  - Standard errors do not reflect uncertainty about the ability of the control group to reproduce the counterfactual of interest
  - Bias due to inexact matching...

Synthetic control methods: Setup

- Consider a simple example, with only two time periods: one before the policy and one after.
- Let Y_{it} be the outcome for unit i in time t, with $i = 1$ the treated unit.
- Suppose there are J possible controls $\{2, ..., J+1\}$ (a "donor pool")
- We aim to estimate the effect of the intervention on the treated unit α_{12} where

$$\alpha_{12} = Y_{12} - Y_{12}^0$$

- We can define the control group in the second period as

$$\sum_{j=2}^{J+1} w_j Y_{j2},$$

  where w_j are **nonnegative weights** that **add up to one** chosen such that a vector of pre-treatment covariates and outcomes for the treated unit are reproduced.
- The SC estimator is

$$\widehat{\alpha}_{12} = Y_{12} - \sum_{j=2}^{J+1} w_j Y_{j2}$$

Synthetic control methods: Weights

- How can we choose the weights?
- The SC approach can be described as a two-step optimization procedure with an "outer" and "inner" level.
  - Inner level: choose the weights for the donor units to minimize the Mean Square Predicted error (MSPE) between pretreatment outcomes of the synthetic control and the treated unit.
  - Outer level: choose the weights for the predictors to minimize the MSPE between the predictors of the synthetic control and the predictors of the treated unit.
- In most studies, predictors include at least some of the outcomes in the pretreatment period.

Synthetic control methods: Weights

- Two kinds of predictors are introduced.
  - The first is given by M linear combinations of Y in the pretreatment periods
  - The second consists of other covariates with explanatory power for Y.
  - All k predictors (with $k = M + r$) are combined in a $(k \times 1)$ vector X_1 for the treated unit and in a $(k \times J)$ matrix X_0 for all control units.
- Optimization procedure:

Synthetic control methods: Weights

- Inner level: find a linear combination of the columns of X_0 that represents X_1 best $\Longrightarrow$ the difference of the predictors' values of the treated and the counterfactual becomes as small as possible.
- The distance metric used to measure this difference is

$$(\mathbf{X}_1 - \mathbf{X}_0\mathbf{W})' \, V \, (\mathbf{X}_1 - \mathbf{X}_0\mathbf{W})$$

  where the vector W are the weights used to construct the synthetic control unit, and the weights of the predictors are given by the nonnegative diagonal matrix V.
- The vector $W^* = (w_2^*, ...w_J^*)$ is chosen to minimize previous expression, subject to weight constraints

Synthetic control methods: Weights

- V takes into consideration that not all predictors have the same predictive power for the outcome variable Y.
- The choice of V is important because the optimal weights W^* depend on it.
- V is diagonal with main diagonal $v_1, ...v_k$. Then, the synthetic control weights $w_2, ...w_J$ minimize:

$$\sum_{m=1}^{k} v_m \left(X_{1m} - \sum_{j=2}^{J+1} w_j X_{jm} \right)^2$$

$$s.t. \sum_{j=2}^{J+1} w_j = 1, w_j \geq 0$$

where v_m is a weight that reflects the relative importance that we assign to the $m-th$ variable when we measure the discrepancy between the treated unit and the synthetic controls.

- Denote the solution to this problem by $W^*(V)$.

Synthetic control methods: Weights

- Example: 2 controls, 2 covariates:

$$
\begin{aligned}
\mathbf{X}_1 &= \begin{pmatrix} X_{11} \\ X_{12} \end{pmatrix} \\
\mathbf{X}_0\mathbf{W} &= \begin{pmatrix} X_{21} & X_{31} \\ X_{22} & X_{31} \end{pmatrix} \begin{pmatrix} w_2 \\ w_3 \end{pmatrix} \\
&= \begin{pmatrix} \sum_{j=2}^{3} w_j X_{j1} \\ \sum_{j=2}^{3} w_j X_{j2} \end{pmatrix}
\end{aligned}
$$

Synthetic control methods: Weights

- Example: 2 controls, 2 covariates:

$$(\mathbf{X}_1 - \mathbf{X}_0\mathbf{W})' \, V \, (\mathbf{X}_1 - \mathbf{X}_0\mathbf{W}) =$$

$$\begin{pmatrix} X_{11} - \sum_{j=2}^{3} w_j X_{j1} & X_{12} - \sum_{j=2}^{3} w_j X_{j2} \end{pmatrix} \begin{pmatrix} v_1 & 0 \\ 0 & v_2 \end{pmatrix} \begin{pmatrix} X_{11} - \sum_{j=2}^{3} w_j X_{j1} \\ X_{12} - \sum_{j=2}^{3} w_j X_{j2} \end{pmatrix}$$

$$= \sum_{m=1}^{k} v_m \left(X_{1m} - \sum_{j=2}^{J+1} w_j X_{jm} \right)^2$$

Synthetic control methods: Weights

- Outer level: find optimal predictor weights.
- Choice of V can be subjective (just uniform weights, or the inverse of the variance of $\mathbf{X}$) or could be based on a pre-treatment regression of Y on X.
- Abadie et al (2010) follow a data-driven approach: V is chosen to minimize the MSPE of the outcome Y over the preintervention periods:

$$\min_{V} \sum_{t \epsilon pre} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^*(V) Y_{jt} \right)^2 ,$$

- An iterative process – end goal is to choose $W^*(V^*)$ and V^* (nested optimization).

Synthetic control methods: Weights

- Cross-validation
  - Divide the pre-treatment period into an initial training period and a subsequent validation period.
  - For any given V, calculate $W(V)$ in the training period.
  - Minimize the MSPE of $W(V)$ in the validation period.

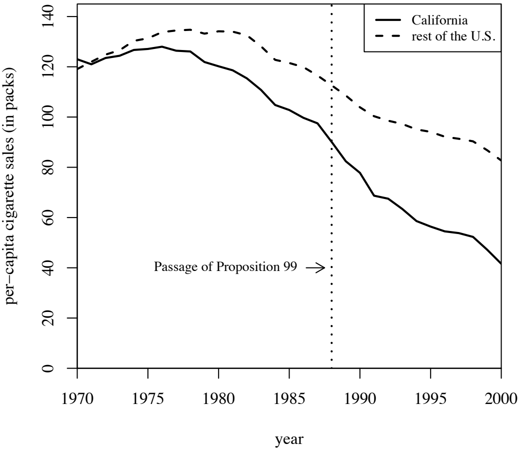
Synthetic control methods: Weights

- In Abadie and Gardeazabal (2003):
  - The optimal weights W^*: Catalonia: 0.851; Madrid: 0.149, and 0 for all other regions.
  - Their results suggest a 10 pp. loss in per capita GDP due to terrorism for the 1980's and 1990's.
- In Abadie, Diamond and Hainmueller (2010):
  - California's Proposition 99: In 1988, California first passed comprehensive tobacco control legislation
    - Increased cigarette tax by 25 cents/pack
    - The optimal weights are: Utah 0.334; Nevada 0.234; Montana 0.199; Colorado 0.164.

Synthetic control methods: ADH (2010)

- Figure 1 plots the trends in per capita cigarette consumption in California and the rest of the United State:
    - The rest of the United States may not provide a suitable comparison group for California:
        - Before treatment the time series of cigarette consumption in California and in the rest of the United States differed notably
- Table 1 compares the pretreatment characteristics of California with that of the synthetic California, as well as with the population-weighted average of the 38 states in the donor pool.
    - The average of states does not seem to provide a suitable control group for California.
    - In contrast, the synthetic California accurately reproduces the values that smoking prevalence and smoking prevalence predictor variables had in California prior to the passage of Proposition 99.

Cigarette Consumption: CA and the Rest of the US



Predictor Means: Actual vs. Synthetic California

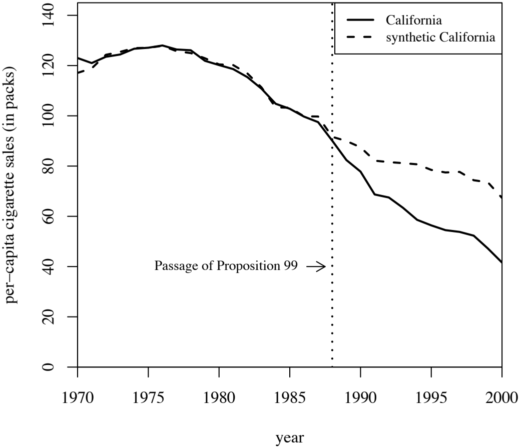
	California		Average of
Variables	Real	Synthetic	38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15-24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20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

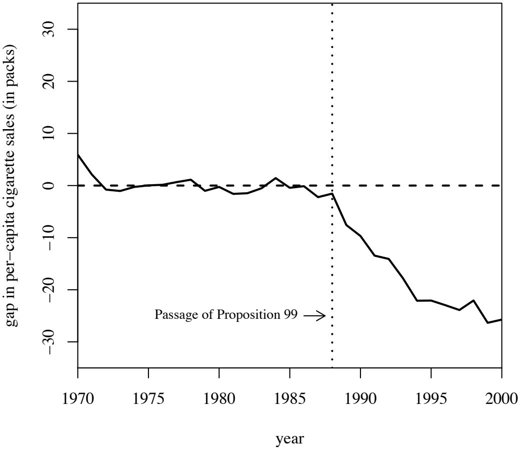
Synthetic control methods: ADH (2010)

- Figure 2 displays per capita cigarette sales for California and its synthetic counterpart during the period 1970–2000.
  - Per capita sales in the synthetic California very closely track the trajectory of this variable in California for the entire pre-Proposition 99 period.
- Figure 3 plots the yearly estimates of the impacts of Proposition 99, that is, the yearly gaps in per capita cigarette consumption between California and its synthetic counterpart:
  - Proposition 99 had a large effect on per capita cigarette sales, and that this effect increased in time.
  - For the entire 1989–2000 period cigarette consumption was reduced by an average of almost 20 packs per capita, a decline of approximately 25%.

Cigarette Consumption: CA and synthetic CA



Smoking Gap between CA and synthetic CA



Synthetic control methods: Predictors

- One question when applying SC methods is which variables should be used when seeking to match across matrices $\mathbf{X}_1$ and $\mathbf{X}_0$.
    - Should we use all pre-intervention outcomes as predictors?
- Ferman et al. (2020) show that different choices in how synthetic controls are generated can lead to cases where a given intervention is found to be significant, or insignificant.
- They suggest that SC should use all pretreatment periods when generating a match.

Synthetic control methods: Predictors

- Kaul et al. (2022) show that using all pretreatment values of the outcome variable as separate predictors leads to every single covariate being ignored:
    - The synthetic control resulting after optimizing the outer objective function will be calculated by ignoring the covariates.
- When ignoring relevant covariates, synthetic controls are not applied as they are intended to be:
    - Gardeazabal and Vega-Bayo (2017): "The synthetic control is primarily designed to use any covariates that help explain the outcome variable as predictors, and not only pretreatment values of the outcome variable".
- Abadie et al (2010) suggest to use data-driven methods for variable selection measuring the predictive power of alternative sets of variables:
    - Divides the pre-intervention periods into an initial training period and a subsequent validation period.
    - Synthetic control weights are computed using data from the training period only.
    - The validation period can then be used to evaluate the predictive power of the resulting synthetic control.

SC: Bias for omitting covariates

- But what are the potential consequences of ignoring the covariates?
- We denote by Y_{jt} the dependent variable's value at time t for unit j. While $t \leq T_0$ indicates pre-treatment values, $t > T_0$ refers to post-treatment values.
- Usually we assume Y_{jt} to depend additively on the covariates by $\theta_t C_j$:

$$Y_{jt} = \widetilde{Y}_{jt} + \theta_t C_j$$

and $\widetilde{Y}_{jt}$ denotes how Y_{jt} would evolve if it was not influenced by the covariates.

SC: Bias for omitting covariates

- Given weights W, the difference between treated unit and its synthetic control with respect to the outcome then is

$$Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt} = \widetilde{Y}_{1t} - \sum_{j=2}^{J+1} w_j \widetilde{Y}_{jt} + \theta_t \left(C_1 - \sum_{j=2}^{J+1} w_j C_j \right)$$

- If the covariates obtain positive weights, then $C_1 - \sum_{j=2}^{J+1} w_j C_j$ will be close to or ideally even exactly zero.

SC: Bias for omitting covariates

- However, if the estimator W ignores the covariates, then $C_1 - \sum_{j=2}^{J+1} w_j C_j$ may be arbitrarily large.
- Therefore, ignoring the covariates introduces an additional uncontrolled small-sample bias to the estimation which is likely to be significant, especially when θ_t takes large values.

SC: Bias for omitting covariates

- Abadie et al. (2010) show that the bias introduced by omitting covariates goes to zero as the number of pre-treatment time periods increases.
- Intuition: if you can closely align the treated and synthetic outcome trajectories for a long enough time before treatment you must also have aligned on the factors (maybe unobserved?) that are relevant for producing that outcome.
  - Therefore, to what extent ignoring the covariates is harmful depends on the length of the pre-treatment timespan as well as the observed covariates' importance for explaining the outcome.

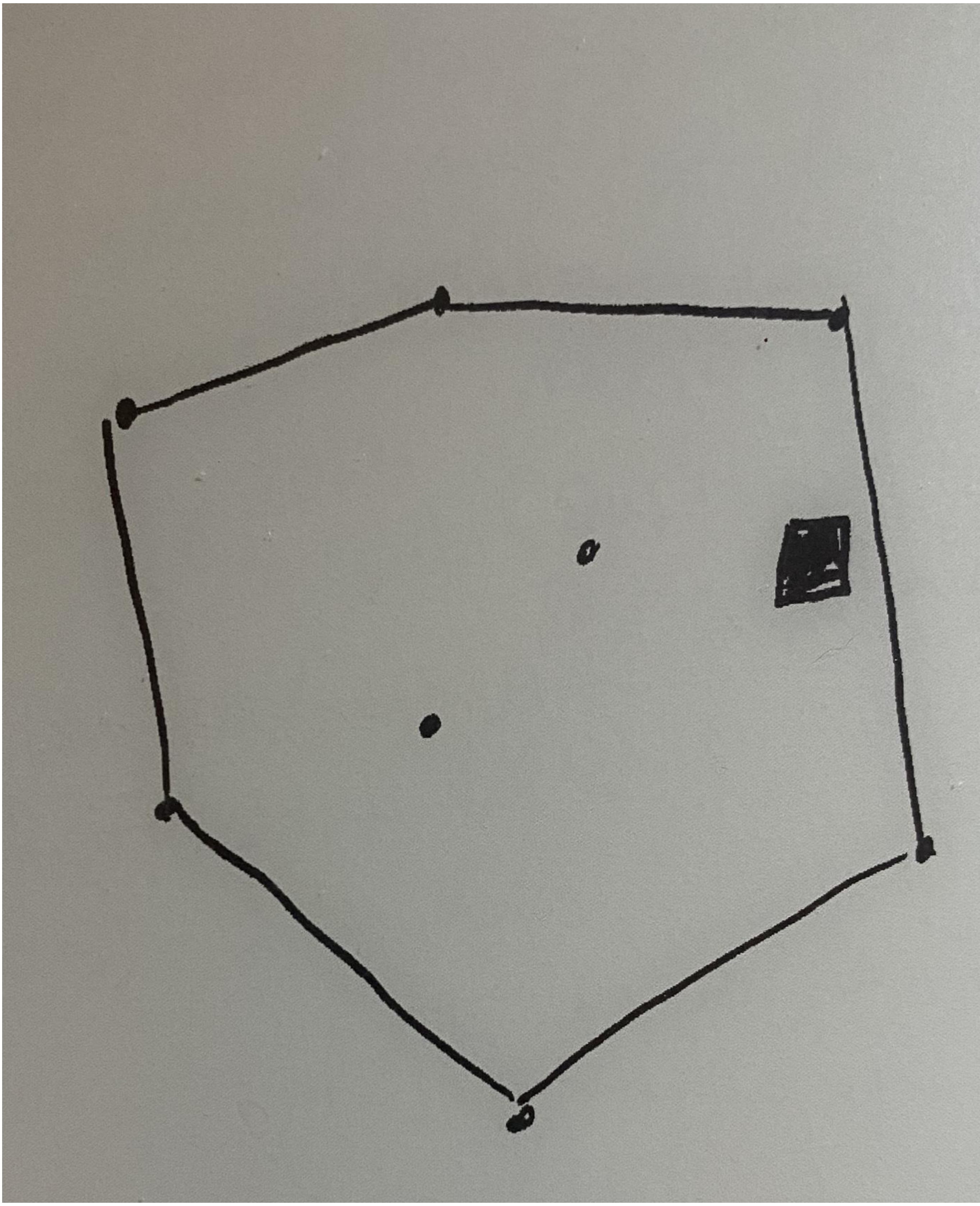
SC: Bias for omitting covariates

- Abadie et al (2015): “The applicability of the SC method requires a sizable number of pre-intervention periods..... **We do not recommend using this method when the pretreatment fit is poor or the number of pretreatment periods is small.** A sizable number of post-intervention periods may also be required in cases when the effect of the intervention emerges gradually after the intervention or changes over time.”
- With a short time series, it would be the case that you predict well with pre-intervention outcomes but just because of the effect of transitory shocks¡

SC: Negative weights?

- With non-negative weights that sum up to 1:
    - We have an interpretable counterfactual as a weighted average of untreated units.
        - Negative wights does not make much sense from a causal inference perspective.
    - Precludes extrapolation: it uses as counterfactual a linear convex combination of the control group units.
        - Similar to common support concept
        - The counterfactual is based on where the data actual is, as opposed to extrapolating beyond the support of the data.
        - Protection from extrapolation is a desirable property, but the convex hull restriction is not a requirement for identification.

Ideal for SC



SC: Negative weights?

- Abadie and L'Hour (2020) and Ben-Michael et al. (2020) propose modifications of the SC estimator in settings where the synthetic control counterfactual is constructed using untreated units with values of the predictors that do not closely reproduce the predictor values for the treated.
    - Abadie and L'Hour (2020) propose adding to the objective function a set of penalty terms that depend on the discrepancies between the characteristics of the affected unit and the characteristics of the individual units included in the synthetic control.
    - Ben-Michael et al (2021, JASA) present a modification to SC in which they allow for negative weights:
- Augmented Synthetic Control: Modification to the original synthetic control model with the inclusion of the penalty term
    - When SC is imbalanced, augmented SC will reduce bias, and when SC is balanced, they are the same
    - The trade-off between the two objectives is controlled by a parameter λ, where a value of zero leads to the original SC
- Stata package `allsynth`

SC: Interpolation bias

- This can arise for instance if a non-linear relationship between the predictors and the outcome exists:
  - ADH (2010) show that when the outcome and the predictors have a non-linear relationship there is a bias that does not decrease as the number of pre-treatment periods increase.
  - Solution_1: restrict the donor pool to regions with similar characteristics to the region exposed to the treatment.
    - Display the results of progressively limiting the pool of potential donors to those with, for example, predictors less than 3 standard deviations from the treated unit, .5 st. dev., and so on.
  - Solution_2: Abadie and L'Hour (2020) develop a solution that does not reduced the pool of donors.
    - Instead it penalizes the use of control units whose predictor values are far from those of the treated unit by adding another term to the minimization problem that penalizes the discrepancies between the characteristics of the treated and the control units.

SC: Shrinking a large pool of donor units

- A large number of donors can cause problems:
    - If pre-treatment outcomes are used as predictors, a large pool of donor units increases the chance that a donor unit is selected because idiosyncratic shocks make the outcome of the donor appear to resemble the treated unit.
        - This form of overfitting can lead to biased results because donor units that are selected may not have outcomes that match the outcomes of the treated unit
    - Solution: increasing the length of time over which the predictor is averaged reduces the impact of a shock in a given year.
    - Limiting the pool of donors to units with predictors most like to those of the treated unit.

SC: Unobserved factors

- What about unobserved factors?
- Comparative case studies are complicated by unmeasured factors affecting the outcome variables
- However, if the number of pre-intervention periods in the data is large, matching on pre-intervention outcomes allows us to control for unobserved factors
- Intuition:
    - Only units that are alike in observed and unobserved determinants of the outcome variable should produce similar trajectories of the outcome variable over extended periods of time.

To sum up:

- 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 pre-treatment observations and large donor pool create substantial risk of interpolation bias and overfitting.
- To reduce them, restrict the donor pool to units that are similar to the treated unit or use Augmented SC methods.

Synthetic control methods: Inference

- The standard errors commonly reported in regression-based comparative case studies measure uncertainty about aggregate data.
  - For example, Card and Krueger (1994) use data on a sample of fast food restaurants in New Jersey and Pennsylvania to estimate the average number of employees in fast food restaurants in these two states around the time when the minimum wage was increased in New Jersey.
  - The standard errors reported reflect only the unavailability of aggregate data on employment (in fast food restaurants in New Jersey and Pennsylvania respectively):

$$Y_{ist} = \gamma_s + \lambda_t + \beta D_{st} + e_{ist}$$

  - This model of inference would logically produce zero standard errors if aggregate data were used for estimation.

Synthetic control methods: Inference

- However, perfect knowledge of aggregate data does not eliminate all uncertainty about the parameters of interest.
    - That is, even if aggregate data are used for estimation, researchers would not believe that there is no remaining uncertainty about the value of the parameters of interest.
- The reason is that not all uncertainty about the value of the estimated parameters comes from lack of knowledge of aggregate data.
    - An additional source of uncertainty derives from ignorance about the ability of the control group to reproduce the counterfactual of how the treated unit would have evolved in the absence of the treatment.
- Abadie, Diamond and Hainmuller (2010) propose permutation methods to perform inference in comparative case studies.

Synthetic control methods: Inference

- To evaluate the significance of the estimates, they pose the question of whether the results could be driven entirely by chance:
  - How often would we obtain results of this magnitude if we had chosen a state at random instead of the treated state?
- They run placebo studies by applying the synthetic control method to states that did not implement the policy during the sample period.
  - This allows to assess whether the effect estimated by the synthetic control for the region affected by the intervention is large relative to the effect estimated for a region chosen at random.

Synthetic control methods: Inference

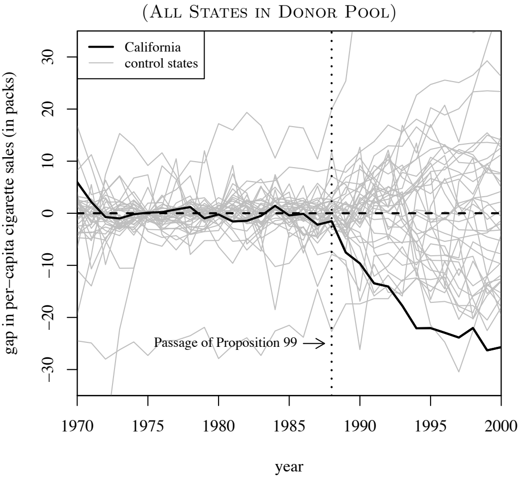
- A permutation distribution can be obtained by iteratively reassigning the treatment to the units in the donor pool and estimating "placebo effects" in each iteration.
- The effect of the treatment on the unit affected by the intervention is deemed to be significant when its magnitude is extreme relative to the permutation distribution.
    - If the placebo studies create gaps of magnitude similar to the one estimated for the treated state, the interpretation is that our analysis does not provide significant evidence of a negative effect of the policy on the treated state.
    - Compare the gap (RMSPE) for the treated to the distribution of the placebo gaps.
    - For example the post-Prop. 99 Root Mean Square Predicted Error is:

$$RMSPE = \left(\frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{\frac{1}{2}}$$

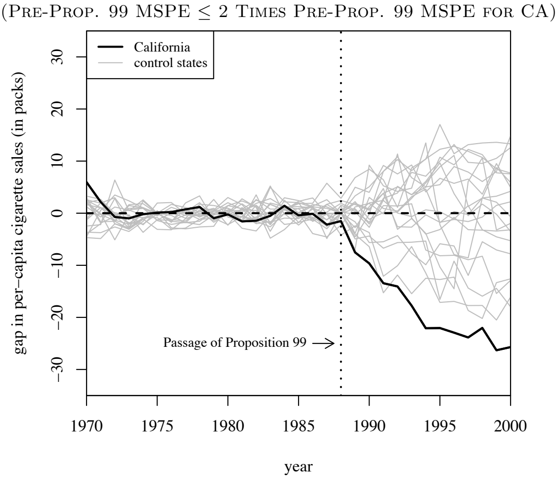
Synthetic control methods: Inference

- Next figure displays the results for the placebo test.
  - The gray lines represent the gap associated with each of the 38 runs of the test. That is, the gray lines show the difference in per capita cigarette sales between each state in the donor pool and its respective synthetic version.
  - The superimposed black line denotes the gap estimated for California.
    - The estimated gap for California during the 1989–2000 period is unusually large relative to the distribution of the gaps for the states in the donor pool.
    - If we focus exclusively on those states that we can fit almost as well as California in the period 1970–1988, still the gap for California appears highly unusual. The negative effect in California is now by far the lowest of all.

Smoking Gap for CA and 38 control states



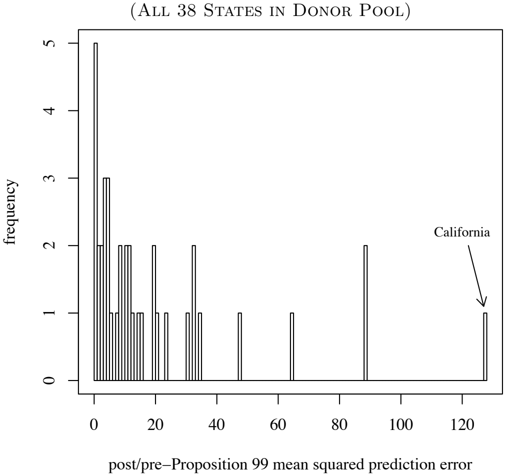
Smoking Gap for CA and 19 control states



Synthetic control methods: Inference

- Another way to evaluate the treated state gap relative to the gaps obtained from the placebo runs is to look at the distribution of the ratios of post/pre-intervention of the MSPE.
- Next figure displays the distribution of the post/pre-Proposition 99 ratios of the MSPE for California and all 38 control states.
- The ratio for California clearly stands out in the figure. No control state achieves such a large ratio. If one were to assign the intervention at random in the data, the probability of obtaining a post/pre-Proposition 99 MSPE ratio as large as California's is extremely small, 0.026. $1/39 = 0.026$.

Ratio Post-Prop. 99 RMSPE to Pre-Prop. 99 RMSPE



Comparison to Regression

- Constructing a synthetic comparison as a linear combination of the untreated units with coefficients that sum to one may appear unusual.
- However in fact a regression-based approach does exactly this, albeit in an implicit way.
- In contrast to the synthetic control method, the regression approach does not restrict the coefficients of the linear combination that define the comparison unit to be in between zero and one, therefore allowing extrapolation outside the support of the data.

Comparison to Regression

- Let:
  - Y_0 be the $(T - T_0) \times J$ matrix of post-intervention outcomes for the units in the donor pool.
  - X_1 : $(k \times 1)$-matrix of covariates for treated unit
  - X_0 : $(k \times J)$-matrix of covariates for control units
- Let

$$\widehat{B} = \left(X_0 X_0'\right)^{-1} X_0 Y_0'$$

  be the matrix of regression coefficients of Y_0 on X_0
  - That is, each column of $\widehat{B}$ contains the regression coefficients of Y_0 on X_0 for a post-intervention period.
- A regression-based counterfactual of the outcome for the treated unit in absence of the treatment is given by $\widehat{B}' X_1$

Comparison to Regression

- Notice that
$$\widehat{B}' X_1 = Y_0 W^{reg},$$
where
$$W^{reg} = X_0' \left(X_0 X_0'\right)^{-1} X_1$$
- Therefore, the regression-based estimate of the counterfactual of interest is a linear combination of post-treatment outcomes for the untreated units, with weights W^{reg}.
- It can be proven that the components of W^{reg} sum to one, but may be outside $(0, 1)$, allowing extrapolation.

Comparison to Regression

- Example: Abadie, Diamond, and Hainmueller (2015), which estimates the effect of the 1990 German reunification on per capita GDP in West Germany. The donor pool consists a set of industrialized countries

Synthetic vs. Regression Weights

Country	Synthetic Control Weight	Regression Weight	Country	Synthetic Control Weight	Regression Weight
Australia	0	0.12	Netherlands	0.10	0.14
Austria	0.42	0.26	New Zealand	0	0.12
Belgium	0	0	Norway	0	0.04
Denmark	0	0.08	Portugal	0	-0.08
France	0	0.04	Spain	0	-0.01
Greece	0	-0.09	Switzerland	0.11	0.05
Italy	0	-0.05	UK	0	0.06
Japan	0.16	0.19	USA	0.22	0.13

Comparison to Regression: Why use synthetic controls?

- **No extrapolation:** SC 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, $X_1 = X_0 W^{reg}$, 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, $X_1 - X_0 W^*$.
- **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.
- **Transparency of the counterfactual**. Synthetic controls make explicit the contribution of each comparison unit to the counterfactual of interest.

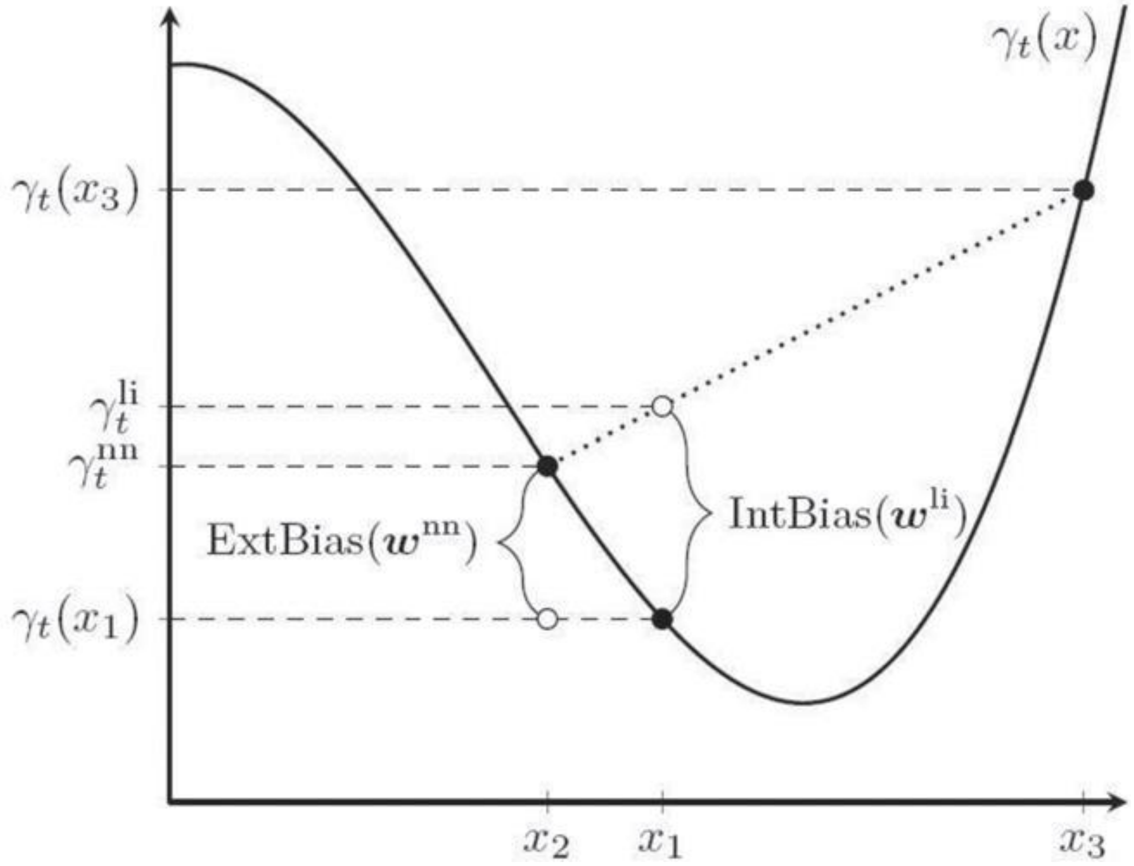
Comparison to matching

- In some ways, synthetic control can be seen as a specific form of matching:
  - Predict unobserved potential outcome using observed outcome of “similar” units
  - Can choose “matches” (i.e., weights) to match untreated outcomes
- Synthetic control differs in how weights are chosen:
  - Matching estimators limit interpolation bias but are sensitive to extrapolation bias
  - SC estimators limit extrapolation bias but are sensitive to interpolation bias.
  - The SC estimator interpolates by using a convex weighted average of untreated units to create a synthetic control with pre-treatment characteristics similar to that of the treated.
    - Interpolation bias arises if the conditional mean of the outcome is non-linear in pre-treatment characteristics.
  - Extrapolation bias arises when the counterfactual constructed for the treated unit is not identical to the treated unit on pre-treatment characteristics.

Comparison to matching: MASC

- Kellogg et. al (2021, JASA): "Combining Matching and Synthetic Control to Trade off Biases From Extrapolation and Interpolation"
- They propose a matching and synthetic control (MASC) as a model averaging estimator that combines the standard SC and matching estimators.
- The MASC estimator is close to the penalized SC estimator by Abadie and L'Hour (2020):
  - Both assign weights to untreated units while taking into consideration their distance from the treated unit in terms of pretreatment characteristics.
  - Kellogg et. al (2021) show that the penalized SC estimator is a particular case of MASC.

Extrapolation vs. interpolation bias



$$\mathrm{ATT}t \equiv Y1t - \gamma\mathrm{t}(x1) = Y1t - \sum_{i>1} wi\,\gamma\mathrm{t}(xi)$$

Where

$$\gamma\mathrm{t}(xi) \equiv \mathbf{E}[Yit|\,\boldsymbol{X}i = \boldsymbol{x}].$$

Consider a simple case in which there are two controls (n = 3), and one x_i (scalar).

Figure 1 plots (x_i, $\gamma_{\mathrm{t}}(x_i)$) for i = 1, 2, 3, as well as $\gamma_{\mathrm{t}}(x)$ as a function of x.

Notice that x_1 lies between x_2 and x_3, so that it is an element of their convex hull.

One way to use the conditional means of the untreated units (*γt(x*2*)* and *γt(x*3*)*) to approximate that of the treated unit (*γt(x*1*)*) is to linearly interpolate between *x*2 and *x*3. The extrapolation bias is 0. However, there is still bias due to interpolation, because *γt(x)* is not a linear function of *x*.

Another way to use the untreated units is to simply use the conditional mean for the unit whose value of *xi* is closest to *x*1. The weights are the nearest neighbor weights (0 weight for x_3), which produce *γt(x*2*)* as an approximation to *γt(x*1*)*. This approach does not interpolate, so However, it does extrapolate, creating bias to the extent that *γt(x*1*) ≠ γt(x*2*)*

Comparison to matching: MASC

- Re-Examining the Spanish terrorism application of Abadie and Gardeazabal (2003).
- The goal is to assess if the alternative estimators yield substantively different estimates:
    - MASC, SC, penalyzed versions of the SC
    - The MASC estimator corresponds exactly to the original SC estimator of Abadie and Gardeazabal (2003).
        - Both average together Catalonia with a weight of 0.85 and Madrid with a weight of 0.15.
    - The Penalized SC places virtually all weight on Catalonia (the nearest neighbor of the Basque Country)
    - The matching places equal weight on Catalonia and Cantabria.

Comparison to matching: MASC

- The MASC and SC imply a cost of conflict of $580 per person per year.
- The Penalyzed SC has a smaller implied cost of conflict ($532 per person per year).
- The matching estimator has much worse fit and implies a positive effect of terrorism on per capita GDP of $331 per person per year.

SC: Recent Developments (Synthetic DD)

- Much of the DD and SC literatures have evolved separately.
- These two strategies are often viewed as targeting different types of empirical applications:
  - DD methods are applied in cases where we have a substantial number of treated units:
    - Minimum wage (Card & Krueger, 1994), $N_{tr} = 321$; $N_{co} = 78$; $T_{pre} = 1$; $T_{post} = 1$
  - SC methods are applied in settings with only a single (or small number) of units exposed:
    - California smoking (ADH, 2010), $N_{tr} = 1$; $N_{co} = 38$; $T_{pre} = 19$; $T_{post} = 12$
- Recent work has begun to try to combine insights from the two literatures

SC: Recent Developments (Synthetic DD)

- Synthetic DD (SDD): Combine SC and DD (Arkhangelsky, et al. 2021, AER).
  - Combine the beneficial features of both synthetic controls and difference in difference estimators.
    - Weight controls and weight time, then applying a type of DD
- To provide intuition for the SDD estimator, it will help to write all estimators in the same manner:

SC: Recent Developments (Synthetic DD)

- Consider a balanced panel with N units and T time periods
- Suppose that the first N_{co} (control) units are never treated, and the last $N_{tr} = N - N_{co}$ (treated) are exposed after time T_{pre}.
- As with SC methods, SDD starts by finding weights $\widehat{w}_i^{SDD}$ that align pre-exposure trends in the outcome of controls with those for treated.
- Time weights $\widehat{\lambda}_t^{SDD}$ balance pre-exposure time periods with post-exposure ones.
- Then these weights are used in a basic TWFE regression with unit and time fixed effects (and with unit and time weights):

$$(\widehat{\tau}^{SDD}, \widehat{\mu}, \widehat{\alpha}, \widehat{\beta}) = \arg\min \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \mu - \alpha_i - \beta_t - D_{it}\tau) \widehat{w}_i^{SDD} \widehat{\lambda}_t^{SDD} \right)^2$$

SC: Recent Developments (SDD)

- In comparison, DD estimates the effect by solving the same TWFE regression without either time or unit weights (unweighted regression)
    - DD with equal weights to all time periods and groups:

$$(\widehat{\tau}^{DD}, \widehat{\mu}, \widehat{\alpha}, \widehat{\beta}) = \arg\min \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \mu - \alpha_i - \beta_t - D_{it}\tau) \right)^2$$

- The use of weights in the SDD estimator makes the TWFE regression "local", in that it puts more weight on units that on average are similar in terms of their past to the treated units, and it emphasizes periods that are on average similar to the treated periods.

SC: Recent Developments (SDD)

- Unit weights $\widehat{w}_i^{SDD}$ are designed so that the average outcome for the treated units is approximately parallel to the weighted average for control units:

$$\widehat{W}^{SDD} = \arg\min_{w} \sum_{t=1}^{T_{pre}} \left(\frac{1}{N_{tr}} \sum_{i=N_{co}+1}^{N} Y_{it} - \sum_{i=1}^{N_{co}} w_i Y_{it} + w_0 \right)^2$$

- $\widehat{w}_i^{SDD}$ are closely related to the $\widehat{w}_i^{SC}$ weights
  - $\widehat{w}_i^{SDD}$ allows for an intercept term w_0, meaning that the weights no longer need to make the controls pre-trends perfectly match the treated ones; rather, it is sufficient that the weights make the trends parallel.

SC: Recent Developments (Synthetic DD)

- $\widehat{\lambda}_t^{SDD}$ are designed so that the average post-treatment outcome for the control units differs just by a constant from the weighted average of the pre-treatment outcomes for the same control units:

$$\begin{aligned} \lambda &= \arg\min_{\lambda} \sum_{i=1}^{N_{co}} \left(\frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} Y_{it} - \sum_{t=1}^{T_{pre}-1} \lambda_t Y_{it} + \lambda_0 \right)^2 \\ s.t.\ \lambda_t &\geq 0, \sum_{t=1}^{T_{pre}-1} \lambda_t = 1 \end{aligned}$$

- Example: one treated, one pre-period:

$$\lambda = \arg\min_{\lambda} \sum_{i=1}^{N-1} \left(Y_{it} - \sum_{t=1}^{T-1} \lambda_t Y_{it} + \lambda_0 \right)^2$$

- Together, these weights make the DD strategy more plausible.
  - This idea is not far from the current empirical practice in which different techniques are used to get parallel trends (i.e adjusting for covariates or selecting appropriate time periods).

SC: Recent Developments (SDD)

- In comparison with the SDD estimator, the SC estimator is a weighted linear regression with no unit FEs:
  - The SC maintains weights w, but does not seek to optimally consider time periods via time weights, and omits unit fixed effects implying that the synthetic control and treated units should maintain approximately equivalent pre-treatment levels, as well as trends:

$$(\widehat{\tau}^{SC}, \widehat{\mu}, \widehat{\beta}) = \arg\min \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \mu - \beta_t - D_{it}\tau) \widehat{w}_i^{SC} \right)$$

  - The argument for including time weights in the SDD estimator is similar as before: $\widehat{\lambda}_t^{SDD}$ can both remove bias and improve precision by eliminating the role of time periods that are very different from the post-treatment periods.
  - Similarly, the argument for the inclusion of the unit fixed effects, α_i, is twofold:
    - First, by making the model more flexible, we strengthen its robustness properties.
    - Second, these α_i can absorb any consistent differences between units and can improve precision.

SC: Recent Developments (SDD)

- SDD offers greater flexibility than both the DD and SC:
  - In the case of DD by permitting a violation of parallel trends in aggregate data (a weighted version of parallel trends)
  - In the case of SC, by both additionally seeking to optimally weight time periods and allowing for level differences between treatment and control groups (the synthetic control can be parallel to treatment, as opposed to identical in pre-treatment period)
- They find that
  - When DD performs well and SC doesn't, SDD performs similarly to or better than DD
  - When SC performs well and DD doesn't, SDD matches or improves upon performance of SC
- Stata commands: `didregress` (for DD), `synth` (for SC), `sdid (for SDD).`

SC: Recent Developments (SDD)

- Conditioning on covariates: in certain settings, it may be of relevance to condition on exogenous time-varying covariates X_{it}.
- Arkhangelsky et al. (2021) note that in this case, we can proceed by applying the SDD algorithm to the residuals:

$$Y_{it}^{res} = Y_{it} - X_{it}\widehat{\beta}$$

- This procedure is different to the logic of SC:
  - In SC when covariates are included the synthetic control is chosen to ensure that these covariates are as closely matched as possible between treated and synthetic control units.
  - However in the SDD, covariate adjustment is viewed as a pre-processing task, which removes the impact of changes in covariates from the outcome prior to calculating the synthetic control.

SC: Recent Developments (SDD)

- Arkhangelsky, et al. (2021) compare $\widehat{\tau}^{DD}$, $\widehat{\tau}^{SC}$, $\widehat{\tau}^{SDD}$ through the evaluation of the California smoking cessation program example of ADH(2010).
- Estimates for average effect of increased cigarette taxes on California per capita cigarette sales over twelve post-treatment years:

	$\widehat{\tau}^{DD}$		$\widehat{\tau}^{SC}$		$\widehat{\tau}^{SDD}$
Estimate	-27.4		-19.8		-13.4
St. error	(16.4)		(7.7)		(7.6)

- As argued in Abadie et al. (2010), the assumptions underlying the DD estimator are suspect here, and the -27.4 point estimate likely overstates the effect of the policy.
- SC provides a reduced (and more credible) estimate of -19.8.
- SDD further attenuates it to -13.4.
- Arkhangelsky, et al. (2021) argue that when $\widehat{\tau}^{DD}, \widehat{\tau}^{Sc,}$ and $\widehat{\tau}^{SDD}$ differ, the latter is often more credible.

SC: Recent Developments (SDD)

- To facilitate direct comparisons, notice that the three estimators can be rewritten as a weighted average difference in adjusted outcomes $\widehat{\delta}$ for appropriate sample weights $\widehat{\omega}_i$.

$$\widehat{\tau} = \widehat{\delta}_{tr} - \sum_{i=1}^{N_{co}} \widehat{\omega}_i \widehat{\delta}_i,$$

  where

$$\widehat{\delta}_{tr} = \frac{1}{N_{tr}} \sum_{i=N_{co}+1}^{N} \widehat{\delta}_i.$$

- The key differences are in the estimation of weights $\widehat{\omega}_i$ and how the adjusted outcomes $_i\widehat{\delta}_i$ are defined.
- DD uses constant weights: $\widehat{\omega}_i = \frac{1}{N_{co}}$, SC and SDD the weights explained above.

SC: Recent Developments (SDD)

- For the adjusted outcomes $\widehat{\delta}_i$, SC uses unweighted treatment period averages:

$$\widehat{\delta}_i^{SC} = \frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} Y_{it}$$

- DD uses unweighted differences between average treatment period and pre-treatment outcomes:

$$\widehat{\delta}_i^{DD} = \frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} Y_{it} - \frac{1}{T_{pre}} \sum_{t=1}^{T_{pre}} Y_{it}$$

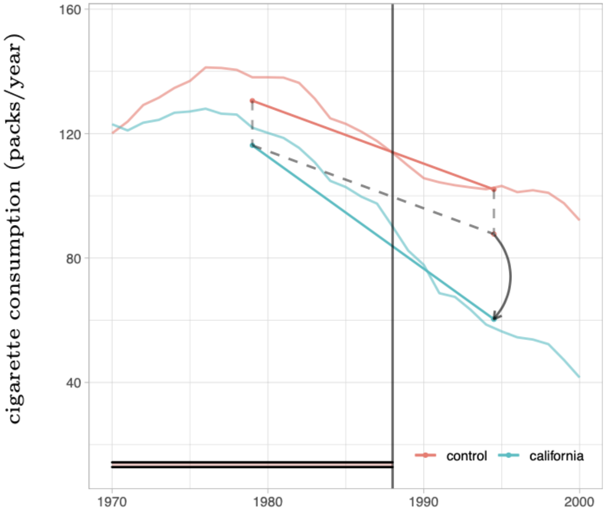
- SDD uses weighted differences of the same:

$$\widehat{\delta}_i^{SDD} = \frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} Y_{it} - \sum_{t=1}^{T_{pre}} \widehat{\lambda}_t^{SDD} Y_{it}$$

SC: Recent Developments (SDD)

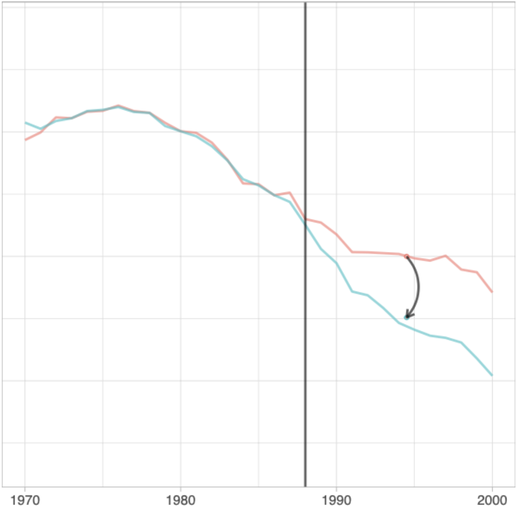
- Next figure illustrates the how each method operates.
- DD relies on the PT assumption absent the intervention.
  - Pre-intervention trends are not parallel, so the DD estimate should be considered suspect.
- SC reweights the control states so that the weighted of outcomes for these states match California pre-intervention.
- SDD re-weights the unexposed control units to make their time trend parallel (but not necessarily identical) to California pre-intervention, and then applies a DD analysis to this re-weighted panel. Moreover, it only focuses on a subset of the pre-intervention time periods:
  - These time periods were selected so that the weighted average of historical outcomes predict average treatment period outcomes for control units, up to a constant.

Difference in Differences



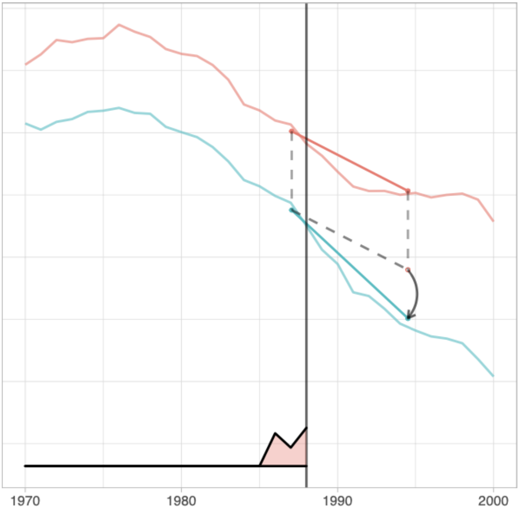
Estimated decrease: -27.3 (17.7)

Synthetic Control



Estimated decrease: -19.6 (9.9); bad fit just prior bc weights are fitting everywhere

Synthetic Diff. in Differences



Estimated decrease: -15.4 (8.4). Jagged line left of 1988 is the weighting of those years

Binary models with PT assumption

- When applying nonlinear models in a DD framework, researchers typically use a linear index structure together with a nonlinear link function.
- The linear index structure is specified as if it would be a specification for the linear regression model.
- However, while the linear regression can be derived from the DD assumptions, such rationalization is generally not possible for nonlinear models.
    - PT assumption holds for the latent variable, but it fails for the binary outcome.

DD in binary models

- Modeling the latent linear index similarly to the linear model case:

$$E\left[Y^*(t) \middle| D \right] = \alpha + \lambda T + \beta D + \gamma DT$$

$$E\left[Y_0^*(2) \middle| D = 1 \right] = \alpha + \lambda + \beta,$$

$$ATT = E\left[Y_1^*(2) - Y_0^*(2) \middle| D = 1 \right]$$

- Under the parallel trend assumption,

$$\begin{aligned} & E\left[Y_0^*(2) \middle| D = 1 \right] \\ = \; & E\left[Y_0^*(1) \middle| D = 1 \right] + \left(E\left[Y_0^*(2) \middle| D = 0 \right] - E\left[Y_0^*(1) \middle| D = 0 \right] \right) \\ = \; & [\alpha + \beta] + [(\alpha + \lambda) - \alpha] = \alpha + \lambda + \beta \end{aligned}$$

$$\gamma = E\left[Y^*\left(2\right) - Y^*\left(1\right) \middle| D = 1 \right] - E\left[Y^*\left(2\right) - Y^*\left(1\right) \middle| D = 0 \right] = ATT$$

DD in binary models

- We observe

$$\begin{aligned} Y &= 1 \text{ if } Y^* > 0 \\ Y &= 0 \text{ if } Y^* < 0 \end{aligned}$$

  Using the link function $G(\cdot)$, the conditional expectations of the binary potential outcomes Y_0 and Y_1 become:

$$E\left[\, Y(t) \middle|\, D\right] = G\left(\alpha + \lambda T + \beta D + \gamma (D \times T)\right)$$

$$E\left[\, Y_0(2) \middle|\, D = 1\right] = G\left(\alpha + \lambda + \beta\right)$$

- This specification resembles the linear model with the exception of the addition of the link function $G(\cdot)$.
- Notice that in this model the effect of T in the absence of the treatment is not constant (different for individuals with $D = 1$ and with $D = 0$) as opposed to what happens in linear model.

DD in binary models

- In this case, the effect of T for the controls:

$$E\left[\, Y_0\,(2) - Y_0\,(1) \middle|\, D = 0\right] = G(\alpha + \lambda) - G(\alpha)\ \left[= \lambda \text{ in linear case}\right]$$

- The effect of T for the treated in the absence of treatment:

$$E\left[\, Y_0\,(2) - Y_0\,(1) \middle|\, D = 1\right] = G(\alpha + \lambda + \beta) - G(\alpha + \beta)\left[= \lambda \text{ in linear case}\right]$$

- Thus, the intuitive specification does not fulfill the common trend assumption (see Blundell and Powell, 2004).
- The common trend assumption would only be held if β were zero.

DD in binary models

- This means that:

$$\begin{aligned} ATT &= E\left[\left. Y_1(2) - Y_0(2) \right| D = 1\right] \\ &\neq E\left[\left. Y(2) - Y(1) \right| D = 1\right] - E\left[\left. Y(2) - Y(1) \right| D = 0\right], \end{aligned}$$

  where

$$\begin{aligned} & E\left[\left. Y(2) - Y(1) \right| D = 1\right] - E\left[\left. Y(2) - Y(1) \right| D = 0\right] \\ = \; & \left[G(\alpha + \lambda + \beta + \gamma) - G(\alpha + \beta)\right] - \left[G(\alpha + \lambda) - G(\alpha)\right] \end{aligned}$$

- In this case, the counterfactual is:

$$\begin{aligned} & E\left[\left. Y_0(2) \right| D = 1\right] \\ = \; & E\left[\left. Y_0(1) \right| D = 1\right] + E\left[\left. Y_0(2) \right| D = 0\right] - E\left[\left. Y_0(1) \right| D = 0\right] \\ = \; & G(\alpha + \lambda + \beta) \neq G(\alpha + \beta) + \left[G(\alpha + \lambda) - G(\alpha)\right] \end{aligned}$$

DD in binary models

- How can we obtain the parameter $ATT = E\left[Y_1(2) - Y_0(2) \middle| D = 1 \right]$?
    - ATT is equivalent to the partial effect of $W = (D \times T)$, evaluated at $D = 1$ and $T = 1$
    - For the treated group the potential outcome under treatment $E\left[Y_1(2) \middle| D = 1 \right]$ is observed:

$$E\left[Y_1(2) \middle| D = 1 \right] = G\left(\alpha + \lambda + \beta + \gamma \right)$$

    - The counterfactual outcome $E\left[Y_0(2) \middle| D = 1 \right]$ is unobserved, but modeled as

$$E\left[Y_0(2) \middle| D = 1 \right] = G\left(\alpha + \lambda + \beta \right)$$

- Thus:

$$\begin{aligned} ATT &= E\left[Y_1(2) - Y_0(2) \middle| D = 1 \right] \\ &= G\left(\alpha + \lambda + \beta + \gamma \right) - G\left(\alpha + \lambda + \beta \right) \end{aligned}$$

- Another alternative is to use a linear specification (LPM), despite its problematic features for outcome variables with bounded support.

```
 . /*** PROBIT MODEL ***/
. gen DT=D*T

.  probit Y i.D i.T i.DT

Iteration 0:   log likelihood = -1999.0668
Iteration 1:   log likelihood = -1986.6359
Iteration 2:   log likelihood = -1986.4776
Iteration 3:   log likelihood = -1986.4776

Probit regression                          Number of obs = 10,335

------------------------------------------------------------------------------
           Y | Coefficient        Std. err.      z    P>|z|        [95% conf. interval]
-------------+----------------------------------------------------------------
         1.D |   .3112768   .0896121     3.47   0.001     .1356404    .4869132
         1.T |   .0826837   .0458724     1.80   0.071    -.0072246     .172592
        1.DT |  -.0425108   .1207327    -0.35   0.725    -.2791425    .1941209
       _cons |  -1.743779   .0340777   -51.17   0.000    -1.810571   -1.676988
------------------------------------------------------------------------------


.
.  /*************MARGINAL EFFECTS OF INTEREST********/
.  margins, dydx(DT) at (D==1 T==1)

Conditional marginal effects                          Number of obs = 10,335
Model VCE: OIM

Expression: Pr(Y), predict()
dy/dx wrt:  1.DT
At: D = 1
    T = 1

------------------------------------------------------------------------------
             |            Delta-method
             |     dy/dx        std. err.      z      P>|z|          [95% conf. interval]
-------------+----------------------------------------------------------------
        1.DT |  -.0066258   .0189542    -0.35   0.727    -.0437753    .0305238
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

```
/************* WRONG MARGINAL EFFECTS*****************/
.  probit Y i.D##i.T


Probit regression                        Number of obs = 10,335

-----------------------------------------------------------------------------
         Y   | Coefficient    Std. err.      z       P>|z|     [95% conf. interval]
-------------+---------------------------------------------------------------
       1.D |   .3112768   .0896121     3.47   0.001     .1356404    .4869132
       1.T |   .0826837   .0458724     1.80   0.071    -.0072246     .172592
             |
     D#T |
      1 1  |  -.0425108   .1207327    -0.35   0.725    -.2791425    .1941209
             |
   _cons |  -1.743779   .0340777   -51.17   0.000    -1.810571   -1.676988
-----------------------------------------------------------------------------

.
.   margins, dydx(T) over (D)

Average marginal effects                          Number of obs = 10,335
Model VCE: OIM

Expression: Pr(Y), predict()
dy/dx wrt:  1.T
Over:       D

-----------------------------------------------------------------------------
            |            Delta-method
            |       dy/dx          std. err.      z     P>|z|       [95% conf. interval]
-------------+---------------------------------------------------------------
0.T         |  (base outcome)
-------------+---------------------------------------------------------------
1.T         |
         D |
         0  |   .0077483   .0042826     1.81   0.070    -.0006454     .016142
         1  |   .0059113   .0163945     0.36   0.718    -.0262213    .0380438
-----------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.


. di .0059113 -.0077483
-.001837
```