Lecture 11: Repeated Observations I

POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

March 8, 2020

Overview

Today:

- One-way fixed effects regression.
- Simple difference-in-differences (DID).
- Simple event studies.
- Standard synthetic control.

Next time:

- Staggered adoption
- Generalized synthetic control and interactive fixed effects.
- Matrix completion and high dimensional approaches.

One-way Fixed Effects

- Units $i = 1, ..., N$ from large population.
- For each i observe (Y_{it}, D_{it}) for time periods: $t = 1, ..., T$.
- Outcome DGP:

$$Y_{it} = \mu + \rho D_{it} + U_i + \varepsilon_{it}.$$

- Assuming homogenous effect, ρ.
- Suppose $\varepsilon_{it} \sim$ [white noise] but $D_{it} = \gamma_0 + \gamma_1 U_i + \nu_{it}$, ν_{it} also white noise.
- OLS of Y on D has expectation $\rho + \gamma_1 (\text{Var}[U] / \text{Var}[D])$.
- Omitted variable bias.

Fixed effects

- Try something else—take one i at a time $\Rightarrow$ mini datasets:

$$\begin{aligned} Y_{i1} &= \mu + \rho D_{i1} + U_i + \varepsilon_{i1} \\ Y_{i2} &= \mu + \rho D_{i2} + U_i + \varepsilon_{i2} \\ &\vdots \\ Y_{iT} &= \mu + \rho D_{iT} + U_i + \varepsilon_{iT} \end{aligned}$$

  and OLS Y on D (for cases where $\text{Var}[D_{it}|i] > 0$).
- $(\mu + U)$ is constant, ε is random noise.
- Yields coef ρ_i with expected value ρ (homog. effects).
- Aggregate over all i and you get a more precise estimate of ρ.
- Maximal precision: weight proportional to $\text{Var}[D_{it}|i]$.
- "Within" regression.

Fixed effects

- Another idea, center on unit (i-specific) means:

$$\begin{aligned} Y_{it} - \bar{Y}_i &= (\mu - \mu) + \rho(D_{it} - \bar{D}_i) + (U_i - U_i) + (\varepsilon_{i1} - \bar{\varepsilon}_i) \\ Y_{it} - \bar{Y}_i &= \rho(D_{it} - \bar{D}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \\ Y^*_{it} &= \rho D^*_{it} + \varepsilon^*_{it} \end{aligned}$$

- OLS of Y^* on D^* has expected value ρ.
- “Sweep” transformation.

Fixed effects

- Now note the following:
- First, from lecture 3, recall using OLS to fit

$$Y_{it} = \mu + \rho_{i} D_i + \sum_{j=1}^{N} I(i=j) + \varepsilon_{it}$$

yields coefficient on D that converges to

$$\rho_R = \frac{\sum_i \rho_i \mathrm{Var}[D_{it}|i] \Pr[i]}{\sum_i \mathrm{Var}[D_{it}|i] \Pr[i]}.$$

- Same as "within" regression above.
- Second, apply FWL to this dummy variable regression:
- Residualizing wrt $1(i=j)$ is subtracting off mean values for unit j, leave other units untouched.
- Same as "sweep" transformation above.

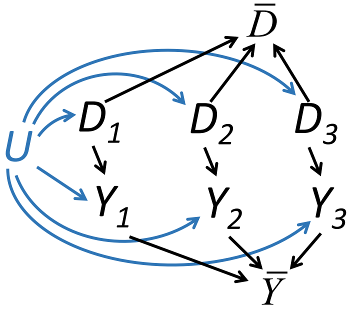
Fixed effects

- Thus, the following are algebraically equivalent:
  - Dummy variable OLS with $1(i = j)$,
  - Variance weighted average of coefs. from "within" OLS by i,
  - OLS after "sweep" transformation with i-specific means.
- This is "one-way fixed effects" regression.
- Addresses "time-invariant" confounders for DGPs like

$$
\begin{aligned}
Y_{it} &= \mu + \rho D_{it} + X_{it}\beta + U_i + \varepsilon_{it} \\
D_{it} &= \gamma_0 + \gamma_1 U_i + X_{it}\lambda + \nu_{it}.
\end{aligned}
$$

(X_{it}, assumed measured, added for more generality).

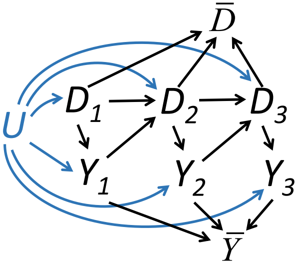
DGPs and FE identification



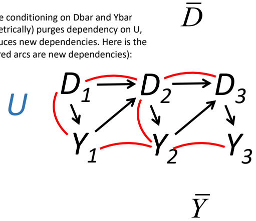
FE on a DAG:

- Partial out U via sweep – implies conditioning on $\bar{D}$ and $\bar{Y}$ that parametrically removes U.
- Identify effects for each time period, and then aggregate.

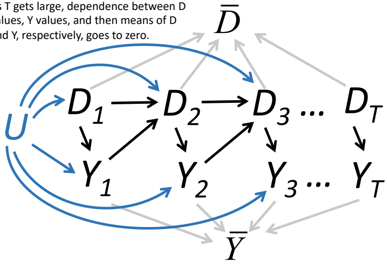
DGPs and FE identification



Suppose conditioning on Dbar and Ybar (parametrically) purges dependency on U, but induces new dependencies. Here is the result (red arcs are new dependencies):



As T gets large, dependence between D values, Y values, and then means of D and Y, respectively, goes to zero.



- Bias from serial correlation (Nickel bias).

Remarks on One-Way FE

- Identification for classical one-way fixed relies on linear separability of time-invariant confounder.
- Recall that if effects are heterogeneous, then the fixed effects regression yields a conditional variance weighted average:

$$\rho_R = \frac{\sum_i \rho_i \text{Var}[D_{it}|i] \Pr[i]}{\sum_i \text{Var}[D_{it}|i] \Pr[i]}.$$

- Centered interaction model can re-target the regression to the ATE (Imbens & Wooldridge 2009, p. 28).
- Cluster-robust standard errors or cluster bootstrap can account for dependencies within FE strata. In Stata, use t reghdfe so that you have the right degrees of freedom adjustment (usual commands like `areg`, `xtreg` are overconservative).

Differences-in-Differences (DID)

Classic difference-in-differences is a two-period, two-way fixed effects DGP:

- Suppose units index by i grouped into 2 sets indexed by $g = 0, 1$ for "treated" and "control" and two time periods labeled $t = 0, 1$.
- $D_{g[i]0} = 0$ for all g, while $D_{11} = 1$ and $D_{01} = 0$.

Differences-in-Differences (DID)

Classic difference-in-differences is a two-period, two-way fixed effects DGP:

- Suppose units index by i grouped into 2 sets indexed by $g = 0, 1$ for "treated" and "control" and two time periods labeled $t = 0, 1$.
- $D_{g[i]0} = 0$ for all g, while $D_{11} = 1$ and $D_{01} = 0$.
- A two-way "fixed effects" model with time and group effects is:

$$
\begin{aligned}
Y_{it} &= \mu + \alpha_{g[i]} + \lambda_t + \delta D_{g[i]t} + \varepsilon_{it} \\
&= \beta_0 + \beta_1 \cdot 1(g[i] = 1) + \beta_2 \cdot 1(t = 1) \\
&\quad + \delta \cdot 1(g[i] = 1) \cdot 1(t = 1) + \varepsilon_{it}
\end{aligned}
$$

where latter is how DID is often estimated.

Differences-in-Differences (DID)

Classic difference-in-differences is a two-period, two-way fixed effects DGP:

- Suppose units index by i grouped into 2 sets indexed by $g = 0, 1$ for "treated" and "control" and two time periods labeled $t = 0, 1$.
- $D_{g[i]0} = 0$ for all g, while $D_{11} = 1$ and $D_{01} = 0$.
- A two-way "fixed effects" model with time and group effects is:

$$\begin{aligned} Y_{it} &= \mu + \alpha_{g[i]} + \lambda_t + \delta D_{g[i]t} + \varepsilon_{it} \\ &= \beta_0 + \beta_1 \cdot 1(g[i] = 1) + \beta_2 \cdot 1(t = 1) \\ &\quad + \delta \cdot 1(g[i] = 1) \cdot 1(t = 1) + \varepsilon_{it} \end{aligned}$$

  where latter is how DID is often estimated.
- With a little algebra,

$$\delta = \mathrm{E}\left[Y_{i1} - Y_{i0} | g[i] = 1\right] - \mathrm{E}\left[Y_{i1} - Y_{i0} | g[i] = 0\right]$$

  which shows how δ is indeed a "difference in differences."
  (NB: 1 and 0 subscripts are time periods here, not potential outcomes.)

Differences-in-Differences

- To avoid confusion, let's call potential outcomes under control Y_{it}^{C} and under treatment Y_{it}^{T}.
- We observe $Y_{i0}^{C} = Y_{i0}$ for everyone.
- We observe $Y_{i1}^{C} = Y_{i1}$ for $g[i] = 0$, and $Y_{i1}^{T} = Y_{i1}$ for $g[i] = 1$.
- We would like the ATT: $\mathrm{E}\left[Y_{i1}^{T} - Y_{i1}^{C} | g[i] = 1\right]$.

Differences-in-Differences

- To avoid confusion, let's call potential outcomes under control Y_{it}^{C} and under treatment Y_{it}^{T}.
- We observe $Y_{i0}^{C} = Y_{i0}$ for everyone.
- We observe $Y_{i1}^{C} = Y_{i1}$ for $g[i] = 0$, and $Y_{i1}^{T} = Y_{i1}$ for $g[i] = 1$.
- We would like the ATT: $\mathrm{E}\left[Y_{i1}^{T} - Y_{i1}^{C} | g[i] = 1\right]$.
- Assumption 1: Suppose a form of mean independence:

$$\mathrm{E}\left[Y_{i1}^{C} - Y_{i0}^{C} | g[i] = 0\right] = \mathrm{E}\left[Y_{i1}^{C} - Y_{i0}^{C} | g[i] = 1\right].$$

Trend in control is equal to what trend *would have been* among treated *had treatment never been applied*.

Differences-in-Differences

- Then, δ is identified for the ATT:

Differences-in-Differences

- Then, δ is identified for the ATT:

$$\delta = \underbrace{\mathrm{E}[Y_{i1}^T - Y_{i0}^C | g[i] = 1] - \mathrm{E}[Y_{i1}^C - Y_{i0}^C | g[i] = 0]}_{\text{observed}}$$
$$= \mathrm{E}[Y_{i1}^T - Y_{i0}^C | g[i] = 1] - \underbrace{\mathrm{E}[Y_{i1}^C - Y_{i0}^C | g[i] = 1]}_{\text{counterfactual}}$$
$$= \mathrm{E}[Y_{i1}^T - Y_{i1}^C | g[i] = 1].$$

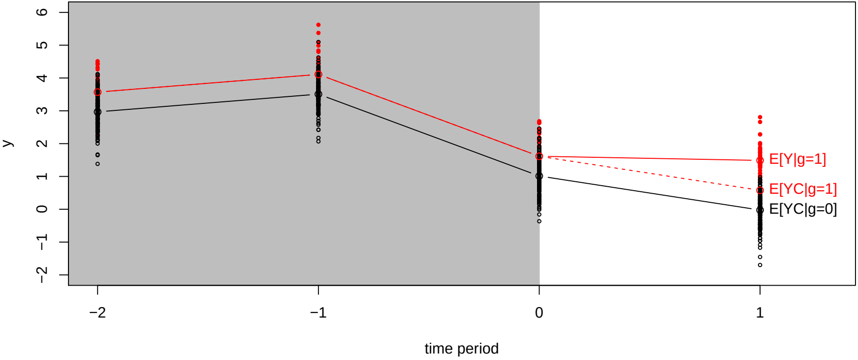
Differences-in-Differences

- Assumption 1 is the DID "parallel trends" assumption.
- Implies control group trend is parallel to what *would have happened* to treatment group members were there no treatment.
- I.e., control group trend is parallel to counterfactual trend for treatment group.

Differences-in-Differences

- Assumption 1 is the DID "parallel trends" assumption.
- Implies control group trend is parallel to what *would have happened* to treatment group members were there no treatment.
- I.e., control group trend is parallel to counterfactual trend for treatment group.
  - NB: Parallel trends *prior to treatment* lend plausibility but do not ensure parallel post-treatment trends for the control and the *counterfactual* of the treated.

Differences-in-Differences



Differences-in-Differences

Some considerations for identification:

- Trend assumptions are sensitive to transformations! (Linear trend in natural scale implies non-linear trend in log scale.)
- Trend assumptions may not be plausible on levels, though perhaps on differences or other higher order differences. Identification is still possible (Mora & Reggio, 2019).
- Trend assumptions may be plausible only for classes of similar units and not for treated and control groups as a whole $\rightarrow$ conditional DID...

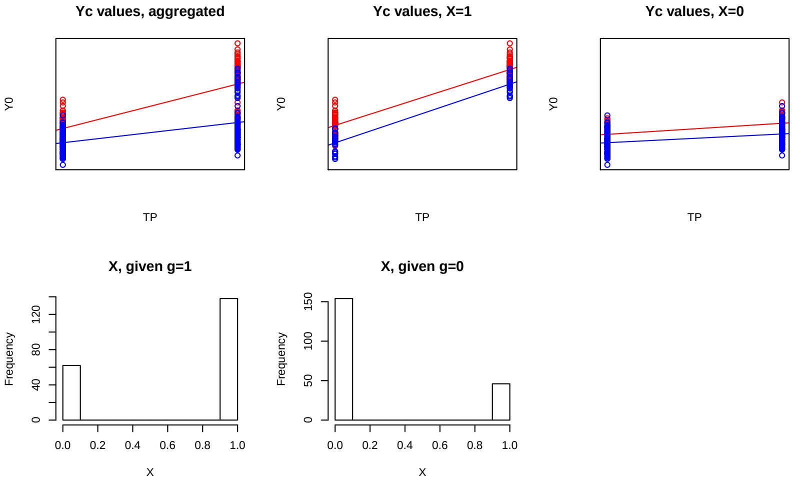
Conditional Differences-in-Differences

- Abadie (2005) considers conditional mean independence:
- Assumption 2:

$$\mathrm{E}\left[Y_{i1}^{C} - Y_{i0}^{C} | g[i] = 0, X_i\right] = \mathrm{E}\left[Y_{i1}^{C} - Y_{i0}^{C} | g[i] = 1, X_i\right],$$

and $\Pr[g[i] = 1 | X_i] < 1$ for all X_i.

Conditional Differences-in-Differences



Conditional Differences-in-Differences

- Then,

$$\begin{aligned}\delta_x &\equiv \mathrm{E}[Y_{i1} - Y_{i0}|g[i] = 1, X_i] - \mathrm{E}[Y_{i1} - Y_{i0}|g[i] = 0, X_i] \\ &= \mathrm{E}[Y_{i1}^T - Y_{i0}^C|g[i] = 1, X_i] - \mathrm{E}[Y_{i1}^C - Y_{i0}^C|g[i] = 0, X_i] \\ &= \mathrm{E}[Y_{i1}^T - Y_{i0}^C|g[i] = 1, X_i] - \mathrm{E}[Y_{i1}^C - Y_{i0}^C|g[i] = 1, X_i] \\ &= \mathrm{E}[Y_{i1}^T - Y_{i1}^C|g[i] = 1, X_i],\end{aligned}$$

and so the ATT is identified, since

$$\begin{aligned}\int_x \mathrm{E}[Y_{i1}^T - Y_{i1}^C|g[i] = 1, X_i = x] f(x|g[i] = 1) dx \\ = \mathrm{E}[Y_{i1}^T - Y_{i1}^C|g[i] = 1].\end{aligned}$$

Conditional Differences-in-Differences

- Three different ways to exploit Assumption 2:
  1. Regression model that incorporate X_i.
     - Consider interactions with time period and group dummies, higher order X_i terms, etc.
     - Key is to trace out outcome trajectories under control.
     - Risk of specification or aggregation biases.
  2. Inverse-propensity score weighting using $e(X_i)$.
  3. Matching on X_i.

Triple differences

- Policy in state (S) targets poor (P) but leaves rich (R) unaffected.

Triple differences

- Policy in state (S) targets poor (P) but leaves rich (R) unaffected.
- Under parallel trends for P and R, could estimate effect as,

$$\delta = \mathrm{E}\left[Y_{i1}^{T} - Y_{i0}^{C} | P\right] - \mathrm{E}\left[Y_{i1}^{C} - Y_{i0}^{C} | R\right].$$

Triple differences

- Policy in state (S) targets poor (P) but leaves rich (R) unaffected.
- Under parallel trends for P and R, could estimate effect as,

$$\delta = \mathrm{E}\left[Y_{i1}^{T} - Y_{i0}^{C} | P\right] - \mathrm{E}\left[Y_{i1}^{C} - Y_{i0}^{C} | R\right].$$

- Controls for any "state" effects.
- But poor may have different trends than rich.

Triple differences

- Policy in state (S) targets poor (P) but leaves rich (R) unaffected.
- Under parallel trends for P and R, could estimate effect as,

$$\delta = \mathrm{E}\left[Y_{i1}^{T} - Y_{i0}^{C} | P\right] - \mathrm{E}\left[Y_{i1}^{C} - Y_{i0}^{C} | R\right].$$

- Controls for any "state" effects.
- But poor may have different trends than rich.
- We can incorporate poor from another state (S'):

$$\delta_3 = \mathrm{E}\left[Y_{i1}^{T} - Y_{i0}^{C} | P, S\right] - \Big(\underbrace{\mathrm{E}\left[Y_{i1}^{C} - Y_{i0}^{C} | R, S\right]}_{\text{state effect}} + \underbrace{\mathrm{E}\left[Y_{i1}^{C} - Y_{i0}^{C} | P, S'\right]}_{\text{poor effect}}\Big)$$

- With P, S, and T as poor, state S, and $t = 1$ indicators, estimate

$$Y_{it} = \beta_0 + \beta_1 P + \beta_2 S + \beta_3 PS + \delta_0 T + \delta_1 PT + \delta_2 ST + \delta_3 PST + \varepsilon_{it}$$

- Incorporate covariates as above.

DID inference

- Generally, “robust” or “cluster robust” within g.
- If clustering is at the level of g, then there is a problem – only two groups!
- Recent contributions on DID inference with few groups: MacKinnon & Webb (2016), Ferman & Pinto (2015).
- A different, and I think especially promising angle, is Doudchenko & Imbens (2017)—more later.

Event Studies

- Another classical design is the "event study".
- Jacobsen et al. (1993) study on earnings losses from employment displacement: "difference between their actual and expected earnings had the events that led to their job losses not occurred."
- Treatment is the displacement event, effects may initiate prior to the displacement itself and may accumulate over time:

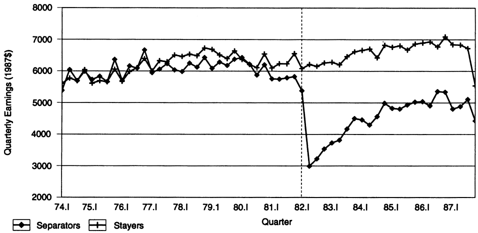


Figure 1. Quarterly Earnings (1987 Dollars) of High-Attachment Workers Separating in Quarter 1982:1 and Workers Staying Through Quarter 1986:4

Event Studies

Two-way FE model with period specific coefficients: for person i in period t,

$$Y_{it} = \alpha_i + \gamma_t + x_{it}\beta + \sum_{k \geq -m} D_{it}^k \delta_k + \varepsilon_{it},$$

- γ_t tracks general trends,
- x_{it} tracks covariate specific variation in the absence of displacement,
- $-m$ is a set of lead periods prior to actual displacement during which effects may nonetheless start,
- k indexes accumulated effects surrounding displacement.

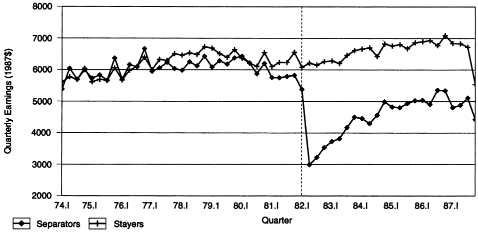


FIGURE 1. QUARTERLY EARNINGS (1987 DOLLARS) OF HIGH-ATTACHMENT WORKERS SEPARATING IN QUARTER 1982:1 AND WORKERS STAYING THROUGH QUARTER 1986:4

Synthetic Control

- Abadie & Gardeazabal (2003) developed a method similar to the event study design for “quantitative case studies” (causal inference with a single treated observation).

Synthetic Control

- Abadie & Gardeazabal (2003) developed a method similar to the event study design for "quantitative case studies" (causal inference with a single treated observation).
- Their application is to estimate the effects of terrorism in the Basque region on the prosperity of the region.
- They do so by creating a "synthetic" Basque region out of the rest of Spain.

Synthetic Control

TABLE 3—PRE-TERRORISM CHARACTERISTICS, 1960'S

	Basque Country (1)	Spain (2)	"Synthetic" Basque Country (3)
Real per capita GDP[a]	5,285.46	3,633.25	5,270.80
Investment ratio (percentage)[b]	24.65	21.79	21.58
Population density[c]	246.89	66.34	196.28
Sectoral shares (percentage)[d]			
Agriculture, forestry, and fishing	6.84	16.34	6.18
Energy and water	4.11	4.32	2.76
Industry	45.08	26.60	37.64
Construction and engineering	6.15	7.25	6.96
Marketable services	33.75	38.53	41.10
Nonmarketable services	4.07	6.97	5.37
Human capital (percentage)[e]			
Illiterates	3.32	11.66	7.65
Primary or without studies	85.97	80.15	82.33
High school	7.46	5.49	6.92
More than high school	3.26	2.70	3.10

Sources: Authors' computations from Matilde Mas et al. (1998) and Fundación BBV (1999).

[a] 1986 USD, average for 1960–1969.

[b] Gross Total Investment/GDP, average for 1964–1969.

[c] Persons per square kilometer, 1969.

[d] Percentages over total production, 1961–1969.

[e] Percentages over working-age population, 1964–1969.

Synthetic Control

- Suppose 1 treated and J potential control regions.

Synthetic Control

- Suppose 1 treated and J potential control regions.
- Let $W = (w_1, ..., w_J)'$ be a vector of non-negative weights for each control region. Let $\sum_k w_j = 1$.
- Control region outcomes are combined using W to create a synthetic counterfactual for the treated region.
- We want to choose W^* to create the best possible synthetic counterfactual.

Synthetic Control

- Let X_1 be covariates for the treated region and $\mathbf{X}_0$ a matrix of covariates for the control region.

Synthetic Control

- Let X_1 be covariates for the treated region and $\mathbf{X}_0$ a matrix of covariates for the control region.
- Choose W^* to minimize weighted squared distance,

$$(X_1 - \mathbf{X}_0 W)' \mathbf{V} (X_1 - \mathbf{X}_0 W),$$

  where $\mathbf{V}$ weights the different covariate discrepancies on the basis of their relative "importance."
- You are free to choose $\mathbf{V}$ as you see fit.
- Abadie & Gardeazabal choose $\mathbf{V}$ to give priority to minimizing distance between the pre-conflict GDP trend.

Synthetic Control

- Let X_1 be covariates for the treated region and $\mathbf{X}_0$ a matrix of covariates for the control region.
- Choose W^* to minimize weighted squared distance,

$$(X_1 - \mathbf{X}_0 W)' \mathbf{V} (X_1 - \mathbf{X}_0 W),$$

  where $\mathbf{V}$ weights the different covariate discrepancies on the basis of their relative "importance."
- You are free to choose $\mathbf{V}$ as you see fit.
- Abadie & Gardeazabal choose $\mathbf{V}$ to give priority to minimizing distance between the pre-conflict GDP trend.
- Then, synthetic control outcomes for the treated region are computed as $\hat{Y^C}_{it} = Y'_{jt} W^*$.
- They use time series techniques (essentially ADL models) for estimation and inference.

Synthetic Control

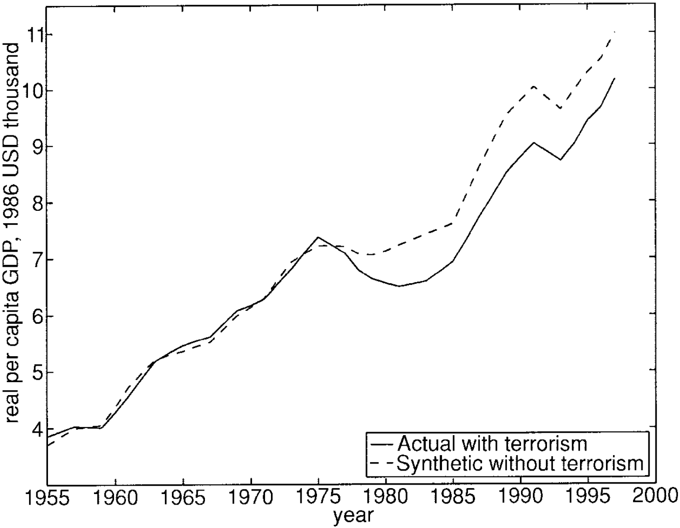


FIGURE 1. PER CAPITA GDP FOR THE BASQUE COUNTRY

Remarks

- Explosion of literature recently on FE, DID. event studies, synth and related methods.
- Next lecture will go through some of these more recent results.
- Many of these paper point to the “contaminated” nature of the comparisons that conventional two-way FE models construct and propose ways to estimate things more cleanly.