STA 640 — Causal Inference

Chapter 8: Causal Inference with time: Difference-in-Differences, Synthetic Control, and Beyond

Fan Li

Department of Statistical Science
Duke University

Treatment-Control Before-After Design

- A treatment-control comparison is not necessarily a causal comparison because of the potential systematic differences between two groups
- A unit is arguably the "best match" for itself
- A before-after comparison (of the same units) is not necessarily a causal comparison because of the potential change in time
- Treatment-control before-after design combines both
- Setup: two or more groups, with units observed in two or more periods. In some periods and some groups are exposed to the treatment (repeated cross-sections, or panel data)

Treatment-Control Before-After: Examples

- Example 1 (labor economics): Evaluate effect of the minimum wage on employment (Card and Krueger, 1994)
  - units: fast-food restaurants in New Jersey and adjacent eastern PA
  - intervention: raise of the state minimum wage; NJ raised the minimum on April 1, 1992, but not PA
  - outcome: employment (average per store), observed in both areas, and both right-before and after the change
- Classic study in labor economics, updated in Card and Krueger (2000) with additional number of years
- In most economics and policy application, the study setting is a policy change that would affect only some but not all clusters (region, age, etc.)

Treatment-Control Before-After: Examples

- Example 2 (transportation safety research): Evaluate effect of safety countermeasures on accident rate
  - units: segments of roads (with similar characteristics)
  - intervention: some new safety countermeasure (e.g. pavement of road, lighting conditions), only some units are treated
  - outcome: counts of accidents, observed before and after for all segments under study
- This setup and associated models is known as the Empirical-Bayes before-after design (Hauer, 1997)

Two-Group Before-After Design: Basic setup

- Two groups of units: $G_i = 0, 1$ ($G_i = 1$ treatment)
- Two periods of time: $T = t, t+1$ (before (t) and after ($t+1$))
- For each unit i, we observe outcome in each period: $Y_{i,t}, Y_{i,t+1}$, and a vector of covariates X_i
- The two-by-two table of the two-group before-after design (drop i for simplicity)

	before	after
control	$Y_{0,t}$	$Y_{0,t+1}$
treatment	$Y_{1,t}$	$Y_{1,t+1}$

- Denote the number of units in each cell of G, T as $N_{G,T}$.

Potential Outcomes

- Two different settings:
  - Panel: the same units being followed for multiple periods ($N_{G,T} = N_G$)
  - Repeated cross-section: each period has a different (possibly overlapped) sample of units, ($N_{G,T}$ are different across G, T)
- Potential outcomes at time T: $Y_T(1), Y_T(0)$
- Since the treatment is only administered at $T = t + 1$, define the treatment status: $Z_{iT} = G_i \cdot 1\{T = t + 1\}$, equal to one only for the treatment group in the after period
- The two-by-two table of the two-group before-after design

	before	after
control	$Y_t(0)$	$Y_{t+1}(0)$
treatment	$Y_t(0)$	$Y_{t+1}(1)$

Causal estimand

- Average treatment effect on the treated (ATT):

$$\tau_{\text{ATT}} = E\{Y_{i,t+1}(1) - Y_{i,t+1}(0) \mid G_i = 1\} = \mu_1 - \mu_0$$

  - $\mu_1 = \mathbb{E}\{Y_{i,t+1}(1) \mid G_i = 1\} = \mathbb{E}(Y_{i,t+1} \mid G_i = 1)$ identifiable
  - $\mu_0 = \mathbb{E}\{Y_{i,t+1}(0) \mid G_i = 1\}$: counterfactual
- Task: infer μ_0 based on observed data
- Two identification strategies, based on different assumptions:
  - Difference-in-differences (DID)
  - Lagged-dependent-variable adjustment (LDV)
- For simplicity, the following discussions are implicitly conditional on X

Before-After Estimator

- Before-after estimator: compare treated units with their "untreated selves" in the previous wave:

$$\hat{\tau}^{BA} = \bar{Y}_{1,t+1} - \bar{Y}_{1,t}$$

where $\bar{Y}_{1,t}$ is the sample mean outcome for units in treatment group at time t

- Identification assumption: $\hat{\tau}^{BA}$ uses outcomes before treatment (at t) as proxy for counterfactual outcomes at $t+1$ in the absence of treatment, i.e.

$$\mathbb{E}[Y_{t+1}(0)|G = 1] = \mathbb{E}[Y_t(0)|G = 1]$$

- Problem: Even in the absence of treatment, outcomes may have changed between t and $t+1$ because of time trends, macro factors, lifecycle effects...

Cross-Sectional Estimator

- Cross-Sectional estimator: compare treated units with control units in the same wave:

$$\hat{\tau}^{CS} = \bar{Y}_{1,t+1} - \bar{Y}_{0,t+1}$$

where $\bar{Y}_{1,t}$ ($\bar{Y}_{0,t}$) is the sample mean outcome for units in treatment (control) group at time t

- Identification assumption: $\hat{\tau}^{CS}$ uses outcomes in the control group at $t+1$ as proxy for counterfactual outcomes at $t+1$ in the absence of treatment, i.e.

$$\mathbb{E}[Y_{t+1}(0)|G = 1] = \mathbb{E}[Y_{t+1}(0)|G = 0]$$

- Problem: Even in the absence of treatment, treatment and control groups may be systematically different in measured and unmeasured way

Strategy 1: Difference-in-differences (DID)

DID solution: leverage the 2×2 design with a central assumption

- "*Parallel trends* " assumption: The treatment group and the control group experience the same trends in the absence of treatment:

$$\mathbb{E}[Y_{t+1}(0) - Y_t(0)|G = 1] = \mathbb{E}[Y_{t+1}(0) - Y_t(0)|G = 0] \quad (1)$$

- The parallel trend assumption (1) is equivalent to the "*constant difference*" assumption: Difference between treatment and control group in the absence of treatment is constant across time

$$\begin{aligned} & \mathbb{E}[Y_{t+1}(0)|G = 1] - \mathbb{E}[Y_{t+1}(0)|G = 0] \\ = \quad & \mathbb{E}[Y_t(0)|G = 1] - \mathbb{E}[Y_t(0)|G = 0] \end{aligned} \quad (2)$$

- Parallel trends is scale-dependent: hold for Y but may not for a nonlinear monotone transformation of Y (Athey and Imbens, 2006)

Graph Illustration

Angrist and Pischke, 2009: Mostly Harmless Econometrics

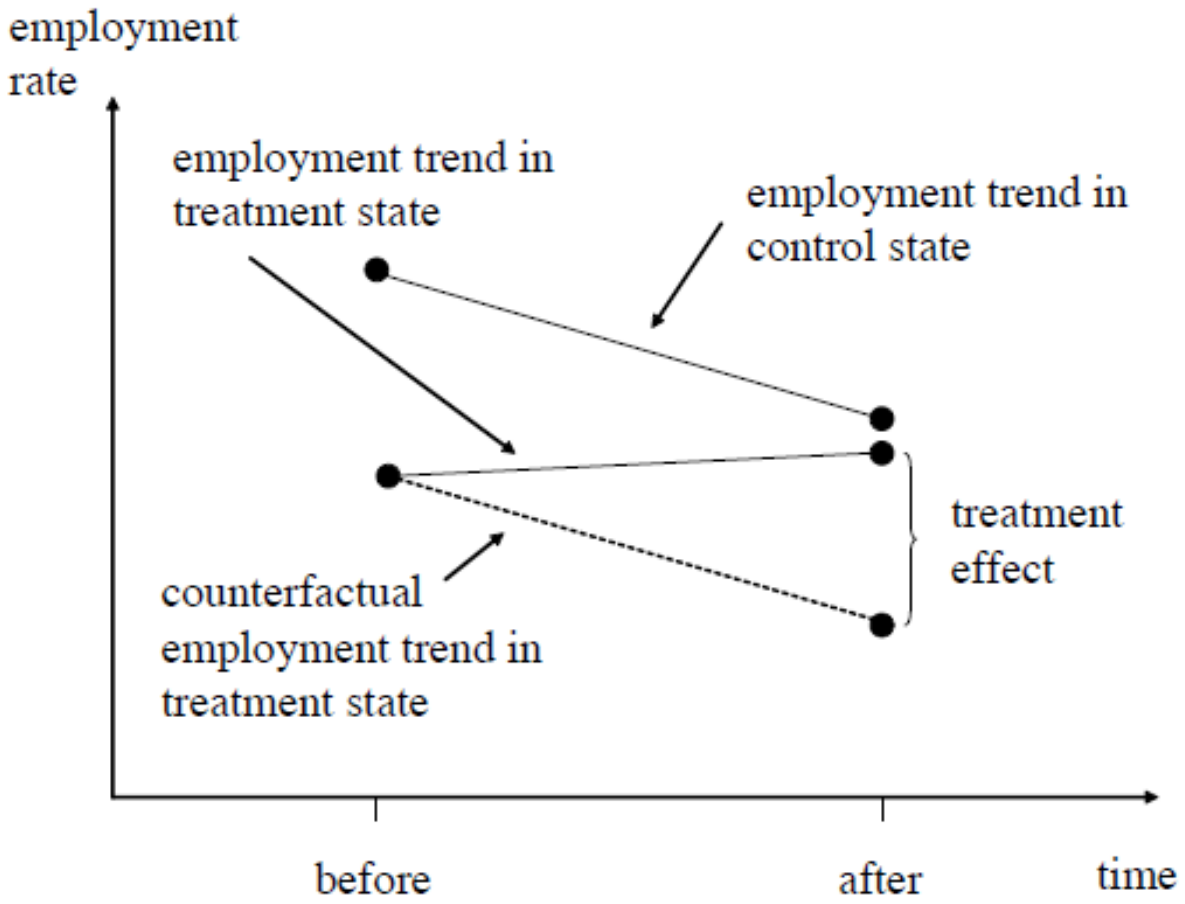


Figure: Causal effects in the difference-in-differences design

DID: Identification and estimation

- Under parallel trends, DID estimand is nonparametrically identified:

$$\tilde{\mu}_{0,\text{DID}} = \mathbb{E}(Y_{it} \mid G_i = 1) + \mathbb{E}(Y_{i,t+1} \mid G_i = 0) - \mathbb{E}(Y_{it} \mid G_i = 0)$$

- DID estimator takes the difference between the change in outcomes for treated individuals and the change for control individuals
- Equivalently, DID estimator can be written as the difference between the before-after estimators of treatment and of control groups

$$\begin{aligned} \hat{\tau}_{\text{DID}} &= (\bar{Y}_{1,t+1} - \bar{Y}_{1,t}) - (\bar{Y}_{0,t+1} - \bar{Y}_{0,t}) \\ &= \hat{\tau}_1^{BA} - \hat{\tau}_0^{BA} \end{aligned}$$

where $\bar{Y}_{g,t}$ are mean outcomes for group g at time t

Why DID works?

- Assuming SUTVA, we have

$$Y_t = Y_t(0), \quad Y_{t+1} = (1-G)Y_{t+1}(0) + GY_{t+1}(1),$$

- Therefore

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_0^{BA}) &= \mathbb{E}(\bar{Y}_{0,t+1} - \bar{Y}_{0,t}) \\
&= \mathbb{E}[Y_{t+1}(0) - Y_t(0)|G=0] \quad \text{(SUTVA)} \\
&= \mathbb{E}[Y_{t+1}(0) - Y_t(0)|G=1] \quad \text{(par. trends)} \\
&= \mathbb{E}[Y_{t+1}(0)|G=1] - \mathbb{E}[Y_t|G=1] \quad \text{(SUTVA)}
\end{aligned}$$

  Similarly, we have

$$\mathbb{E}(\hat{\tau}_1^{BA}) = \mathbb{E}[Y_{t+1}(1)|G=1] - \mathbb{E}[Y_t|G=1]$$

- Subtracting the two, we have

$$\mathbb{E}(\hat{\tau}_{\text{DID}}) = \mathbb{E}(\hat{\tau}_1^{BA} - \hat{\tau}_0^{BA}) = \mathbb{E}[Y_{t+1}(1) - Y_{t+1}(0)|G=1] = \tau_{\text{ATT}}$$

DID Estimator: Alternative View

- Alternatively, DID estimator can be written as the difference between the cross-sectional estimators at times $t+1$ and t

$$\begin{aligned}\hat{\tau}_{\text{DID}} &= (\bar{Y}_{1,t+1} - \bar{Y}_{0,t+1}) - (\bar{Y}_{1,t} - \bar{Y}_{0,t}) \\ &= \hat{\tau}^{CS}_{t+1} - \hat{\tau}^{CS}_{t}\end{aligned}$$

- Based on the "*constant difference*" interpretation of the parallel trend assumption (equation (2)), easy to show the above DID estimator is unbiased for τ

- DID uses double difference to eliminate the bias of the one-group before-after estimator (eliminate time trend) and of the one-time cross-sectional estimator (eliminate group difference)

DID estimator: Variance

- Under homoscedasticity, variance of the DID estimator is

$$\mathbb{V}(\hat{\tau}_{\text{DID}})_{homo} = \sigma^2 \cdot \left(\frac{1}{N_{11}} + \frac{1}{N_{10}} + \frac{1}{N_{01}} + \frac{1}{N_{00}} \right)$$

- Under heteroscedasticity, robust variance of the DID estimator is

$$\mathbb{V}(\hat{\tau}_{\text{DID}})_{hetero} = \frac{\sigma^2_{11}}{N_{11}} + \frac{\sigma^2_{10}}{N_{10}} + \frac{\sigma^2_{01}}{N_{01}} + \frac{\sigma^2_{00}}{N_{00}}$$

where the within-cell variance σ^2_{gt} can be estimated by

$$S^2_{gt} = \frac{1}{N_{gt} - 1} \sum_{i:G_i=g, T_i=t} (Y_i - \bar{Y}_{gt})^2$$

DID: A Regression Perspective

- Above shows the nonparametric identification of the DID estimator
- For many complex situations (multiple groups, multiple periods, covariates), a parametric/modeling perspective is more flexible
- In fact, in the econ literature, DID is usually tied toa regression model with fixed effects for the potential outcomes in the absence of treatment (Angrist and Pischke, 2009, Chapter 5)

DID: A Regression Perspective

- The core of the regression DID model is an additive fixed effects model for $Y(0)$ (omitting X):

$$Y_{iT}(0) = \alpha + \gamma G_i + \delta_T + \epsilon_{iT}, \tag{3}$$

  with $\mathbb{E}(\epsilon_{iT}|G_i, T) = 0$; here γ is the group fixed effect and δ_T is the time effect
- Denote $\alpha_i \equiv \alpha + \gamma G_i$: individual time-fixed effect,
- Model (3) implies the parallel trends assumption: for $g = 0, 1$

$$\mathbb{E}[Y_{t+1}(0)|G = g] - \mathbb{E}[Y_t(0)|G = g] = \delta_{t+1} - \delta_t$$

- Similarly, Model (3) implies the constant difference assumption: for any T

$$\mathbb{E}[Y_T(0)|G = 1] - \mathbb{E}[Y_T(0)|G = 0] = \gamma$$

DID: A Regression Perspective

- Besides Model (3), postulate the following model for the observed outcomes in DID (omitting X here):

$$Y_{iT} = \underbrace{\alpha + \gamma G_i}_{\alpha_i} + \delta_T + \underbrace{\tau \cdot Z_{iT}}_{\tau G_i \cdot 1\{T=t+1\}} + \epsilon_{iT} \tag{4}$$

  with $\mathbb{E}(\epsilon_{iT}|G_i, T) = 0$

- Coefficient τ in Model (4) is the ATT estimand

- Easy to show that double-differencing Model (4) gives τ_{ATT}:

$$\tau = \mathbb{E}[(\bar{Y}_{1,t+1} - \bar{Y}_{1,t}) - (\bar{Y}_{0,t+1} - \bar{Y}_{0,t})] = \tau_{\text{ATT}}$$

  So the logic follows the nonparametric DID idea

Causal Interpretation

- Model (4) translates into a model for the potential outcomes $Y(1)$:

$$Y_{iT}(1) = \alpha_i + \delta_T + \tau + \epsilon_{iT}, \tag{5}$$

- Model (5) (joint with Model 3) effectively assumes homogeneous treatment effect τ for all units for all T

$$\tau = Y_{iT}(1) - Y_{iT}(0) = \mathbb{E}[Y_{iT}(1) - Y_{iT}(0)]$$

- So τ is a causal effect

- Is τ ATT or ATE? Depending on what population one imposes Model (5)

Causal Interpretation: Remarks

- Though Model (5) can be posed on all units, it is sufficient to pose it only on the treated units to identify the ATT, which is the DID estimand in the nonparametric setup

- In fact, because $Y(1)$ is only observable for the treated units in the second (treated) period, one would only be able to make causal claims for that subpopulation (ATT), regardless of model assumptions

- The additivity (of group-specific and time-specific effects) assumption of the fixed effects model (3)-(5) is not required in the nonparametric setup

DID: A Regression Perspective

- Pros of regression-based DID:
  - easy inference (point estimate and standard errors)
  - easy to incorporate covariates
  - easy to extend to multiple periods and treatments
- Cons of regression-based DID:
  - parametric assumptions, subject to misspecification
- Athey and Imbens (2006, Econometrica) extended DID to non-linear models (called change-in-change model) for both continuous and discrete data
- Abadie (2005, RESTAT) proposed a semiparametric approach based on inverse probability weighting (IPW)
- Double-robust DID is also available (Callaway and Sant'Anna, 2021; Li and Li, 2019)

Parallel Trends Assumption

- So far, the discussion assumed away $X \Rightarrow$ conditional DID within strata of X
- The conditional parallel trends (Heckman, 1997)

$$\mathbb{E}[Y_{t+1}(0) - Y_t(0)|X, G = 1] = \mathbb{E}[Y_{t+1}(0) - Y_t(0)|X, G = 0]$$

- DID permits unobserved confounders (U) to affect assignment as long as their impact on $Y(0)$'s is separable and time-invariant (Lechner, 2001).

Parallel Trends Assumption

- If the following holds

$$\mathbb{E}[Y_{t+1}(0) - Y_t(0)|X, U, G = 1] = \mathbb{E}[Y_{t+1}(0) - Y_t(0)|X, U, G = 0]$$

then parallel trends holds under

$$\mathbb{E}[Y_T(0)|X, U, G = g] = \mathbb{E}[Y_T(0)|X, G = g] + h_g(U, X), \ \forall \ g, T$$

- Inference on causal parameters is still feasible with additive and time-fixed confounding, as long as the parallel trends is assumed:

$$Y_{it} = \alpha_i + \delta_t + \tau \cdot Z_{it} + X_i'\beta + \epsilon_{it} \tag{6}$$

Recall $\alpha_i \equiv \alpha + \gamma G_i$

Parallel Trends Assumption

- DID rests entirely on the validity of the parallel trends assumption
- The parallel trends assumption may be violated in practice:
  - Time trends or macro factors may affect the two groups differently
  - There may be time-varying factors affecting only one group (e.g., “Ashenfelter’s dip" (Ashenfelter (1978)): often trainees have a temporary drop in earnings before they take up training course

$\Rightarrow$ In DID tests, much effort is devoted to show that trends are indeed parallel in different groups

- Use negative control outcomes to indirectly test the comparability
- At least, need to show that in the before period the comparison groups are similar in key covariates

Multiple Control Groups

- DID can be viewed a case of multiple (three) control groups (Rosenbaum, 2000)
- Essentially DID uses three control groups to estimate the potential outcome $Y(0)$ of the treatment group
- Recall we have discussed before using multiple control groups to indirectly test the unconfoundedness assumption, similar logic applies here
- If outcomes in the three control groups were similar, the results would be more credible than the case where both groups are very different initially, and the change over time in the control group is more substantial

DID with Multiple Periods: Serial Correlation

Bertrand et al., 2004, QJE

- DID setup often involves multiple periods
- Serial correlation: only few observations are "truly independent." Particularly severe if:
  - Outcomes are serially correlated
  - Treatment variable changes little over time (e.g, just once)
  - Panel uses many years of data
- One solution: collapse data into one observation per unit "before", and one "after" treatment
  - Takes care of serial correlation, but power declines fast
- No collapsing: conventional variance estimator often led to over-rejection, and cluster-robust (sandwich) variance should become standard practice

Serial correlation: simple fixes

- Collapse data into one observation per firm "before" and one "after" treatment
  - Works if treatment is passed at same time for all treated groups
  - Takes care of serial correlation
- If treatments are staggered (i.e. passed at different times), "before" and "after" are no longer the same for each treated state and not even defined for control states
  - First, regress Y_{gt} on group fixed effects, time dummies, and any relevant covariates
  - Then divide residuals of treatment states only into two groups: residuals from years before, and residuals from years after
  - Then estimate effects and its standard error from an OLS regression in this two-period panel

DID with multiple-periods and staggered adoption

- Using the fixed effects model, straightforward to extend to the setting of multiple-periods multiple-groups: add fixed effects of time and group
- But causal interpretation is tricky
- Callaway and Sant'Anna (2021): Rigorously formulated DID with multiple time periods with staggered adoption using potential outcomes
  - Define "Group-time average treatment effect": the average treatment effect for group g at time t, where a "group" is defined by the time period when units are first treated
- One of the most cited econometrics/causal papers in recent years
- More details in the "additional notes" at the end of the lecture

Strategy 2: Lagged-dependent-variable adjustment

- Assumption: Ignorability (or unconfoundedness) conditional on LDV (omitted X here)

$$Y_{i,t+1}(0) \perp G_i \mid Y_{it}$$

- Nonparametric identification under ignorability:

$$\begin{aligned} \tilde{\mu}_{0,\text{LDV}} &= \mathbb{E}\{\mathbb{E}(Y_{t+1} \mid G = 0, Y_t) \mid G = 1\} \\ &= \int \mathbb{E}(Y_{t+1} \mid G = 0, Y_t = y) F_t(\mathrm{d}y \mid G = 1) \end{aligned}$$

- $F_t(y \mid G = g) = \Pr(Y_t \leq y \mid G = g)$: the cumulative distribution function of Y_t for units in group g $(g = 0, 1)$
- Ignorability is scale-free
- Estimation strategy: use lagged outcome (confounder) Y_t to predict $Y_{t+1}(0)$

Semiparametric identification

- Under either DID (i.e. parallel trends) or LDV (i.e. ignorability), we can use inverse probability weighting (IPW) to semiparatrically identify the estimand τ_{ATT}
  - Under DID, define the propensity score as $e = \Pr(G = 1)$ (ignore X here), a semiparametric identification formula for μ_0 is (Abadie, 2005)

$$\tilde{\mu}_{0,\text{DID}} = \mathbb{E}\left\{GY_t + \frac{e(1-G)(Y_{t+1} - Y_t)}{1-e}\right\} \Big/ \Pr(G = 1), \quad (7)$$

  - Under LDV, define the propensity score as $e = \Pr(G = 1 \mid Y_t)$ (ignore X here), a semiparametric identification formula for μ_0 is (essentially ATT weighting)

$$\tilde{\mu}_{0,\text{LDV}} = \mathbb{E}\left\{\frac{e(Y_t)}{1-e(Y_t)}(1-G)Y_{t+1}\right\} \Big/ \Pr(G = 1),$$

LDV: model-based estimation

- Under the ignorability assumption, one can simply impute the missing potential outcome $Y_{i,t+1}(0)$ by a regression model of Y_{t+1} with the LDV Y_t as a predictor:

$$Y_{i,t+1} = \alpha + \beta \cdot Y_{it} + \lambda_{t+1} + \tau \cdot Z_{i,t+1} + \epsilon_{i,t+1} \tag{8}$$

- Coefficient τ in Model (8) is the causal estimand in LDV approach
- OLS estimate of τ in Model (8) is the LDV estimator $\hat{\tau}_{\text{LDV}}$
- Another approach is via matching of pre-treatment variables

DID vs. LDV

- The LDV model (8) is very different from the fixed effects model (4)
- A fixed effects model (parallel/independent trend) assumes:

$$\{Y_{i,t+1}(0) - Y_{it}(0)\} \perp Z_{i,t+1} | \alpha_i \tag{9}$$

  where α_i is a subject-specific fixed effects (equation (6))

- How about combine DID and LDV, i.e., add the lagged outcome to a fixed effects model?

$$Y_{i,t+1} = \alpha_i + \beta \cdot Y_{it} + \lambda_{t+1} + \tau \cdot Z_{i,t+1} + \epsilon_{i,t+1} \tag{10}$$

- Unfortunately, OLS estimates of Model (10) are not consistent (e.g. Angrist and Pischke, 2009, Chap 5.3)

DID vs. LDV

- Parallel trends and ignorability: two different assumptions, not nested
- In practice, at best one hopes: one of the two holds; which one is unknown
- Two popular methods, which one to use?
- Some simulations were done (e.g. O'Neil et al. 2016)
- Angrist and Pischke (2009): A bracketing relationship under a specific type of linear models
- Can we say something more intrinsic and general, and with practical implications?

A bracketing relationship between DID and LDV

Angrist and Pischke (2009), Ding and Li (2019)

- Angrist and Pischke (2009, Chapter 5 Appendix) showed a bracketing relationship between DID and LDV estimator in the setting of a linear regression model: if the true effect is positive, then
  - If LDV model (8) is correct, but you mistakenly use fixed effects model (4), you will overestimate the true effect
  - If the fixed effects model (4) is correct, but you mistakenly use the LDV model (8), you will underestimate the true effect
- The relationship reversed for a negative true effect
- Ding and Li (2019): extended AP's results to model-free setting: under mild conditions, the bracket relationship holds regardless of the model assumptions on the outcome (more technical details in additional notes)

Practical guide

- Graphical checks: draw the scatterplot of before-after outcome of control group, see how far the data cloud is from a linear fit
- Report results from both DID and LDV, giving a bracket that bounds the true effect if one of the two assumptions is correct
- Conduct sensitivity analysis to see what happens if neither is correct
- Note: The bracketing result does not answer: if neither assumption holds, whether the true effect falls inside or, if outside, which side of the bracket? It depends.

Example: Minimum wage and employment (revisited)

(Card and Krueger, 1994)

- Effect of minimum wage increase on employment
- Employment information in New Jersey and Pennsylvania before and after a minimum wage increase in NJ in 1992
- Units: fast food restaurant
- Outcome:# employees at each restaurant

Example: Graphical checks

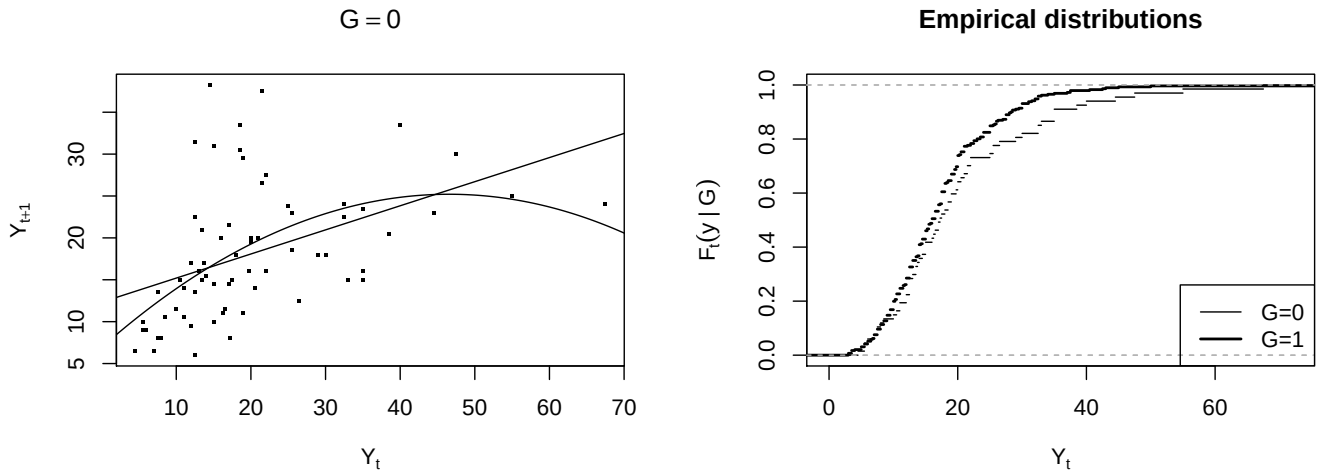


Figure: Left: linear and quadratic fitted lines of $E(Y_{t+1} \mid G = 0, Y_t)$. Right: empirical $F_t(y \mid G = g)$ $(g = 0, 1)$ satisfies Stochastic Monotonicity (1).

- Coefficients of the lag outcome: $\hat{\beta} = 0.288 < 1, \hat{\beta}' = 0.475 < 1$
- Theory predicts $\hat{\tau}_{\text{DID}} > \hat{\tau}_{\text{LDV}}$
- Empirical estimates match theoretical prediction

  $\hat{\tau}_{\text{DID}} = 2.446, \quad \hat{\tau}_{\text{LDV}} = 0.302, \quad \hat{\tau}'_{\text{LDV}} = 0.865$
- The same conclusion under a quadratic model

Example 2: Short-term electoral returns

(Bechtel and Hainmueller, 2011)

- Goal: evaluate causal effect of a disaster relief aid due to the 2002 Elbe flooding in Germany
- Before period: 1998; after period: 2002
- Units: electoral districts
- Treatment: whether a district is affected by the flood
- Outcome: vote share of the Social Democratic Party

Example 2: Graphical checks

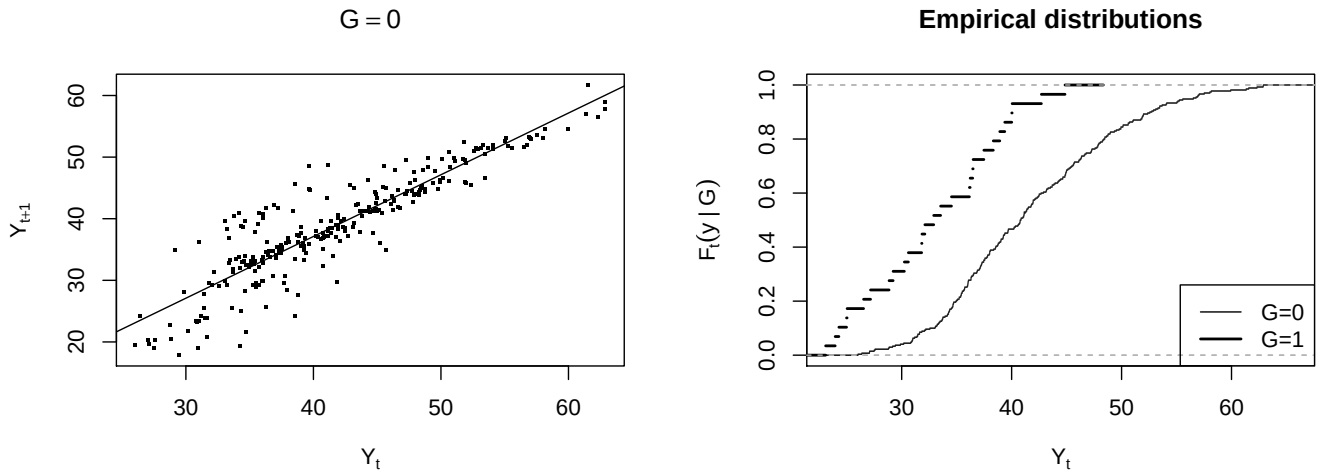


Figure: Left: linear fitted lines of $E(Y_{t+1} \mid G = 0, Y_t)$. Right: empirical $F_t(y \mid G = g)$ $(g = 0, 1)$ satisfy Stochastic Monotonicity (1).

- Coefficients of the lag outcome $\hat{\beta} = 1.002 > 1$ and $\hat{\beta}' = 0.997 < 1$
- Theory predict $\hat{\tau}_{\text{DID}} \approx \hat{\tau}_{\text{LDV}}$
- Empirical estimates (almost identical) match theoretical prediction: $\hat{\tau}_{\text{DID}} = 7.144, \quad \hat{\tau}_{\text{LDV}} = 7.160, \quad \hat{\tau}'_{\text{LDV}} = 7.121$

Comparative case studies

- Policy interventions often take place at an aggregate level, and affect aggregate entities, such as schools, or geographic or administrative areas
- Comparative case studies: the evolution of aggregate outcomes (such as mortality rates, average income, crime rates, etc.) for a unit affected by a particular intervention and compare it to the evolution of the same aggregates estimated for some control group of unaffected units
- Data feature: (1) only one or a few treated units, and many more control units; (2) long time series both before and after
- Shape of data matrix (rows are units, columns are time): short and wide
- The synthetic control (SC) method (Abadie and Gardeazabal, 2003; Abadie, Diamond, Hainmuller, 2010) is the most popular method for evaluating comparative case studies

Example: California's Proposition 99

Abadie, Diamond, Hainmuller, 2010

- In 1988, California first passed comprehensive tobacco control legislation
  - increased cigarette tax by 25 cents/pack
  - earmarked tax revenues to health and anti-smoking budgets
  - funded anti-smoking media campaigns
  - spurred clean-air ordinances throughout the state
  - produced more than $100 million per year in anti-tobacco projects
- 38 other US states were included in the set of control units (donor pool, excluding states passing similar programs)
- Interested in the cause effect of Proposition 99 on per-capita cigarette sales in CA

Cigarette Consumption: CA and Other US States

Abadie, Diamond, Hainmuller, 2010

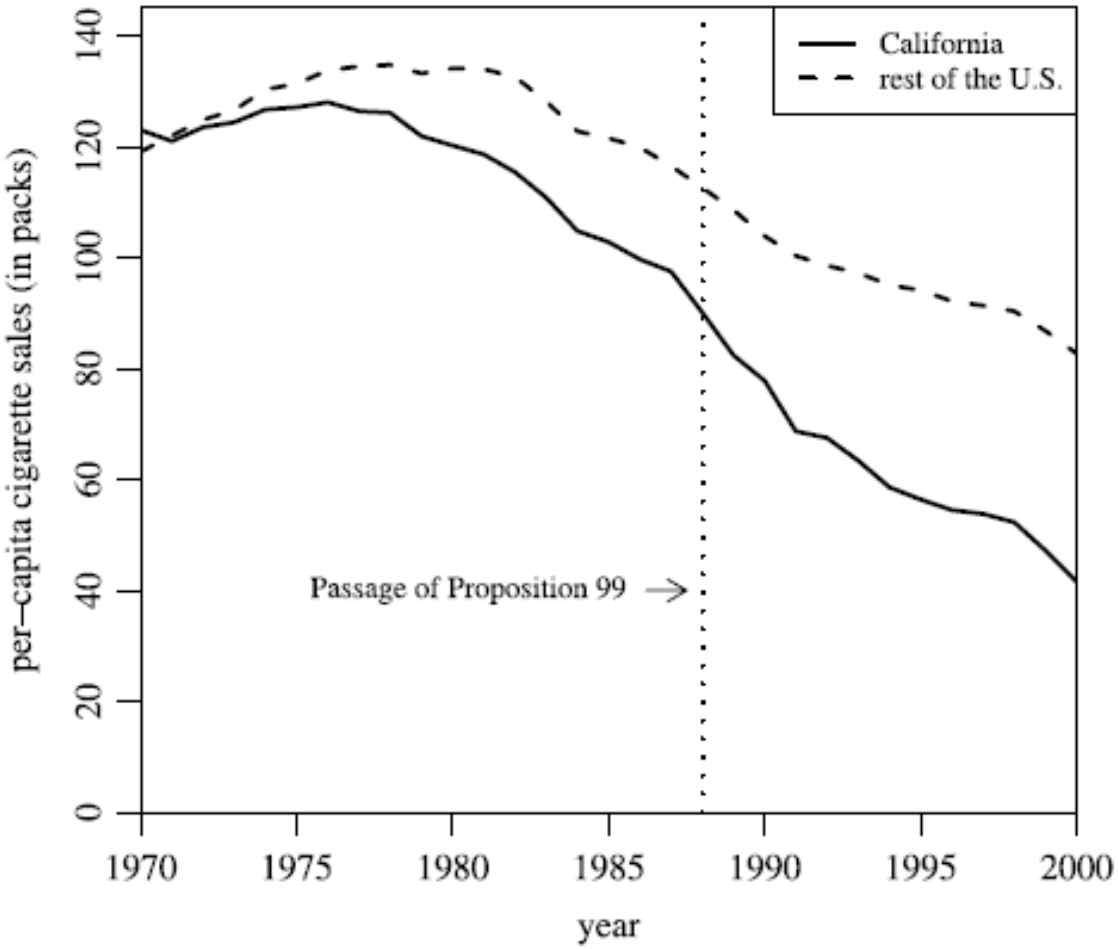


Figure: Trends in per-capita cigarette sales: California vs. the rest of the United States.

Cigarette Consumption: CA and SC-CA

Abadie, Diamond, Hainmuller, 2010

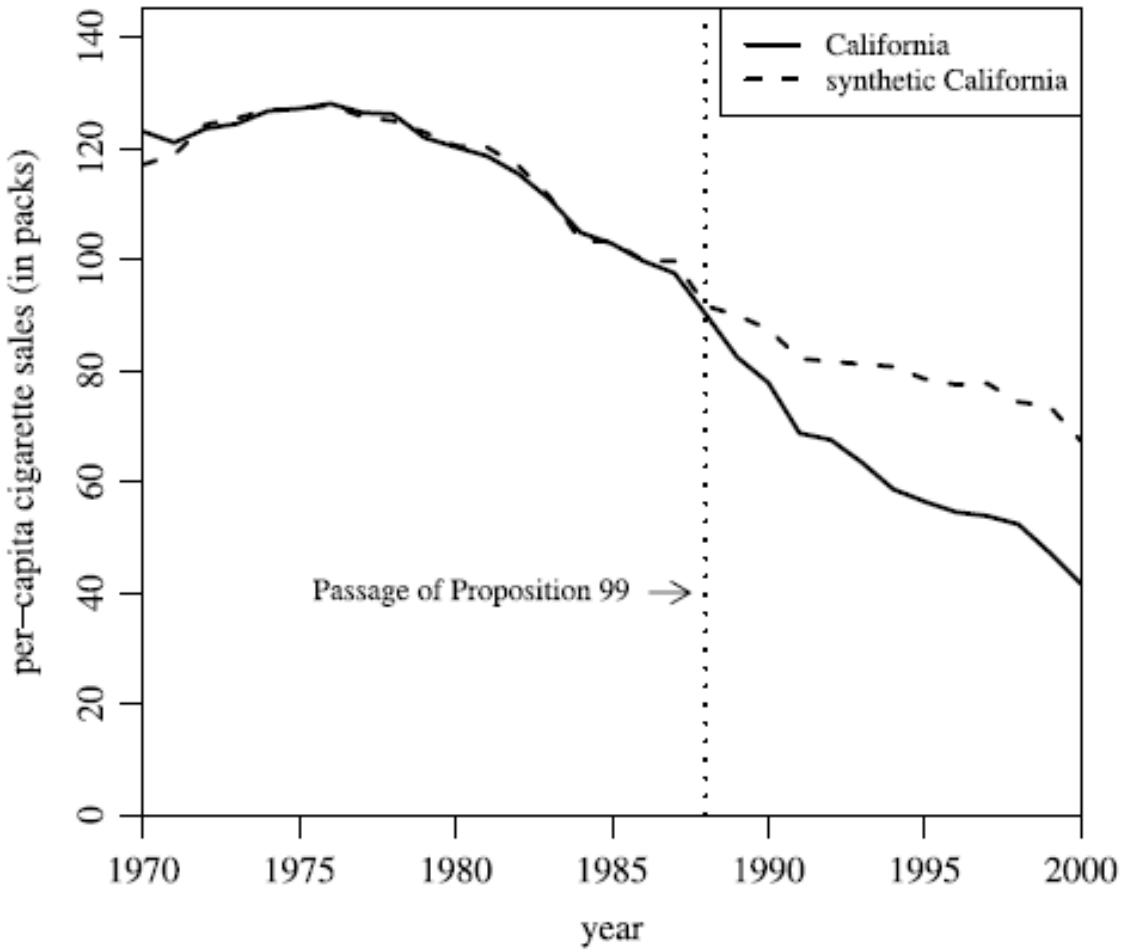


Figure: Trends in per-capita cigarette sales: California vs. synthetic California.

Synthetic Control (SC) Method

Abadie and Gardeazabal, 2003; Abadie, Diamond, Hainmuller, 2010

- What if the unobserved confounding is separable but has time-varying effects?
- Consider the following factor model (notation: $t = 1, \ldots, T$):

$$Y_{it}(0) = \mu_i' \lambda_t + \delta_t + X_i' \beta + \epsilon_{it}, \tag{11}$$

where μ_i: r-vector of unobserved confounders; λ_t: corresponding time-varying coefficients; X_i: p-vector of observed covariates.

- Model (11) generalizes the usual fixed-effects model for DID, where $\mu_i' \lambda_t$ is replaced by α_i, known as the interactive fixed effects model (Bai, 2003) in econ, known as latent factor model in statistics

SC: Formal Setup

- Suppose there are $J + 1$ units across periods $t = 1, \ldots, T$
- Further suppose the first unit is exposed to the intervention only after period T_0 $(1 \leq T_0 < T)$
- Potential outcome $Y_{it}(0)$ given by model (11), and $Y_{1t}(1) = Y_{1t}(0) + \tau_{1t} Z_{1t}$
- Interested in estimating the set of ATT estimands $\{\tau_{1,T_0+1}, \ldots, \tau_{1T}\}$, where

$$\tau_{1t} = Y_{1t}(1) - Y_{1t}(0) = Y_{1t} - Y_{1t}(0) \quad \text{for} \quad t \geq T_0 + 1$$

- The central task of SC is to estimate the unobserved $Y_{1t}(0)$ by a convex combination of the observed outcomes for the control units

SC: Theory

- Let $W = (w_2, \ldots, w_{J+1})'$ with $w_j \geq 0$ and $\sum_{j=2}^{J+1} w_j = 1$. Each choice of W represents a potential synthetic control
- Suppose we can choose W^* such that the pre-treatment covariates and outcomes for the treated unit are reproduced

$$\sum_{j=2}^{J+1} w_j^* X_j = X_1, \sum_{j=2}^{J+1} w_j^* Y_{j1} = Y_{11}, \ldots, \sum_{j=2}^{J+1} w_j^* Y_{jT_0} = Y_{1T_0}$$

- Assuming factor model (11) and fairly standard conditions, could show $Y_{1t}(0) - \sum_{j=2}^{J+1} w_j^* Y_{jt} \approx 0$ if # of pre-treatment periods is large relative to the residual variance
- An approximately unbiased estimator of τ_{1t} is

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

SC: Choosing W^*

- Let $Z_1 = (X_1', Y_{11}, \ldots, Y_{iT_0})'$ be the vector of pre-treatment covariates and outcomes for the treated unit
- Similarly define Z_0 be a matrix with each column being the same variables for each control unit
- The weights W^* is chosen to minimize some discrepancy metric $||Z_1 - Z_0 W||$, subject to the sum-to-unity constraint
- A typical metric is given by

$$||Z_1 - Z_0 W||_V = \sqrt{(Z_1 - Z_0 W)' V (Z_1 - Z_0 W)},$$

  where V is some symmetric and positive semidefinite matrix
- Usually choose V to be diagonal so that each element represents the relative importance of each variable

SC: Choosing Optimal V

- The inferential procedure is valid for any V, and V affects the mean square error (MSE) of the estimator
- An optimal choice of V assigns importance weights to discrepancies in $Z_1 - Z_0 W$ and minimizes MSE of the SC estimator
- A typical strategy is to choose V among positive definite and diagonal matrices s.t. the mean squared prediction error (MSPE) of the pre-treatment outcomes is minimized
- Let U_1, U_0 be the remainders of Z_1, Z_0 after removing X, choose (up to normalization constant)

$$V^* = \arg\min_V \{U_1 - U_0 W^*(V)\}' \{U_1 - U_o W^*(V)\}$$

- An iterative process – end goal is to choose $W^*(V^*)$ and V^* to minimize $||Z_1 - Z_0 W||_V$

SC: Permutation Inference

Abadie et al, 2010

- The distribution of a test statistic is computed under each random permutations of the unit's assignment
- Then assess how extreme the observed statistic is relative to the permutational distribution
- Requires a large number of control units, typical in comparative case studies
- Estimation and inference may be extended when more than one unit receives treatment (an open question, what if each treated unit receives treatment in a staggered fashion?)

Recent developments

- Synthetic control is very popular, owing to its simplicity and intuition
- But the setting SC targets at is inherently challenging: only one or at most a few treated units - lack of information (e.g. overlap)
- Recent developments
  - Relax convexity constraint of SC (Doudchenko and Imbens, 2016)
  - Combine SC and DID: synthetic DID (Arkhangelsky, et al. 2018)
  - Matrix completion method (Athey et al. 2017)
  - Augmented SC (Ben-Michael, Feller, Rothstein, 2018)

Synthetic DID: Setup

Arkhangelsky et al. 2021

- Another recent work looks to combine the beneficial features of both synthetic controls and difference in difference estimators – synthetic difference in difference (SDID)
- To provide intuition for the SDID estimator, it will help to write all estimators in the same manner
- Assume we have N units and T time periods
- The first N_{co} units never receive treatment
  - $Z_{it} = 0$ for $i = 1, \ldots, N_{co}$ and all t
- The remaining N_{tr} units are all exposed after time T_{pre}
- $T_{post} = T - T_{pre}$ time periods post treatment

Synthetic DID: A General Weighted Form

- The SC, DID, and SDID estimators can all be written as

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

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 ω_i and how the adjusted outcomes δ_i are defined

Synthetic DID: DID expression

- DID uses constant weights $\omega_i = N_{co}^{-1}$
- DID adjusted outcomes are given by

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

- Easy to see that if we only have two time periods with treatment given in the second period, this estimator simplifies to the DID estimator we saw earlier
  - Unweighted difference of differences

Synthetic DID: SC expression

- SC on the other hand uses weighting instead of differencing to estimate the average potential outcome in the absence of treatment

- SC uses weights that balance pre-treatment outcomes between treated and control units

$$\widehat{\omega} = \arg\min_{\omega \in \Omega} \left\{ \sum_{t=1}^{T_{pre}} \left(\sum_{i=1}^{N_{co}} \omega_i Y_{it} - \frac{1}{N_{tr}} \sum_{i=N_{co}+1}^{N} Y_{it} \right)^2 \right\}$$

subject to the constraint that

$$\sum_{i=1}^{N_{co}} \omega_i = 1, \quad \omega_i = \frac{1}{N_{tr}} \text{ for all } i = N_{co}+1, \ldots, N$$

Synthetic DID: SC expression

- Unlike DID, SC does not use differencing in the creation of adjusted outcomes
- SC uses

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

- Only really uses the pre-treatment period for estimation of weights that balance the two groups
- We can see that the estimate of the causal effect is just the difference between the average treated outcomes (post-treatment) and the weighted post-treatment outcome of the controls

Synthetic DID: Combining DID and SC

- SDID combines both of these ideas
- First construct unit-specific weights using pre-treatment periods via

$$(\widehat{\omega}_0, \widehat{\omega}) = \arg\min_{\omega_0, \omega} \left\{ \sum_{t=1}^{T_{pre}} \left(\omega_0 + \sum_{i=1}^{N_{co}} \omega_i Y_{it} - \frac{1}{N_{tr}} \sum_{i=N_{co}+1}^{N} Y_{it} \right)^2 + \zeta^2 ||\omega||_2^2 \right\}$$

subject to the constraint that

$$\sum_{i=1}^{N_{co}} \omega_i = 1, \quad \omega_i = \frac{1}{N_{tr}} \text{ for all } i = N_{co} + 1, \ldots, N$$

Synthetic DID: Penalty and Intercept

- These look really similar to the SC weights with two exceptions
- Add a penalty to the weights
  - Spreads weight across units
  - No units with very large weight / reduce variability
- Added an intercept term ω_0
  - Easier to make pre-treatment trends parallel instead of exactly equal
  - This will be allowed, because we will use differencing as in the DID estimator which will remove this constant difference

Synthetic DID: Analytical Form

- SDID also constructs time-specific weights, $\widehat{\lambda}_t$ aimed at balancing pre-exposure time periods and post-exposure ones

$$(\widehat{\lambda}_0, \widehat{\lambda}) = \arg\min_{\lambda_0, \lambda} \left\{ \sum_{i=1}^{N_{co}} \left(\lambda_0 + \sum_{t=1}^{T_{pre}} \lambda_t Y_{it} - \frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} Y_{it} \right)^2 \right\}$$

subject to the constraint that

$$\sum_{t=1}^{T_{pre}} \lambda_t = 1, \quad \lambda_t = \frac{1}{T_{post}} \text{ for all } t = T_{pre} + 1, \ldots, T$$

Synthetic DID

- SDID then uses the following adjusted outcome

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

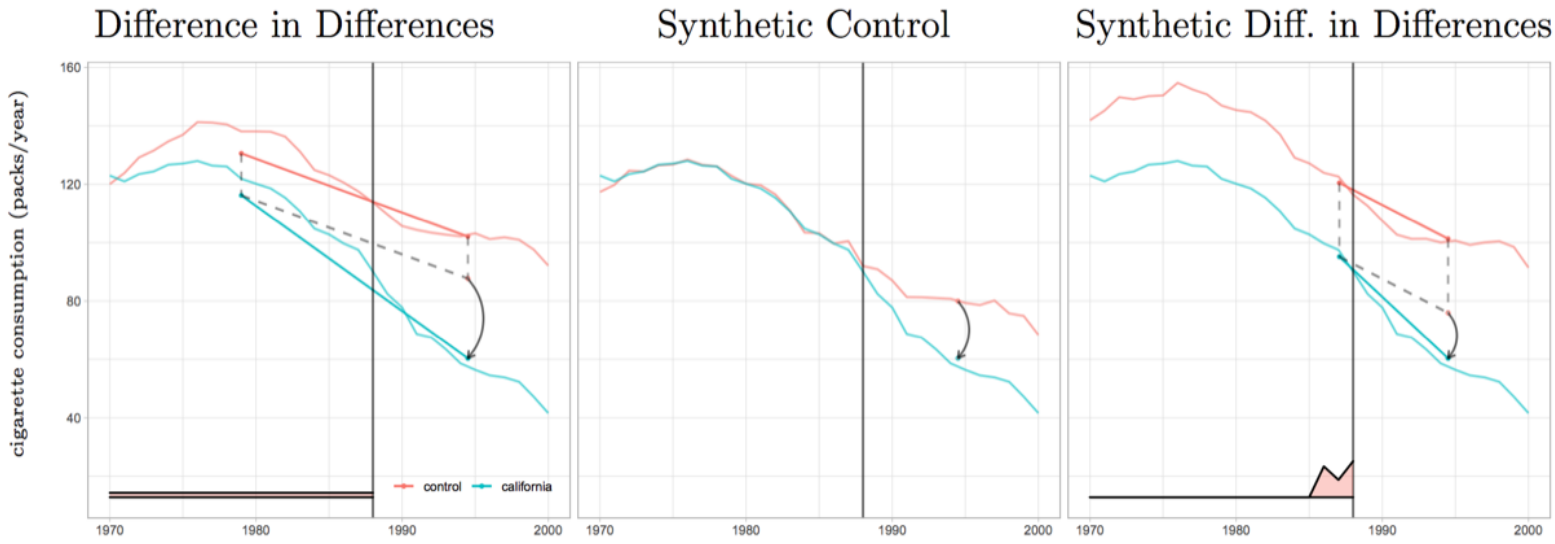
- Very similar to the DID adjusted outcome except the pre-treatment periods are weighted unequally
- More flexible than DID, which assigns equal weight to each pre-treatment period
  - A weighted version of parallel trends
  - More recent time periods might be more reasonable than earlier ones

Synthetic DID: Doubly-robust

- Clearly this estimator is a generalization of SC and DID
  - More flexible
- Authors show this estimator is doubly robust in that it will provide consistent estimates if the assumptions underlying either DID or SC are satisifed
- When DID performs well and SC doesn't, they find SDID performs similarly to or better than DID
- When SC performs well and DID doesn't, SDID matches or improves upon performance of SC

Synthetic DID: Illustration

- Here is an illustration of how each method estimates causal effects in the California smoking cessation program in Arkhangelsky et al. (2021)
- Parallel trends looks suspect, and DID overestimates the causal effect
  - SDID has smallest estimate overall



Matrix Completion

Athey et al. (2021, JASA)

- A lot of ideas for panel data methods were generalized in Athey et al. (2021)
- The key insight is that panel data causal inference problems are essentially matrix completion problems
- If we're interested in the matrix of potential outcomes under control, $Y_{it}(0)$, we only get to observe some of these values
- Fill in this matrix using matrix completion methods, which have their own rich literature

Matrix completion

- We can write our potential outcome and treatment matrices as

$$\boldsymbol{Y}(0) = \begin{pmatrix} Y_{11} & Y_{12} & \cdots & ? & ? \\ Y_{21} & ? & \cdots & ? & ? \\ \vdots & \vdots & \ddots & & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nT-1} & Y_{nT} \end{pmatrix}$$

$$\boldsymbol{Z} = \begin{pmatrix} 1 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & 1 & \cdots & 1 & 1 \end{pmatrix}$$

- Our aim is to impute the missing data (question marks in the matrix)

- From now on, simply use $\boldsymbol{Y}$ to refer to $\boldsymbol{Y}(0)$

Matrix completion

- One can assume that the data are defined as

$$\boldsymbol{Y} = \boldsymbol{L}^* + \boldsymbol{\epsilon}$$

  and we want to learn $\boldsymbol{L}^*$

- Let $\mathcal{O}$ represent the set of indices (i, t) where we observe the outcome, i.e. control units/times
  - $\mathcal{M}$ is for treated units/times

- For any matrix, $\boldsymbol{Z}$ define $P_{\mathcal{O}}(\boldsymbol{Z})$ and $P_{\mathcal{O}}^{\perp}(\boldsymbol{Z})$ as

$$P_{\mathcal{O}}(\boldsymbol{Z})_{it} = \begin{cases} Z_{it} & (i,t) \in \mathcal{O} \\ 0 & (i,t) \notin \mathcal{O} \end{cases} \qquad P_{\mathcal{O}}^{\perp}(\boldsymbol{Z})_{it} = \begin{cases} 0 & (i,t) \in \mathcal{O} \\ Z_{it} & (i,t) \notin \mathcal{O} \end{cases}$$

Matrix completion

- Our first thought might then be to minimize the squared error

$$\min_{\boldsymbol{L}} \frac{1}{|O|} \sum_{(i,t)\in O} (Y_{it} - L_{it})^2 = \min_{\boldsymbol{L}} \frac{1}{|O|} ||P_O(\boldsymbol{Y} - \boldsymbol{L})||_F$$

- But this doesn't help us solve the problem
  - Trivially we get $L_{it} = Y_{it}$ for $(i,t) \in O$
  - Doesn't help us for the important points $(i,t) \in \mathcal{M}$

- To fix these issues, we can regularize the matrix
  - Individual values
  - The rank
  - Magnitude of eigenvalues

Matrix completion

- The proposed estimator is given by $\boldsymbol{L}^* = \widehat{\boldsymbol{L}} + \widehat{\Gamma}\mathbf{1}_T^T + \mathbf{1}_N\widehat{\Delta}^T$ where the unknown values are found via

$$\arg\min_{\boldsymbol{L},\boldsymbol{\Gamma},\boldsymbol{\Delta}} \left\{ \frac{1}{|O|} ||P_O(\boldsymbol{Y} - \widehat{\boldsymbol{L}} - \widehat{\Gamma}\mathbf{1}_T^T - \mathbf{1}_N\widehat{\Delta}^T)||_F^2 + \lambda||\boldsymbol{L}||_* \right\}$$

  where $||\boldsymbol{L}||_*$ is the nuclear norm of $\boldsymbol{L}$ and is given by

$$||\boldsymbol{L}||_* = \sum_i \sigma_i(\boldsymbol{L})$$

- where $\sigma_i(\boldsymbol{L})$ is the i^{th} eigenvalue of $\boldsymbol{L}$

Matrix completion

- Note that they don't regularize the individual or time-specific fixed effects
- Unlike standard uses of matrix completion, we typically have a fair amount of observed data
  - Have enough data to estimate these parameters
- Effectively this estimates a fixed effects model for the potential outcomes, but allows for more complex structure through $\boldsymbol{L}$
  - Place structure on $\boldsymbol{L}$ through the penalty

Matrix completion

- Interestingly, they showed that this generalizes many of the commonly used estimators in the panel data literature
- Estimators such as difference in differences, synthetic controls, and others can be framed in this way
  - Typically without the penalty
  - Usually with a constraint on $\boldsymbol{L}$ or the fixed effects
- Model can be extended to include covariates as well
- Inference is challenging in such a model, though the authors propose a resampling procedure
  - Unclear how well this works

Same Root Different Leaves

Shen et al. 2024, Econometrica

- Key ideas for panel data methods are crystallized in a beautiful paper by Shen et al. (2024): *Same Root Different Leaves: Time Series and Cross-Sectional Methods in Panel Data*
- Slides by Dennis Shen **here (click)**

Additional notes on DID with multiple periods

DID with multiple-periods: setup

Callaway and Sant'Anna, 2021

- Consider the case with T time periods
- $D_{i,t} = 1$ unit i is treated in period t and 0 otherwise.
- Assumption 1 (Staggered Treatment Adoption): If $D_{t-1} = 1$ then $D_t = 1$
- Under staggered treatment adoption, we define G_g as a dummy that is equal to 1 if a unit is first treated in period g: $G_{i,g} = 1(G_i = g)$
- Estimand: **Group-time average treatment effect** $ATT(g,t) = E[Y_t(g) - Y_t(0)|G_g = 1]$

Identification assumptions

- Assumption 2 (A2: Limited Treatment Anticipation) Allow for a fixed period δ of anticipation. ($\delta = 1$: one period of anticipation)
- Two alternative assumptions for the definition of control group:
  - A3: *Conditional Parallel Trends* based on a "Never-Treated" Group

    $$E[Y_t(0) - Y_{t-1}(0)|X, G_g = 1] = E[Y_t(0) - Y_{t-1}(0)|X, C = 1]$$

    where $C = 1$ if a unit never participants in the treatment
  - A3': *Conditional Parallel Trends* based on "Non-Yet-Treated" Groups

    $$E[Y_t(0) - Y_{t-1}(0)|X, G_g = 1] = E[Y_t(0) - Y_{t-1}(0)|X, D_s = 0, G_g = 0]$$

- A4: Overlap. $\Pr(G_g = 1) > \epsilon$

Non-parametric identification

- Two main challenges when generalizing DiD to multiple groups and time periods
  1. Most appropriate time reference? Most recent time period when untreated potential outcomes are observed for group g
  2. Most appropriate comparison group? comes down to choosing between A3 (never-treated) and A3' (not-yet-treated).
- When pre-treatment covariates play no role, under A3:

$$ATT^{nev}_{unc}(g,t;\delta) = E[Y_t - Y_{g-\delta-1}|G_g = 1] - E[Y_t - Y_{g-\delta-1}|C = 1]$$

Under A3':

$$ATT^{ny}_{unc}(g,t;\delta) = E[Y_t - Y_{g-\delta-1}|G_g = 1] - E[Y_t - Y_{g-\delta-1}|D_{t+\delta} = 0]$$

Summarizing group-time ATEs

- The general form for parameters to aggregate the individual $ATT(g,t)$:

$$\theta = \sum_{g \in \mathcal{G}} \sum_{t=2}^{\mathcal{T}} w(g,t) \cdot ATT(g,t)$$

- One can use these parameters to answer three important questions:
  - How do average treatment effects vary with length of exposure to treatment?
  - How do average treatment effects vary across groups?
  - What is the cumulative average treatment effect of the policy across all groups until time $\tilde{t}$?

Estimation and inference

- Outcome Regression (OR), Inverse Probability Weighting (IPW), and Doubly Robust (DR) are all valid means for estimating average treatment effect of the treated
  - OR: requires researchers to correctly model the outcome evolution of the comparison group (connected to the conditional parallel trends assumption).
  - IPW: requires researchers to correctly model the conditional probability of unit i being in group g given the covariates X.
  - Doubly-Robust: cobmine OR and IPW, more flexible

Additional notes on bracketing between DID and LDV

Additional notes on bracketing between DID and LDV

Ding and Li (2019)

- If the true outcome model is $\mathbb{E}(Y_{t+1} \mid G, Y_t) = \alpha + \tau G + \beta Y_t$, then $\tau = \tau_{\text{ATT}}$
- Two versions of LDV
  - Fit OLS $\hat{\mathbb{E}}(Y_{t+1} \mid G = 0, Y_t = y) = \hat{\alpha} + \hat{\beta} Y_t$ to control units:

$$\hat{\tau}_{\text{LDV}} = (\bar{Y}_{1,t+1} - \bar{Y}_{0,t+1}) - \hat{\beta}(\bar{Y}_{1,t} - \bar{Y}_{0,t})$$

  - Fit OLS $\hat{\mathbb{E}}(Y_{t+1} \mid G, Y_t) = \hat{\alpha} + \hat{\tau}'_{\text{LDV}} G + \hat{\beta}' Y_t$ to all units:

$$\hat{\tau}'_{\text{LDV}} = (\bar{Y}_{1,t+1} - \bar{Y}_{0,t+1}) - \hat{\beta}'(\bar{Y}_{1,t} - \bar{Y}_{0,t})$$

- Compared to DID

$$\hat{\tau}_{\text{DID}} = (\bar{Y}_{1,t+1} - \bar{Y}_{0,t+1}) - (\bar{Y}_{1,t} - \bar{Y}_{0,t})$$

- DID and LDV estimators are identical if $\hat{\beta} = 1$ or $\hat{\beta}' = 1$

Interpreting the bracketing relationship under linear models

- Consider the case with $\hat{\beta}$ or $\hat{\beta}'$ smaller than 1
- The sign of $\hat{\tau}_{\text{DID}} - \hat{\tau}_{\text{LDV}}$ or $\hat{\tau}_{\text{DID}} - \hat{\tau}'_{\text{LDV}}$ depends on the sign of $\bar{Y}_{1,t} - \bar{Y}_{0,t}$
- Treatment group has smaller Y_t on average $\Longrightarrow \hat{\tau}_{\text{DID}} > \hat{\tau}_{\text{LDV}}$
- Treatment group has larger Y_t on average $\Longrightarrow \hat{\tau}_{\text{DID}} < \hat{\tau}_{\text{LDV}}$
- How much $\hat{\beta}$ or $\hat{\beta}'$ deviates from 1
  $\Longrightarrow$ how different the DID and LDV estimates are
- Numeric result, no stochastic assumption, more general than Angrist and Pischke (2009)

Additional notes on Bayesian approach to panel data and time series modeling

Bayesian approaches and time series modeling

- All of the aforementioned approaches take a frequentist approach to inference
- Recent ideas have exploited complex Bayesian time series models to estimate time-varying treatment effects
  - Inference is automatic
  - Easy to account for difficult trends such as seasonality
  - Easy to allow effects to vary over time
- These ideas are generally useful with a large amount of pre-treatment data
- More information is provided in the additional notes at the end

Bayesian approaches and time series modeling

- Our data is simply Y_t for $t = 1, \ldots, T$
- Some treatment is imposed after time T_{pre} where T_{pre} is relatively large
- Want to estimate the impact of this treatment at each time point after T_{pre}
- We may (though not necessarily) have some contemporaneous time series X_t for $t = 1, \ldots, T$
  - Unaffected by treatment
  - Could be control units, or other variables

Bayesian approaches and time series modeling

- Idea is simple. Model Y_t for $t \leq T_{pre}$ and use it to forecast what would have happened in the absence of the treatment
- Difference between the observed outcome and this forecast is the causal effect
- Uncertainty automatically accounted for by the posterior distribution of these forecasts
- A real benefit of this approach is it allows us to use complex time-series models that have been shown to work in a range of settings
- Approach relies on a stationarity assumption for the potential outcome time series

Bayesian approaches and time series modeling

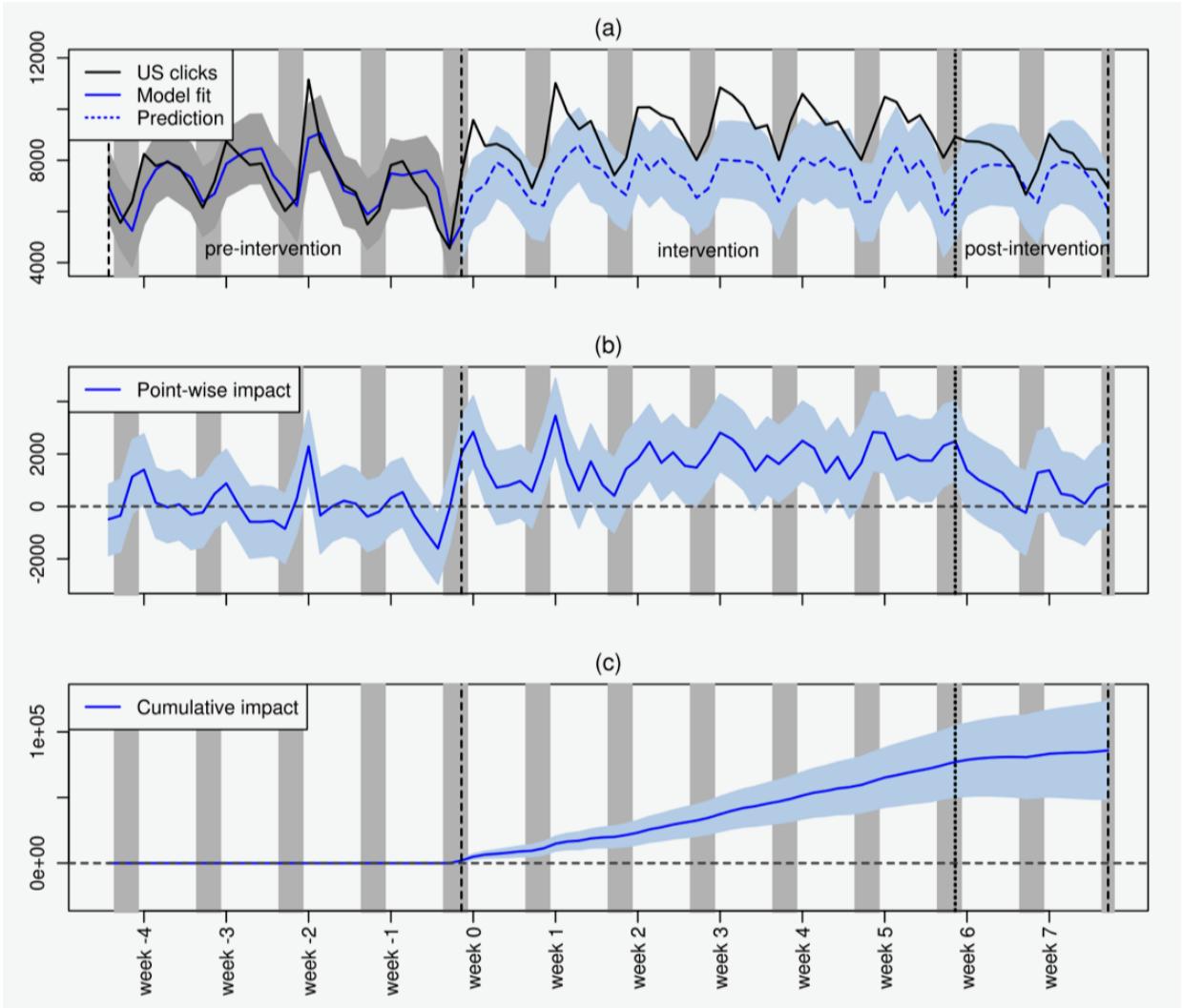
- One nice model is a Bayesian structural time series model

$$Y_t = Z_t^T \alpha_t + \epsilon_t$$

$$\alpha_{t+1} = T_r \alpha_t + R_t \eta_t$$

- α_t are some unobserved states at time t and the first equation describes how the observed outcome relates to these
- The second equation shows how the states develop over time
- Difficult to see in this general formulation, but this allows for lots of flexibility
  - Seasonal effects
  - Nonlinear time trends
  - Contemporaneous covariates or control outcomes

Bayesian approaches and time series modeling



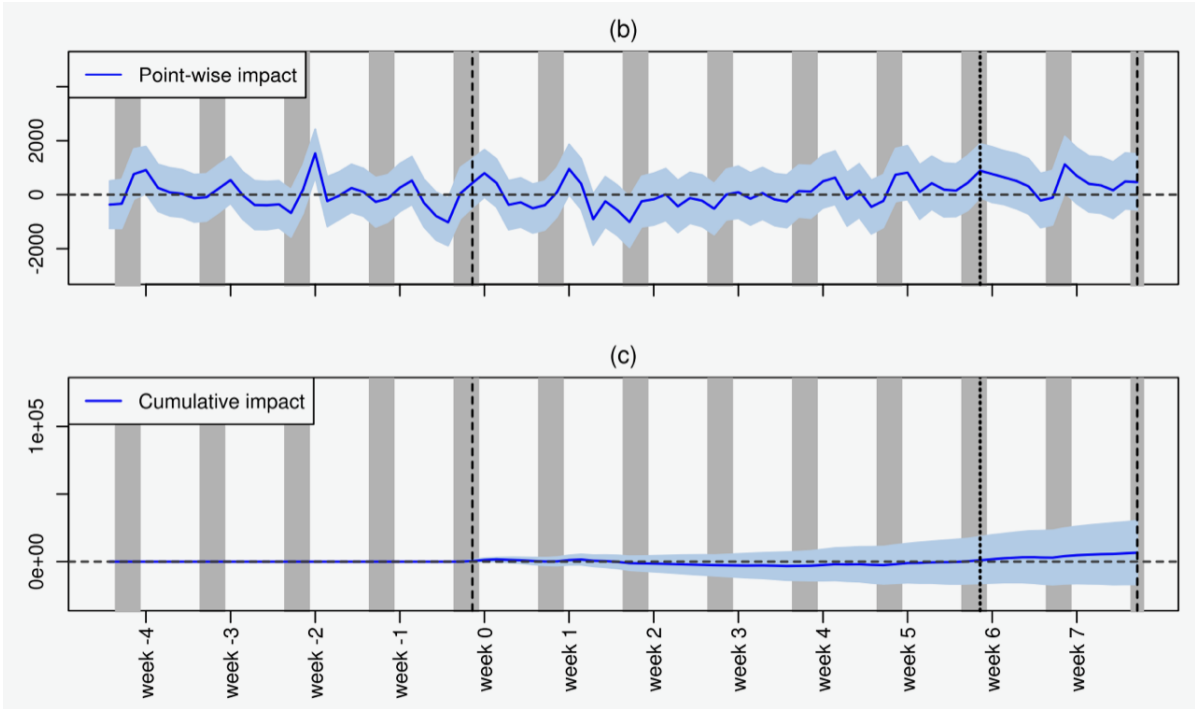
Brodersen, Kay H., et al. "Inferring causal impact using Bayesian structural time-series models." The Annals of Applied Statistics 9.1 (2015): 247-274.

Bayesian approaches and time series modeling

- The previous example had certain difficult aspects
- Some clear, strong seasonal effects
  - How do the previous estimators we've studied address this?
- In this example, the effect was relatively constant, but the approach automatically estimates differing effects over time
- Can be extended to more difficult settings of heterogeneous treatment effects with multiple units (Antonelli and Beck, 2020) or interference (Manchetti and Bojinov, 2020)

Placebo tests and lots of control periods

- We've seen previously that one form of sensitivity analysis is to estimate the causal effect of the treatment on an outcome that can't be affected by treatment (negative control)
- We can do that in panel data settings as well



Brodersen, Kay H., et al. "Inferring causal impact using Bayesian structural time-series models." The Annals of Applied Statistics 9.1 (2015): 247-274.

Placebo tests and lots of control periods

- We don't always have an outcome available to us for this test
- In panel data settings with lots of pre-treatment data we can run tests on the pre-treatment data when we know there should be no effect as the policy hasn't been implemented yet
  - Estimate causal effect at earlier time point
- Other things testable in pre-treatment periods as well
  - Parellel trends
  - Stationarity assumptions

References

- Card, D. and Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. American Economic Review 82, 772–793.
- Card, D. and Krueger, A. B. (2000). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Reply. American Economic Review, 90(5), 1397–1420.
- Hauer, E. (1997). Observational before-after studies in road safety: Estimating the effect of highway and traffic engineering measures on road safety, Oxford, OX, U.K., Pergamon: Emerald Group Publishing Limited.
- Angrist, J. D. and Pischke, J.-S. (2009), Mostly Harmless Econometrics: an Empiricists Companion, Princeton, NJ: Princeton University Press.
- Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. Econometrica. 74(2), 431–497.
- Abadie, A. (2005). Semiparametric difference-in-differences estimators. Review of Economic Studies 72, 1–19.
- Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. The Review of Economic Studies, 64, 605–654.
- Arkhangelsky D, Athey S, Hirshberg DA, Imbens GW, Wager S. Synthetic difference-in-differences. American Economic Review. (2021). 111(12):4088-118.

References

- Menchetti, Fiammetta, and Iavor Bojinov. Estimating causal effects in the presence of partial interference using multivariate Bayesian structural time series models. Harvard Business School, 2020.
- Antonelli, Joseph, and Brenden Beck. "Heterogeneous causal effects of neighborhood policing in New York City with staggered adoption of the policy." arXiv preprint arXiv:2006.07681 (2020).
- Callaway, B. and Sant'Anna, P. (2021). Difference-in-differences with multiple time periods. J Econometrics. 225(2), 200-230.
- Lechner, M. (2011). The Estimation of Causal Effects by Difference-in-Difference Methods. Foundations and Trends in Econometrics, 4, 165–224.
- Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. The Review of Economics and Statistics, 47–57.
- Rosenbaum, P. R. (2002). Observational studies. Springer, New York, NY.
- Bertrand, M., Duflo, E., Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?. The Quarterly Journal of Economics, 119(1), 249–275.

References

- Peng, D. and Li, F. (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis*. 27(4), 605-615.
- 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.
- 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.
- Ben-Michael, Eli, Avi Feller, and Jesse Rothstein. "The augmented synthetic control method." Journal of the American Statistical Association just-accepted (2021): 1-34.
- Athey, Susan, et al. "Matrix completion methods for causal panel data models." Journal of the American Statistical Association (2021): 1-15.
- Brodersen, Kay H., et al. "Inferring causal impact using Bayesian structural time-series models." The Annals of Applied Statistics 9.1 (2015): 247-274.