CodeChella Madrid 2024

Roadmap

Imputation in 2x2

Imputation with Staggered Treatment Timing

Modeling non-parallel trends

Synthetic Control

More General Factor Model Imputation

DID Estimator

Let's start with our standard DID estimator

$$\mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 1] - \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0]$$

DID Estimator

Let's start with our standard DID estimator

$$\mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 1] - \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0]$$

We can rewrite this in an odd-looking way

$$\mathbb{E}[Y_{i1} - (Y_{i0} + \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0]) \mid D_i = 1]$$

DID Estimator

Let's start with our standard DID estimator

$$\mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 1] - \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0]$$

We can rewrite this in an odd-looking way

$$\mathbb{E}[Y_{i1} - (Y_{i0} + \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0]) \mid D_i = 1]$$

- $\mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0]$ is the average change in outcome for control units
- Take a treated unit's initial value Y_{i0} and add to it this change in outcome

DID Estimator

Let's start with our standard DID estimator

$$\mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 1] - \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0]$$

We can rewrite this in an odd-looking way

$$\mathbb{E}\Big[Y_{i1}(1) - \underbrace{(Y_{i0} + \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0])}_{\hat{Y}_{i1}(0)} \mid D_i = 1\Big]$$

- $\mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0]$ is the average change in outcome for control units
- Take a treated unit's initial value Y_{i0} and add to it this change in outcome

Imputation

$$\mathbb{E}\left[Y_{i1}(1) - \underbrace{(Y_{i0} + \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0])}_{\hat{Y}_{i1}(0)} \mid D_i = 1\right]$$

I think this makes it clear what DID is doing:

- We can observe untreated PO in the pre-period, but not the post-period.
- We predict how outcome would change using the untreated group

Imputation

$$\mathbb{E}\left[Y_{i1}(1) - \underbrace{(Y_{i0} + \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0])}_{\hat{Y}_{i1}(0)} \mid D_i = 1 \right]$$

I think this makes it clear what DID is doing:

- We can observe untreated PO in the pre-period, but not the post-period.
- We predict how outcome would change using the untreated group

Identification relies on $\hat{Y}_{i1}(0)$ being an unbiased estimate of $Y_{i1}(0)$

- This is the parallel (counterfactual) trends assumption
- We are 'imputing' missing potential outcome (Imbens, RESTAT 2004)

Imputation with Covariates

The regression adjustment estimator is also an imputation estimator:

$$\mathbb{E}\left[Y_{i1}(1) - \underbrace{(Y_{i0} + \hat{\mu}_{D=0,t=1}(X_i) - \hat{\mu}_{D=0,t=0}(X_i))}_{\hat{Y}_{i1}(0)}, \mid D_i = 1 \right]$$

where $\hat{\mu}_{D=0,t}$ is our estimates of $Y_{it}(0)$ given X_i estimated with the untreated units only.

Imputation with Covariates

Similarly, the IPW estimator is an imputation estimator but you take a weighted average of

$$\mathbb{E}\left[Y_{i1}(1) - \underbrace{\left(Y_{i0} + \mathbb{E}\left[\frac{\hat{p}(X_i)}{\mathbb{P}[D_i = 1]} \left(Y_{i1} - Y_{i0} \right) \mid D_i = 0 \right] \right)}_{\hat{Y}_{i1}(0)} \mid D_i = 1 \right],$$

where $\hat{p}(X_i)$ is our estimated propensity score

Roadmap

Imputation in 2x2

Imputation with Staggered Treatment Timing

Modeling non-parallel trends

Synthetic Control

More General Factor Model Imputation

TWFE Model

Unit i at time t has outcome y_{it} given by the TWFE model with heterogeneous effects:

$$y_{it} = \mu_i + \eta_t + \tau_{it} d_{it} + \varepsilon_{it},$$

with $\mathbb{E}[\varepsilon_{it}] = 0$. η_t are common time shocks, μ_i are unit-specific time-invariant shocks, and d_{it} is a treatment dummy for being actively under treatment.

Units can vary in what year they get treated, denoted by g.

Researchers care about the overall ATT, $\mathbb{E}[\tau_{it} \mid d_{it} = 1]$, or some different average of τ_{it}.

TWFE Model

Treatment effect heterogeneity

τ_{it} in most settings is heterogeneous

Treatment effects may depend on when you start treatment

- *e.g.*, groups that benefit more from a policy implement it earlier

Treatment effects may depend on treatment duration (event study!)

- *e.g.*, policy doesn't affect everyone right away

Ignoring heterogeneous effects

Say we ignore τ_{it} and simplify our model to

$$y_{it} = \mu_i + \eta_t + \tau d_{it} + \underbrace{u_{it}}_{\varepsilon_{it} + (\tau_{it} - \tau) d_{it}}$$

The error term now contains the treatment effect heterogeneity.

- If there is any systematic heterogeneity like we described before, then the covariates (fixed-effects) are correlated with the error term and OLS estimates suffer from omitted variables bias.

Estimating overall average treatment effect

If we knew μ_i and η_t, then we could move terms around in our model:

$$y_{it} - \mu_i - \eta_t = \tau_{it} d_{it} + \varepsilon_{it}$$

Estimating overall average treatment effect

If we knew μ_i and η_t, then we could move terms around in our model:

$$y_{it} - \mu_i - \eta_t = \tau_{it} d_{it} + \varepsilon_{it}$$

If we regress $y_{it} - \mu_i - \eta_t$ on the dummy variable d_{it}, OLS algebra tells us that $\hat{\tau}$ will estimate a simple average of τ_{it} (plus noise):

$$\hat{\tau} = \frac{1}{N_{post}} \sum_{(i,t)} (\tau_{it} d_{it} + \varepsilon_{it}) = \text{ATT} + \text{noise},$$

where N_{post} is the number of post-treatment observations

Estimating overall average treatment effect

If we knew μ_i and η_t, then we could move terms around in our model:

$$y_{it} - \mu_i - \eta_t = \tau_{it} d_{it} + \varepsilon_{it}$$

If we regress $y_{it} - \mu_i - \eta_t$ on the dummy variable d_{it}, OLS algebra tells us that $\hat{\tau}$ will estimate a simple average of τ_{it} (plus noise):

$$\hat{\tau} = \frac{1}{N_{post}} \sum_{(i,t)} (\tau_{it} d_{it} + \varepsilon_{it}) = \text{ATT} + \text{noise},$$

where N_{post} is the number of post-treatment observations

Similarly, if we had a set of *mutually-exclusive* event-study dummy variables, d_{it}^{ℓ}, we would get averages of τ_{it} for (i,t) with $g_i - t = \ell$.

Too bad we don't know μ and η

Since we don't know μ and η, we have to estimate them. Using the FWL theorem our OLS estimate is equivalent to:

$$y_{it} - \hat{\mu}_i - \hat{\eta}_t = \tau \tilde{d}_{it} + u_{it},$$

where $\tilde{d}_{it}$ is the *residualized* treatment dummy.

Using OLS to estimate μ and η

$$y_{it} - \hat{\mu}_i - \hat{\eta}_t = \tau \tilde{d}_{it} + u_{it},$$

Under our model, the left-hand side equals:

$$y_{it} - \hat{\mu}_i - \hat{\eta}_t = \tau_{it} d_{it} + (\mu_i - \hat{\mu}_i) + (\eta_t - \hat{\eta}_t) + \varepsilon_{it}.$$

Using OLS to estimate μ and η

$$y_{it} - \hat{\mu}_i - \hat{\eta}_t = \tau \tilde{d}_{it} + u_{it},$$

Under our model, the left-hand side equals:

$$y_{it} - \hat{\mu}_i - \hat{\eta}_t = \tau_{it} d_{it} + \underbrace{(\mu_i - \hat{\mu}_i)}_{\text{mean } 0} + \underbrace{(\eta_t - \hat{\eta}_t)}_{\text{mean } 0} + \varepsilon_{it}.$$

We have a noisy estimate for τ_{it}, but we are averaging over observations. So what's the problem?

Using OLS to estimate μ and η

$$y_{it} - \hat{\mu}_i - \hat{\eta}_t = \tau \tilde{d}_{it} + u_{it},$$

Under our model, the left-hand side equals:
We have a noisy estimate for τ_{it}, but we are averaging over observations. So what's the problem?

All of the modern diff-in-diff problems (yes all of them) show different ways to interpret the same problem: $\tilde{d}_{it}$.

- Since we have residualized d_{it}, OLS no longer computes the simple average of τ_{it}'s.

Fixing the problem

$$y_{it} - \hat{\mu}_i - \hat{\eta}_t = \tau_{it}\tilde{d}_{it} + u_{it},$$

Okay, so why don't you just regress $y_{it} - \hat{\mu}_i - \hat{\eta}_t$ on d_{it}?

Fixing the problem

$$y_{it} - \hat{\mu}_i - \hat{\eta}_t = \tau_{it}\tilde{d}_{it} + u_{it},$$

Okay, so why don't you just regress $y_{it} - \hat{\mu}_i - \hat{\eta}_t$ on d_{it}? Great intuition!

Two-stage difference-in-differences

Proposed in concurrent work by Borusyak et. al. (2024) and Gardner (2021). I adopt the nomenclature of Gardner (2021).

Stage 1: Estimate μ_i and η_t using untreated/not-yet-treated observations ($d_{it} = 0$).

Two-stage difference-in-differences

Proposed in concurrent work by Borusyak et. al. (2024) and Gardner (2021). I adopt the nomenclature of Gardner (2021).

Stage 1: Estimate μ_i and η_t using untreated/not-yet-treated observations ($d_{it} = 0$).

- Don't include $d_{it} = 1$ since the treatment effects will bias estimates of the fixed effects.

Stage 2: Regress $Y_{it} - \underbrace{(\hat{\mu}_i + \hat{\eta}_t)}_{\hat{Y}_{it}(0)}$ on d_{it}.

Two-stage difference-in-differences

Proposed in concurrent work by Borusyak et. al. (2024) and Gardner (2021). I adopt the nomenclature of Gardner (2021).

Stage 1: Estimate μ_i and η_t using untreated/not-yet-treated observations ($d_{it} = 0$).

- Don't include $d_{it} = 1$ since the treatment effects will bias estimates of the fixed effects.

Stage 2: Regress $Y_{it} - \underbrace{(\hat{\mu}_i + \hat{\eta}_t)}_{\hat{Y}_{it}(0)}$ on d_{it}.

Inference is complicated because the outcome variable in the second-stage is a 'generated regressor'.

- Inference is taken care for you in `did2s` (or you can block bootstrap!)

Pseudocode

R Code:

```
library(fixest)
fs = feols(y ~ 0 | unit + time, data = df[df$d == 0, ])
df$y_resid = df$y - predict(fs, newdata = df)
ss = feols(y_resid ~ d, data = df)
```

Stata Code:

```
reg y i.unit i.time if d == 0
predict y0_hat, xb
gen y_resid = y - y0_hat
reg y0_hat i.d
```

Essential ingredients

1. Posit a model for $Y_{it}(0)$; this can include time-interacted covariates, linear time-trends, etc.
2. Estimate that model using observations that are not impacted by treatment ($d_{it} = 0$)
3. Take observed Y_{it} and subtract the (*out of sample*) predictions from the model.

Essential ingredients

1. Posit a model for $Y_{it}(0)$; this can include time-interacted covariates, linear time-trends, etc.
2. Estimate that model using observations that are not impacted by treatment ($d_{it} = 0$)
3. Take observed Y_{it} and subtract the (*out of sample*) predictions from the model.

Our parallel trends assumption comes in that our fitted model is an unbiased estimator for $Y_{it}(0)$ in the post-treatment periods *for the treated units*.

- E.g. our time fixed-effects are fit using the control group in the post-periods. Need these to be the same time effects as the treated group.

Stacking

I think the usefulness of the imputation procedure really shows itself in ‘more advanced treatment’ settings.

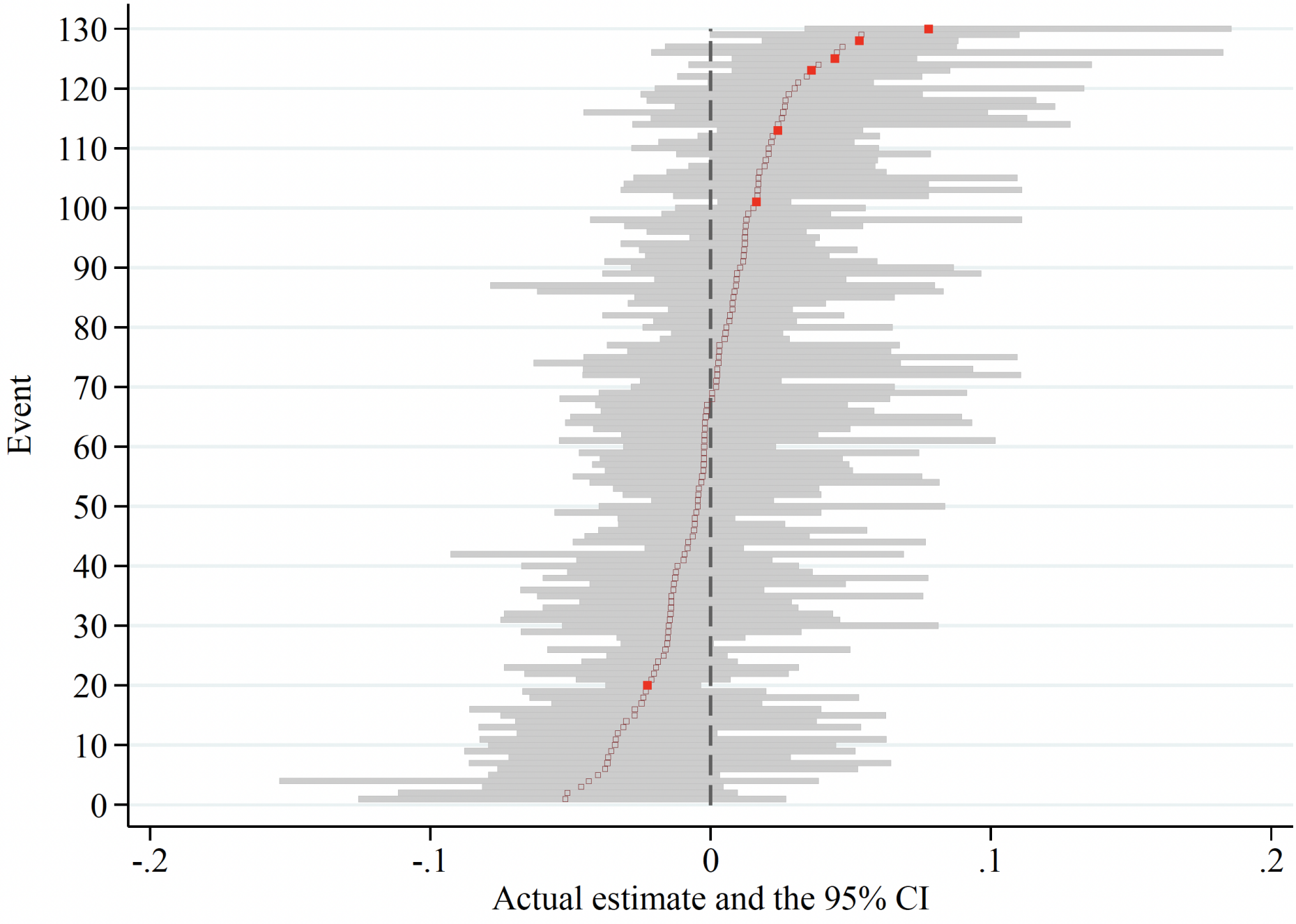
For example, Cengiz, Dube, Lindner, and Zipperer (QJE, 2019) think about state-level minimum wage changes.

- Problem: Not a binary absorbing treatment since states change minimum wage more than once.
- But there are few enough minimum wage changes, there should be a valid control group that can be used.

Minimum Wage Example Continued.

For each minimum wage change (an 'event'), the authors create a filtered dataset and run a DID on the subset

- Treated counties and a set of 'clean control states' that did not have a minimum wage change within 4 years of the 'event'
- Subset to ± 4 years from event


Event
130
120
110
100
90
80
70
60
50
40
30
20
10
0
-.2
-.1
0
.1
.2
Actual estimate and the 95% CI

Stacked overall estimate

They, then want to estimate an 'overall' estimate by pooling across events. To do so they 'stack' the dataset (`dplyr::bind_rows` or `append using` in a for loop) and then run the following event-study regression:

$$Y_{e,it} = \mu_{e,i} + \eta_{e,t} + \sum_{\ell} \tau^{\ell} d_{it}^{\ell} + u_{e,it},$$

where e denotes the event-dataset e.

- Note the treatment dummies are not interacted with e so we estimate an overall average effect

Stacking Pros and Cons

I think the stacking approach is popular because:

- Tailor the control group to a 'good' set of control units (e.g. bordering counties, matching counties with similar X, etc.).
  - $\rightarrow$ All the other packages use all the control units for every treated group by default (there are workarounds)
- You write an explicit model for $Y_{it}(0)$

But it comes with some costs:

- Annoying and error-prone; interacting a bunch of terms; bunch of for loops; make sure you filter dataset right; hard to write up and hard to read; etc.
- Identifies a convex-average of τ_{it} but the weights are non-standard and data-dependent (e.g. can change by adding more pre-periods)
- I don't know of any paper that discusses appropriate standard errors

Imputation and Stacking

In a new WIP (by me), make this procedure much simpler (using the imputation strategy):

First Stage:
For each event, e,

- Fit model of $Y_{it}(0)$ using the pre-treatment observations and the 'clean controls' for that event.
- For event-e treated units, impute their post-e outcomes.

Second Stage:
Regress $Y_{it} - \hat{Y}_{it}(0)$ on d_{it} or d_{it}^{ℓ}.

Imputation and Stacking

Advantages include:

- Identify ATT (not a weirdly weighted average)
- Inference is easy
- Easy to write-up: For each event, here is our model of $Y_{it}(0)$ and here are the control units we use

Imputation Estimators

Imputation estimators offer a few benefits:

- They allow you to transparently write down your model for the untreated potential outcome and estimate that model
- They naturally adapt to more complex settings beyond the two-way fixed effect model (next sections)

Roadmap

Imputation in 2x2

Imputation with Staggered Treatment Timing

Modeling non-parallel trends

Synthetic Control

More General Factor Model Imputation

Covariates and Conditional PTs

Let i denote workers and t denote the year. X_i be the amount of schooling the worker has.

A general model for wages might look like:

$$w_{it}(0) = \mu_i + \eta_t + g_t(X_i) + u_{it},$$

where $g_t(X_i)$ changes over time and causes non-parallel trends. Conditional parallel-trends implies that $\mathbb{E}[u_{it}] = 0$.

Comparing two units with the same $X_i = x$, then $g_t(x)$ is the same for those units receive the same time shocks $\eta_t + g_t(x)$.

- This is the essence of conditional parallel-trends.

Covariates and Conditional PTs

A common choice is $g_t(X_i) = X_i\beta_t$

- Trends are caused by changing returns to characteristic X_i, $\beta_t - \beta_{t-1}$.
- Linearity is the default for `did/DRDID` Sant'Anna and Zhao implementation

Covariates and Conditional PTs

A common choice is $g_t(X_i) = X_i\beta_t$

- Trends are caused by changing returns to characteristic X_i, $\beta_t - \beta_{t-1}$.
- Linearity is the default for `did`/`DRDID` Sant'Anna and Zhao implementation
- Linearity is strong. In a period, going from 8 to 9 years of schooling has the same increase in wages as 14 to 15 years. Maybe you think the profile is non-linear (e.g. quadratic)

Covariates and Conditional PTs

A common choice is $g_t(X_i) = X_i\beta_t$

- Trends are caused by changing returns to characteristic X_i, $\beta_t - \beta_{t-1}$.
- Linearity is the default for `did`/`DRDID` Sant'Anna and Zhao implementation
- Linearity is strong. In a period, going from 8 to 9 years of schooling has the same increase in wages as 14 to 15 years. Maybe you think the profile is non-linear (e.g. quadratic)

Could make more flexible with polynomials like $X_i\beta_{1,t} + X_i^2\beta_{2,t}$ or by binning X_i and interacting with time.

Covariate-specific trends and non-PTs

Let's go with the linear model:

$$w_{it}(0) = \mu_i + \eta_t + X_i\beta_t + u_{it},$$

Covariate-specific trends and non-PTs

Let's go with the linear model:

$$w_{it}(0) = \mu_i + \eta_t + X_i\beta_t + u_{it},$$

Say you don't include X_i in your regression, when would this create a problem?

$$\mathbb{E}[w_{i1} - w_{i0} \mid D_i = d] = (\eta_1 - \eta_0) + \mathbb{E}[X_i \mid D_i = d]\,(\beta_1 - \beta_0)$$

- In the linear case, if treated and control units don't have different average values of X_i, then unconditional PTs hold
- Conversely, if units enter treatment based on X_i, then non-parallel trends is induced.

Non-observed covariates

In the worker's wage example, we might suspect that something like computer-skill might be unobservable and have a time-varying impact.

That is both X_i and β_t are unobserved!

- If the job-training program attracts people with more computer-skills (e.g. excel workshop), then PTs does not hold.
- We can not 'compare two individuals with the same value of X_i' since we do not observe them.

(Linear) Factor Models

We've just walked through the motivation for a 'factor model', sometimes called 'interactive-fixed effects model'.

There are a set of ρ unit characteristics $\lambda_{i,r}$ and a set of corresponding time-specific shocks $f_{t,r}$ whose impact on Y is given by:

$$Y_{it} = \sum_{r=1}^{\rho} \lambda_{i,r} f_{t,r} + \varepsilon_{it}$$

- Same logic as before, λ_i measures the unit's characteristic and $f_{t,r}$ is the shock to returns of the characteristic.

(Linear) Factor Models

We've just walked through the motivation for a 'factor model', sometimes called 'interactive-fixed effects model'.

There are a set of ρ unit characteristics $\lambda_{i,r}$ and a set of corresponding time-specific shocks $f_{t,r}$ whose impact on Y is given by:

$$Y_{it} = \sum_{r=1}^{\rho} \lambda_{i,r} f_{t,r} + \varepsilon_{it}$$

- Same logic as before, λ_i measures the unit's characteristic and $f_{t,r}$ is the shock to returns of the characteristic.
- This nests the two-way fixed effects model $f_t = 1$ implies unit fixed-effects, $\lambda_i = 1$ implies time fixed-effects

(Linear) Factor Models

$$Y_{it} = \sum_{r=1}^{\rho} \lambda_{i,r} f_{t,r} + \varepsilon_{it}$$

Let's give an example using county-level aggregate employment.

- λ_i might consist of (i) manufacturing share and (ii) share of college-educated
- In each period, shocks to the national economy change manufacturing demand and (ii) technological change drives returns to college degree

(Linear) Factor Models

$$Y_{it} = \sum_{r=1}^{\rho} \lambda_{i,r} f_{t,r} + \varepsilon_{it}$$

Let's give an example using county-level aggregate employment.

- λ_i might consist of (i) manufacturing share and (ii) share of college-educated
- In each period, shocks to the national economy change manufacturing demand and (ii) technological change drives returns to college degree

We might have ideas of what are the primary characteristics are, but we might not have data on it. If you do, then we are back in X_i land.

Imputation and (Linear) Factor Models

We have a more-general model for $Y_{it}(0)$ now that allows some forms of non-parallel-trends:

$$Y_{it}(0) = \mu_i + \eta_t + \sum_{r=1}^{\rho} \lambda_{i,r} f_{t,r} + \varepsilon_{it}$$

Can we estimate this and use our imputation procedure:

- The short-answer is yes. There is a bunch of different approaches but they're not as simple as TWFE.
- The factor model is much more data-hungry than fixed effects, usually requiring both a large number of units *and* a large number of time-periods.

Roadmap

Imputation in 2x2

Imputation with Staggered Treatment Timing

Modeling non-parallel trends

Synthetic Control

More General Factor Model Imputation

Synthetic Control

The standard synthetic control method considers a single treated unit (country, state, firm, etc.). We will let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT})'$ be the vector of outcomes for unit i. The treated unit is $i = 0$.

Synthetic Control

The standard synthetic control method considers a single treated unit (country, state, firm, etc.). We will let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT})'$ be the vector of outcomes for unit i. The treated unit is $i = 0$.

Synthetic control imputes $\boldsymbol{Y}_0(0)$ for the treated unit using a weighted average of the control units:

$$\hat{\boldsymbol{Y}}_0(0) = \sum_{i=1}^{N} w_i \boldsymbol{Y}_i$$

- Take a part of this state, a part of that state, and then average them together into a 'synthetic control' unit
- In most cases, we use convex weights $1 \geq w_i \geq 0$.

Choosing weights

We wnat our synthetic control unit to do a good job at approximating the pathway of outcomes for the treated unit: $\boldsymbol{Y}_0(0)$.

Choosing weights

We wnat our synthetic control unit to do a good job at approximating the pathway of outcomes for the treated unit: $\boldsymbol{Y}_0(0)$.

We only observe $Y_{0t}(0)$ for the treated unit up until period T_0 which is the period prior to treatment.

- Synthetic control should try to match the treated unit's outcome path during the pre-period and *HOPEFULLY* that will mean the synthetic control will do a good job in the post-period. It's a leap of faith

Choosing weights

More formally, the weights are selected by trying to minimize the following:

$$\underset{\{w_i\}}{\operatorname{argmin}} \sum_{t=1}^{T_0} \left(Y_{0t} - \sum_{i=1}^{N} w_i Y_{it} \right)^2$$

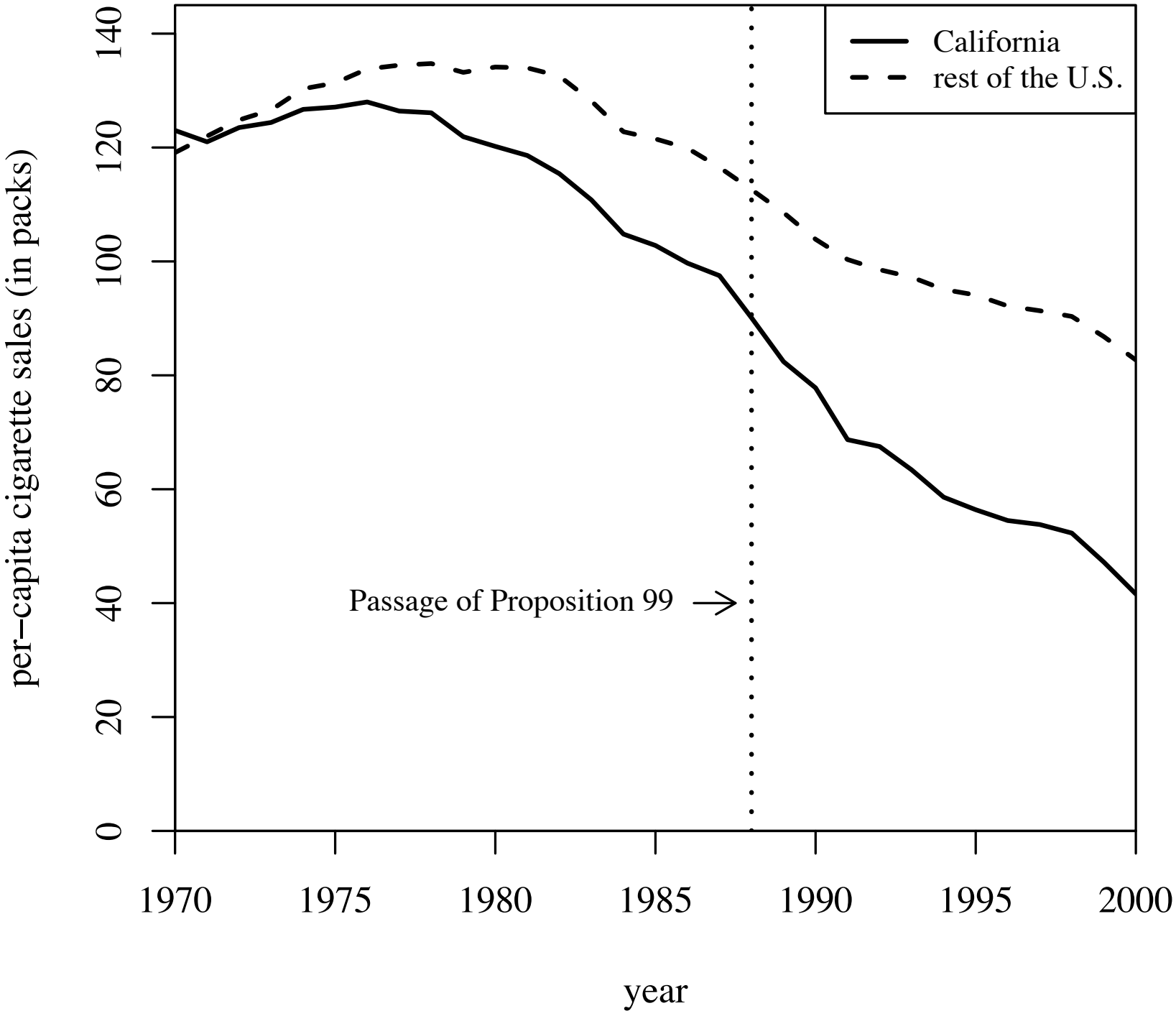
- Minimizing the pre-treatment sum of squared prediction errors between treated unit and the synthetic control unit.
- With convex weights, $1 \geq w_i \geq 0$, we add these as a constrained optimization problem

Example: California's Proposition 99

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

Other states that subsequently passed control programs are excluded from donor pool of controls (AK, AZ, FL, HI, MA, MD, MI, NJ, OR, WA, DC)


California
rest of the U.S.
Passage of Proposition 99
per−capita cigarette sales (in packs)
0
20
40
60
80
100
120
140
1970
1975
1980
1985
1990
1995
2000
year

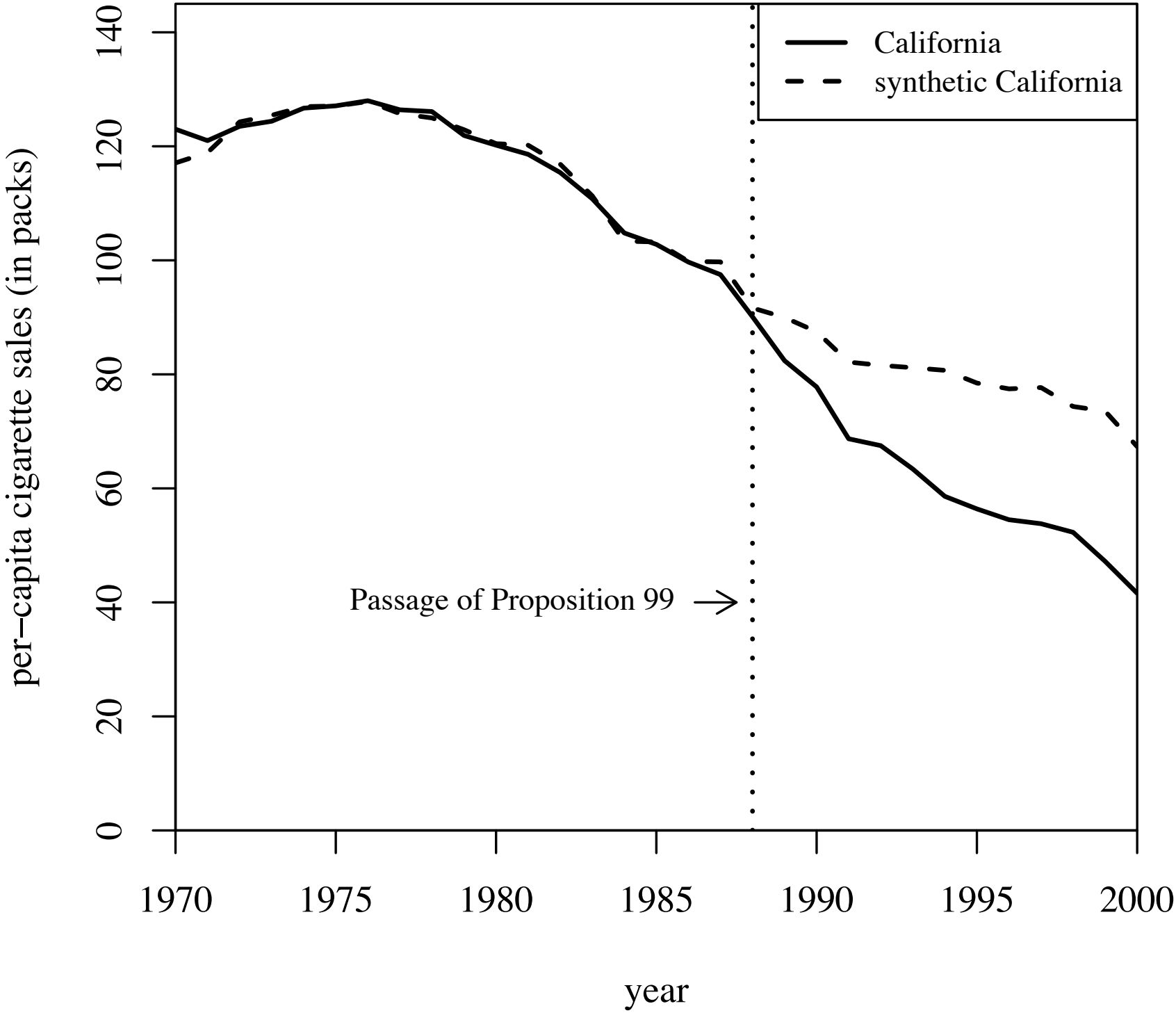

California
synthetic California
Passage of Proposition 99
per−capita cigarette sales (in packs)
0
20
40
60
80
100
120
140
1970
1975
1980
1985
1990
1995
2000
year

Table 2. State weights in the synthetic California

State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0

Beware of Overfitting

Intuitively, synthetic control is ‘believable’ when the synthetic control unit does a great job at approximating the outcome in the per-period

- We think that the synthetic control must be picking up on underlying economic structure in order to co-move with the treated state.
- E.g. if you see a set of runners whose times go up and go down during the same races, you might think they train together. You wouldn’t think that chance alone made them have the same ups and downs.

Beware of Overfitting

Intuitively, synthetic control is 'believable' when the synthetic control unit does a great job at approximating the outcome in the per-period

- We think that the synthetic control must be picking up on underlying economic structure in order to co-move with the treated state.
- E.g. if you see a set of runners whose times go up and go down during the same races, you might think they train together. You wouldn't think that chance alone made them have the same ups and downs.

The number one concern you should have when reading or writing a paper that uses a synthetic control method is that of overfitting.

- If I have 1000 control units and 4 pre-periods, I can probably well approximate Y_{0t} in the pre-period by just random chance.

Synthetic Control and Factor Model

Recall our factor model:

- There are a set of characteristics that the units have and in each period a set of macroeconmic shocks that change the marginal effect of those characteristics
- If we could observe these characteristics, we would want to match on them or use them in conditional PTs

Synthetic Control and Factor Model

Recall our factor model:

- There are a set of characteristics that the units have and in each period a set of macroeconmic shocks that change the marginal effect of those characteristics
- If we could observe these characteristics, we would want to match on them or use them in conditional PTs

Synthetic control, under conditions we will discuss, will create a synthetic control unit that has the same average characteristics as the treated unit!

- Since they are subject to the shocks the same amount, we think their outcome evolutions will match.

Synthetic Control and Factor Model

Okay, so I've pained a very rosy picture of synthetic control. When it works well, we fix the problem of non-parallel trends and we are able to impute the untreated potential outcome well.

However, I want to make clear that this method is **not a panacea**. The original paper is very general and makes it hard to know when it works and when it does not.

- More recent advancements all base discussion on a factor model for outcomes
- Hollingsworth and Wing working paper has great discussion

Bias of Synthetic Control

The original Abadie, Diamond and Hainmueller paper derive the bias of synthetic control when outcomes are generated by a linear factor-model.

The bias arises from **over-fitting on noise**. It is more common to over-fit on data when:

- There are fewer pre-treatment periods T_0
- There are many control units
- The 'convex hull' assumption is unlikely to ˙old

‘Convex Hull’ Assumption

When constraining the weights to be convex ($0 \leq w_i \leq 1$), the synthetic control assumption requires the ‘convex hull’ assumption:

- This is basically an assumption that says ‘we can approximate $\boldsymbol{Y}_0$ using a convex weighted average of control $\boldsymbol{Y}_i$’

‘Convex Hull’ Assumption

When constraining the weights to be convex ($0 \leq w_i \leq 1$), the synthetic control assumption requires the ‘convex hull’ assumption:

- This is basically an assumption that says ‘we can approximate $\boldsymbol{Y}_0$ using a convex weighted average of control $\boldsymbol{Y}_i$’

With a factor model, we can make this assumption a lot clearer:

- The ‘convex hull’ assumption is equivalent to the assumption that the treated unit’s ‘factor loadings’ are a convex average of the other units’ ‘factor loadings’
  - $\rightarrow$ That is, the treated unit can not have an extreme value in any of the $\lambda_{i,r}$ (e.g. huge manufacturing share)

Inference in Synthetic Control

Inference in the classical synthetic control setting is really difficult

- Only one treated unit; you're not averaging over units so the estimate is subject to random shocks

Inference in Synthetic Control

Inference in the classical synthetic control setting is really difficult

- Only one treated unit; you're not averaging over units so the estimate is subject to random shocks

Two forms of randomization inference are typically used:

- Randomly shuffle treatment to control units and reestimate synthetic control; want the treated unit to look more extreme than the placebo estimates. The so-called 'spaghetti plot'
- Randomly shuffle time-periods around for treated unit and reestimate synthetic control

Implementing Synthetic Control

My recommendation for synthetic control is **`scpi`** package (on R, Stata, and Python)

- Covers all the basic method and include inference methods
- Journal of Statistical Software paper is super readable: `https://nppackages.github.io/references/Cattaneo-Feng-Palomba-Titiunik_2024_JSS.pdf`

Roadmap

Imputation in 2x2

Imputation with Staggered Treatment Timing

Modeling non-parallel trends

Synthetic Control

More General Factor Model Imputation

‘Extensions’ of the Synthetic Control Model

Synthetic Control with Lasso/Ridge Penalty

People seem to like when the synthetic control is made up of a few units

- Makes the control unit more ‘interpretable’

‘Extensions’ of the Synthetic Control Model

Synthetic Control with Lasso/Ridge Penalty

Can modify the weights optimization problem to penalize weights being too large:

$$\underset{\{w_i\}}{\operatorname{argmin}} \sum_{t=1}^{T_0} \left(Y_{0t} - \sum_{i=1}^{N} w_i Y_{it} \right)^2 + \lambda \|w\|_k$$

- Add a term that punishes when weights are non-zero; λ is a ‘tuning-parameter’ to chose how much to punish
- Can add convex-weights constraint
- Lasso is $k = 1$; Ridge is $k = 2$

‘Extensions’ of the Synthetic Control Model

Augmented Synthetic Control

Augmented control is a method to help with imperfect pre-treatment fit by estimating a bias and subtracting it off.

- Looks similar to regression adjustment (DRDID without weights). We estimate the trend using covariates and then use that model to bias correct.

‘Extensions’ of the Synthetic Control Model

Augmented Synthetic Control

1. Calculate the standard synthetic control method weights (or with lasso)
2. Estimate a model of $m_t(X_i) = \mathbb{E}[\boldsymbol{Y}_{it} \mid X_i]$ using untreated units where X_i is pre-treatment characteristics (lagged Y or covariates)

Form synthetic control estimate as

$$\left(Y_{0t} - \sum_{i=1}^{N} w_i Y_{it}\right) - \underbrace{\left(m_t(X_0) - \sum_{i=1}^{N} w_i m_t(X_i)\right)}_{\text{‘bias correction’}}$$

‘Extensions’ of the Synthetic Control Model

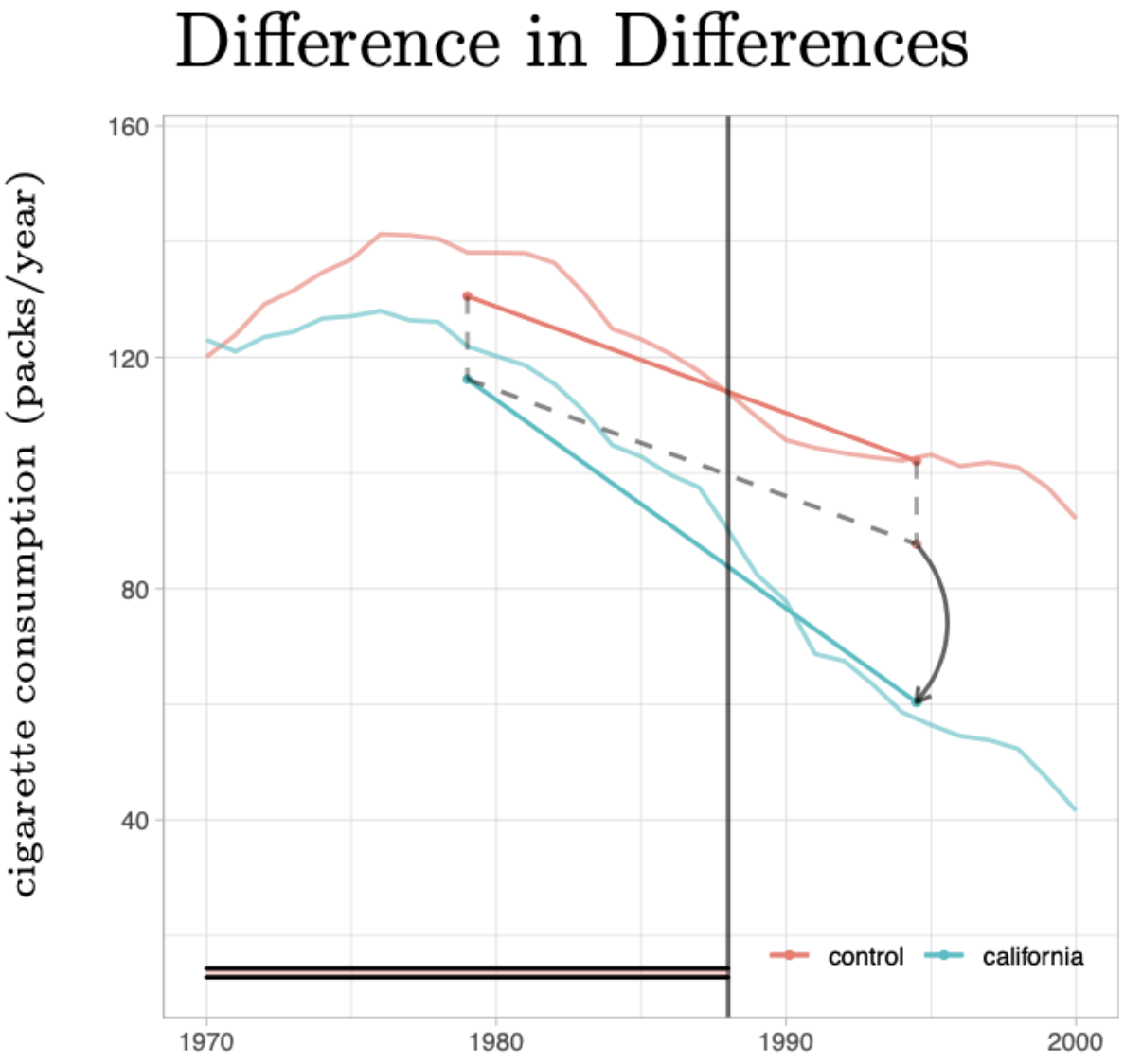
Synthetic Difference-in-Differences

Synthetic Control is based on the idea of ‘finding control units that follow the same fluctuations as the treated unit in the pre-period’

- Some periods are more informative than others (typically more recent years), so we should prefer to fit on those

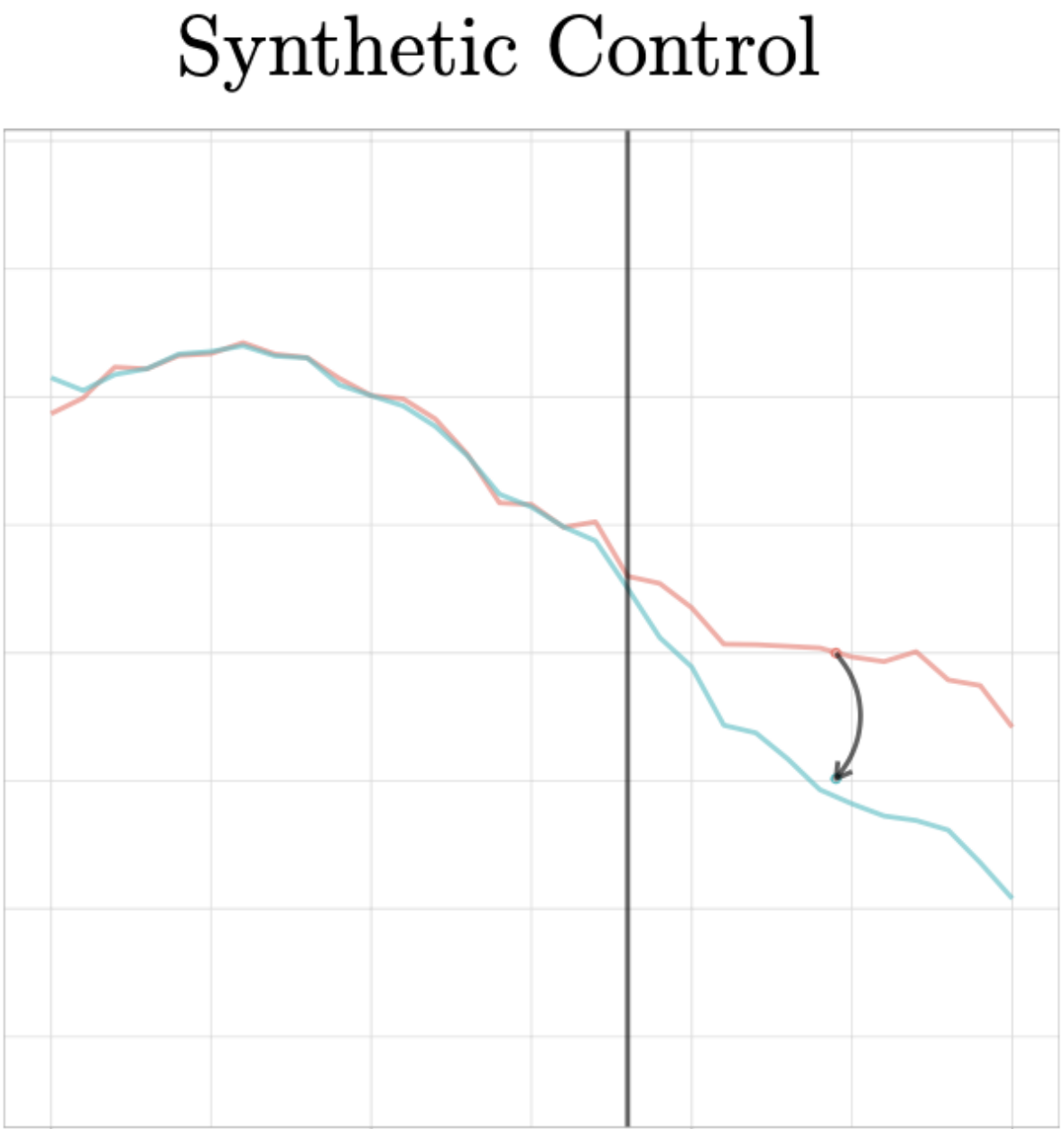
This

Example Synthetic DID



Estimated decrease:
-27.3 (17.7)

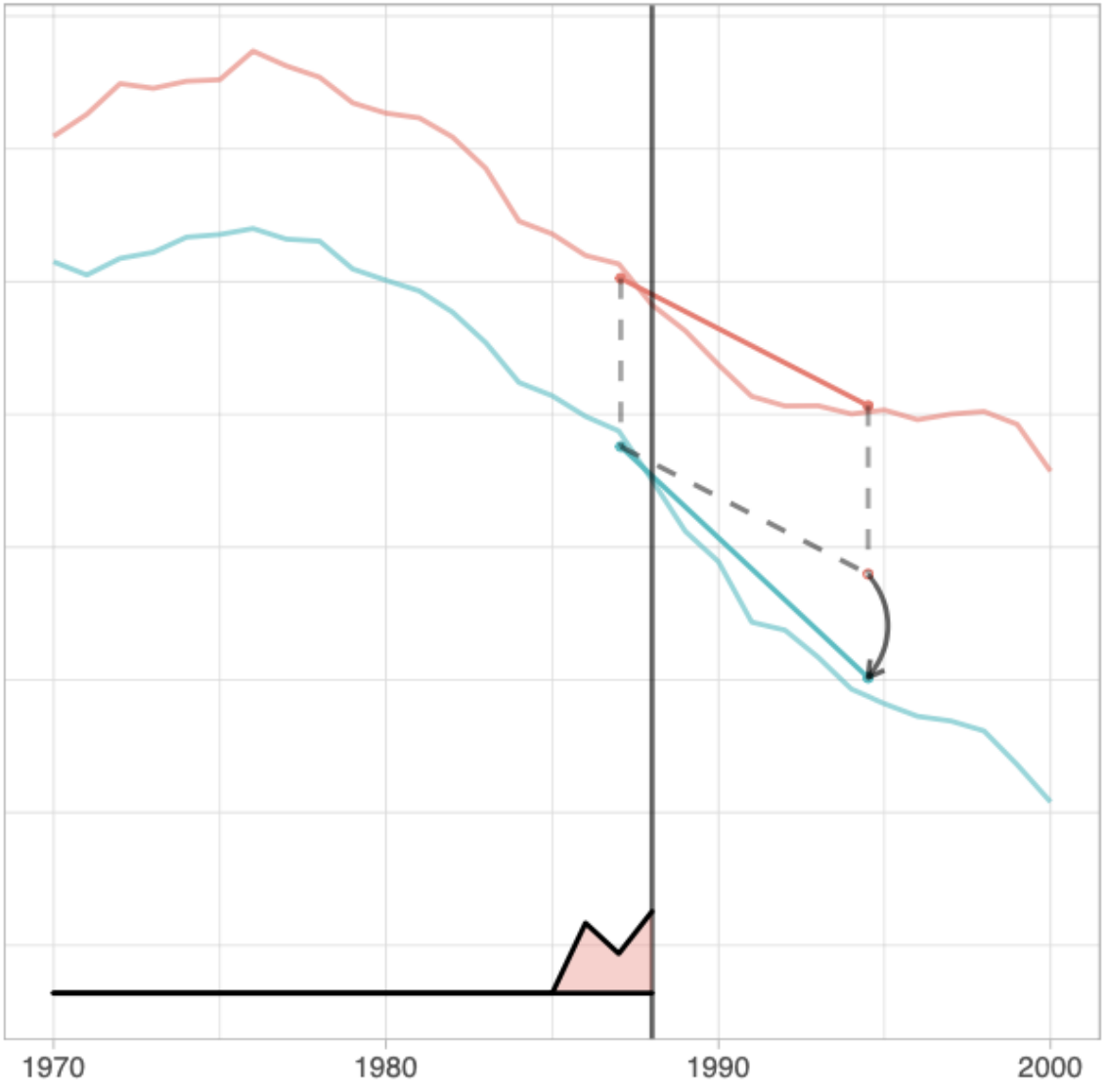
Example Synthetic DID



Estimated decrease: -19.6 (9.9)

Bad fit just prior because all time periods are weighted equally

Synthetic Diff. in Differences



Estimated decrease:
-19.6 (9.9)

Most weights on
most recent years

‘Extensions’ of the Synthetic Control Model

Generalized Imputation Estimator

Xu (2017) and Gobillon and Magnac (2016) both answer a simple question:

- Why not just estimate the factor model directly and impute untreated potential outcomes?

Intuition is to estimate $\lambda_{i,r}$ and $f_{t,r}$ using untreated/not-yet-treated observations and then impute:

$$\hat{Y}_{it}(0) = \hat{\mu}_i + \hat{\lambda}_t + \sum_{r=1}^{\rho} \hat{\lambda}_{i,r} \hat{f}_{t,r}$$

- We estimate the non-parallel trending via the factor model and then subtract it off

Factor Model Imputation

Unlike imputation of the two-way fixed effect model, this approach is very data hungry:

- Requires both a large number of time-periods and units

Intuitively, you need a long number of time-periods to estimate a unit's $\lambda_{i,r}$. You need a large number of units to estimate a time-period's $f_{t,r}$

Factor Model Imputation

Unlike imputation of the two-way fixed effect model, this approach is very data hungry:

- Requires both a large number of time-periods and units

Intuitively, you need a long number of time-periods to estimate a unit's $\lambda_{i,r}$. You need a large number of units to estimate a time-period's $f_{t,r}$

This estimator is likely more efficient when the underlying model is a factor model, but biased if a different underlying model is true

- E.g. a non-linear factor model

Problem with large-T

These new tools are all really powerful, but all rely on having access to many years of data. In a lot of applied work, the data is just not available

Problem with large-T

These new tools are all really powerful, but all rely on having access to many years of data. In a lot of applied work, the data is just not available

But I think there is a more subtle problem at play with this assumption:

- Data from many years ago might not be very useful at understanding the underlying confounders at play in this economy
- Imagine saying ”I use data from 1960 to inform me which counties would be a good control group for housing prices in 2000”. A lot has happened since then !!!!!!

Small-T methods

There is a nascent literature thinking about adapting factor-model estimators in settings with very few pre-periods available:

- Basically (1) my work with Nicholas Brown and (2) work by Brantly Callaway and coauthors

To conclude today, I'll discuss briefly my work

- But, it's been a long workshop, so I'll keep the self-promo short :-)

Small-T methods

It turns out that estimation of treatment effects in short panels with a large number of treated units only requires estimation of $f_{t,r}$

- The only reason we needed large panels was to estimate $\lambda_{i,r}$, so we can do estimation in $f_{t,r}$

Small-T methods

It turns out that estimation of treatment effects in short panels with a large number of treated units only requires estimation of $f_{t,r}$

- The only reason we needed large panels was to estimate $\lambda_{i,r}$, so we can do estimation in $f_{t,r}$

We develop a general approach to impuation in short panels that sounds a lot like the two-way fixed effects version:

1. Using just the untreated group, estimate $f_{t,r}$
2. Perform imputation of outcome variable (in paper)
3. Take averages of $Y_{it} - \hat{Y}_{it}(0)$

I'll discuss two main aproaches to estimate $f_{t,r}$ in short-panels

Estimation of $f_{t,r}$

Instrument Approach

We've discussed quite a bit during this workshop on thinking about the underlying characteristics that we think could cause non-parallel trends

- This is the same thought exercise as chosing covariates X_i

Estimation of $f_{t,r}$

Instrument Approach

We've discussed quite a bit during this workshop on thinking about the underlying characteristics that we think could cause non-parallel trends

- This is the same thought exercise as chosing covariates X_i

In some cases, you might be able to form some *noisy* measure of the exposure variables $\lambda_{i,r}$ (e.g. baseline value of share of college educated). These can be used as an instrumental variable to estimate $f_{t,r}$

- The relevancy condition is that $\lambda_{i,r}$ is correlated with the instrument

Estimation of $f_{t,r}$

Time-varying covariates approach

Alternatively, we might have a set of time-varying covariates, w_{it} that we think are impacted by the same macroeconomic shocks $f_{t,r}$ that impacts the outcome variable

- In this case, we can learn about the underlying shocks by movements in w_{it}

Factor Models

I think factor models are a really great tool for an applied econometrician

- We often want to study treatment in panel settings where we know treated units are selected
- But, if we think the treated units are on different trends based on broad macroeconomic shocks (and not location-specific shocks), then a factor model can estimate these broad trends and remove them from biasing our estimates

A lot of these estimators require long panels to estimate treatment effects consistently

- We have discussed why even long panels might be undesirable