A Practical Guide to Counterfactual Estimators for Causal Inference with Panel Data

Licheng Liu (MIT), Ye Wang (UNC), Yiqing Xu (Stanford)

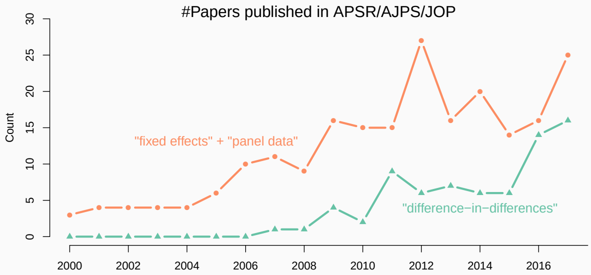
University of Warwick — March 2022

Motivations

Motivations

Two-way fixed effects models are one of the most commonly used statistical routines for TSCS data in the social sciences.

- Accounting for unobserved unit and time heterogeneity
- Flexible, e.g. a treatment can switch on and off
- Easy to estimate and interpret



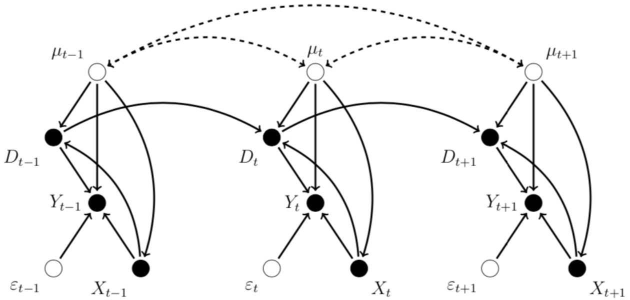
Motivations

$$Y_{it} = \tau D_{it} + X'\beta + \alpha_i + \xi_t + \varepsilon_{it}$$

in which D_{it} is dichotomous

- This approach has shortcomings (Imai and Kim 2019):
  - Strict exogeneity implies: no time-varying confounders and no feedback from past outcome to treatment
  - Functional form implies treatment effect homogeneity and no carryover effects
  - Recent literature focuses on the homogeneity assumption, whose failure will lead to “negative weighting,” hence, biases (Chernozhukov et al. 2013; Goodman-Bacon 2018; de Chaisemartin and D’Haultfœuille 2018; Borusyak, Jaravel & Spiess 2021)

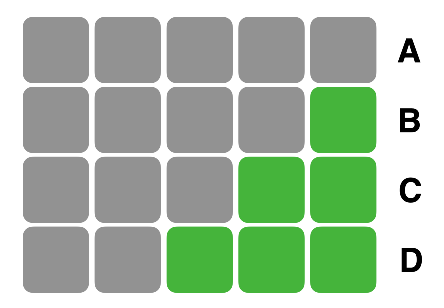
Interpreting Strict Exogeneity



Note: Unit indicies are dropped for simplicity. Vector μ_t represents unobserved time-invariant and decomposable time-varying (for IFEct and MC) confounders.

Recall: no feedback; no time-varying confounder; no anticipation effect; no carryover effects (can be relaxed)

What's Negative Weighting?



- Question: Can TWFE get at (some weighted) ATT when the treatment effects are heterogeneous
- Probably not! (Goodman-Bacon 2018; de Chaisemartin and D'Haultfœuille 2018)
- Early adopters (e.g. D) serves as controls for late adopters (e.g. B)

⇒ Some treated observations receive negative weights

This Paper

1. Propose a simple framework of counterfactual estimation for TSCS data to relax the homogeneity assumption and account for decomposable time-varying confounders

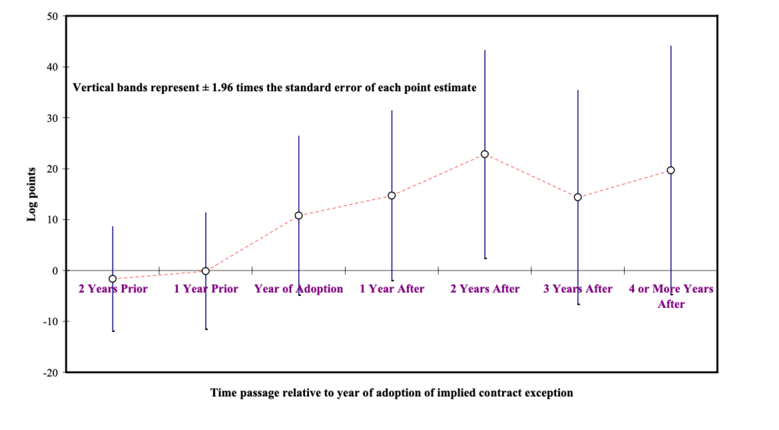
2. Discuss three counterfactual estimators, which directly imputes treated counterfactuals:
   - Fixed effects counterfactual (FEct)
   - Interactive fixed effects counterfactual (IFEct)
   - Matrix completion (MC)

3. Main advantage: accommodate general treatment patterns

4. Provide a set of diagnostic tools to gauge the validity of strict exogeneity assumption
   - A new plot for dynamic treatment effects

Current Practice for Evaluating Assumptions: *Angrist & Pischke* Chapter 5

Plot for "Dynamic Treatment Effects"



Common practices have drawbacks (Sun and Abraham 2020; Borusyak, Jaravel & Spiess 2021)

This Paper

1. Propose a simple framework of counterfactual estimation for TSCS data to relax the homogeneity assumption and account for decomposable time-varying confounders

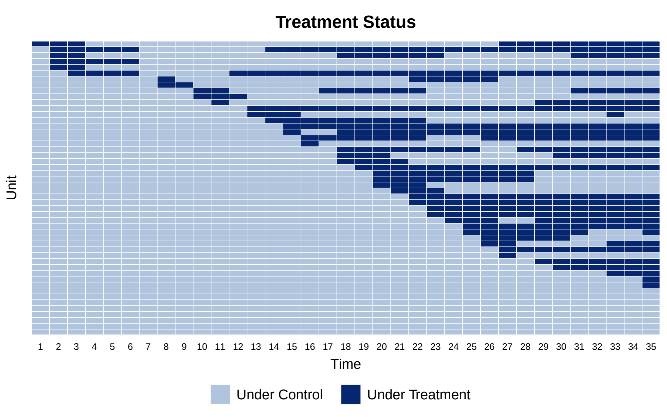
2. Discuss three counterfactual estimators, which directly imputes treated counterfactuals:
   - Fixed effects counterfactual (FEct)
   - Interactive fixed effects counterfactual (IFEct)
   - Matrix completion (MC)

3. Main advantage: accommodate general treatment patterns

4. Provide a set of diagnostic tools to gauge the validity of strict exogeneity assumption
   - A new plot for dynamic treatment effects
   - A placebo test
   - Extension: a test for (no) carryover effects
   - Extension: a test for (no) pretrend

Intuition

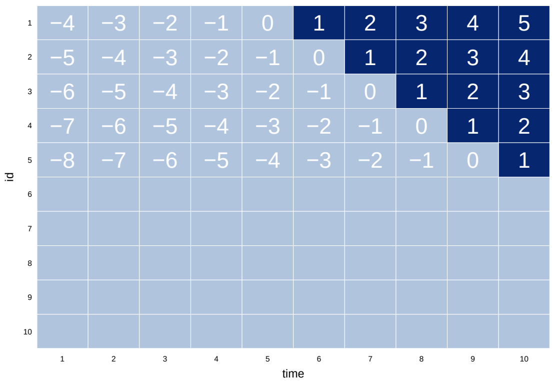
- In a TSCS setting, treat $Y(1)$ as missing data
- Use untreated data to build a model
- Estimate ATT by averaging differences between $Y(1)$ and $\hat{Y}(0) \Rightarrow$ no negative weighting!



Intuition

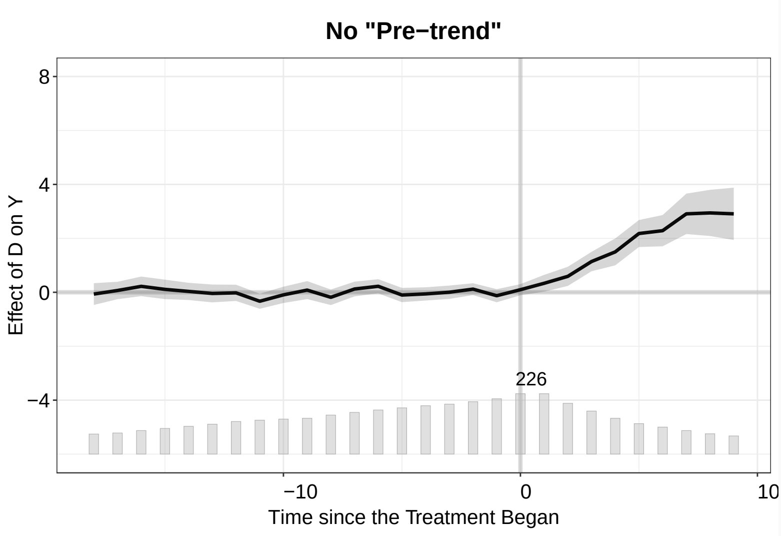
$$\widehat{ATT} = \hat{\mathbb{E}}[\hat{\tau}_{it} | D_{it} = 1, C_i = 1]$$

$$\widehat{ATT}_s = \hat{\mathbb{E}}[\hat{\tau}_{it} | D_{i,t-s} = 0, \underbrace{D_{i,t-s+1} = D_{i,t-s+2} = \cdots = D_{it} = 1}_{s \text{ periods}}, C_i = 1].$$

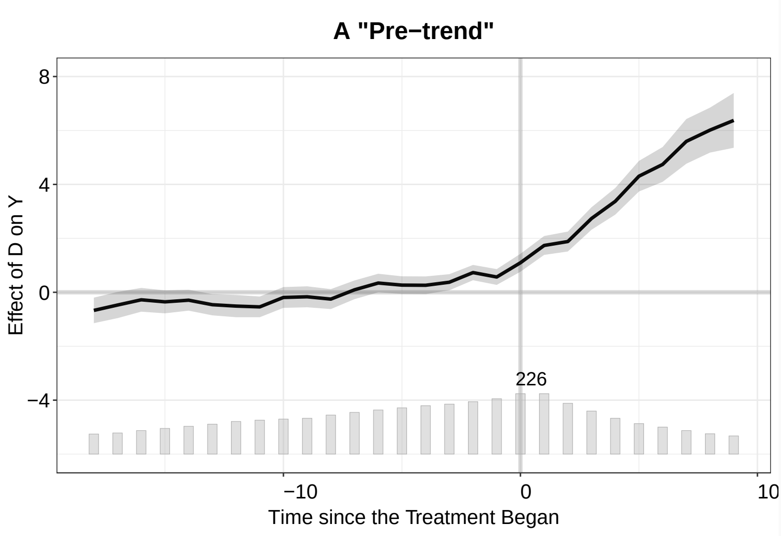


	1	2	3	4	5	6	7	8	9	10
1	−4	−3	−2	−1	0	1	2	3	4	5
2	−5	−4	−3	−2	−1	0	1	2	3	4
3	−6	−5	−4	−3	−2	−1	0	1	2	3
4	−7	−6	−5	−4	−3	−2	−1	0	1	2
5	−8	−7	−6	−5	−4	−3	−2	−1	0	1
6										
7										
8										
9										
10										

A New Plot for Dynamic Treatment Effects



A New Plot for Dynamic Treatment Effects



Plan

1. Motivation
2. Estimators
    - FEct, IFEct, MC
    - Remarks on Properties and Inference
3. Diagnostics
    - A New Plot
    - Placebo Test
    - Test for No Carryover Effects
    - Test for No Pre-trend
4. Empirical Examples
    - Hainmueller & Hangatner (2015)
    - Fouirnaies & Mutlu-Eren (2015)

Estimators

Examples of Counterfactual Estimators

We review three estimation strategies:

- FEct (this paper; Borusyak et al 2020; Gardner 2021):
$$\hat{Y}_{it}(0) = X_{it}\hat{\beta} + \hat{\alpha}_i + \hat{\xi}_t$$
- IFEct (Gobillon&Magnac 2016; Xu 2017):
$$\hat{Y}_{it}(0) = X_{it}\hat{\beta} + \hat{\lambda}_i'\hat{F}_t$$
- Matrix Completion (MC) (Athey et al. 2018):
$$\hat{Y}_{it}(0) = X_{it}\hat{\beta} + \hat{L}_{it},$$
where matrix $\{L_{it}\}_{N\times T}$ is a lower-rank matrix approximation of $\{Y(0)\}_{N\times T}$ with missing values

Remarks:

- DiD is a special case of FEct
- FEct is a special case of `gsynth`
- Both IFEct and MC are estimated via iterative algorithms
- Cross-validation to choose the tunning parameter

Key Assumptions

Assumption 1 (Functional form → additive separability)

For any $i = 1, 2, \cdots, N$ and $t = 1, 2, \cdots, T$,

$$Y_{it}(0) = f(\mathbf{X}_{it}) + h(\mathbf{U}_{it}) + \varepsilon_{it},$$

in which $f(\cdot)$ and $h(\cdot)$ are known, parametric functions.

Assumption 2 (Strict exogeneity → baseline assignment; no anticipation or feedback)

For any $i, j = 1, 2, \cdots, N$ and $t, s = 1, 2, \cdots, T$,

$$\varepsilon_{it} \perp\!\!\!\perp \{D_{js}, \mathbf{X}_{js}, \mathbf{U}_{js}\}, \textit{ for all } i, j \in \{1, 2, \ldots, N\} \textit{ and } s, t \in \{1, 2, \ldots, T\}.$$

Assumption 3 (Low-dimensional decomposition → feasibility)

There exists a low-dimensional decomposition of $h(\mathbf{U}_{it})$: $h(\mathbf{U}_{it}) = L_{it}$, and rank$(\mathbf{L}_{N\times T}) \ll \min\{N, T\}$.

Overview of Properties – FEct and IFEct

Proposition 1 (Unbiasedness and Consistency of FEct)

Under Assumptions 1-3 as well as some regularity conditions,

$$\mathbb{E}[\widehat{ATT}_s] = ATT_s \text{ and } \mathbb{E}[\widehat{ATT}] = ATT;$$
$$\widehat{ATT}_s \xrightarrow{p} ATT_s \text{ and } \widehat{ATT} \xrightarrow{p} ATT \text{ as } N \to \infty.$$

Proposition 2 (Consistency of IFEct)

Under Assumptions 1-3 as well as some regularity conditions,

$$\widehat{ATT} \xrightarrow{p} ATT \text{ as } N, T \to \infty.$$

IFE vs. MC

- Singular value decomposition of L

$$\mathbf{L}_{N\times T} = \mathbf{S}_{N\times N}\boldsymbol{\Sigma}_{N\times T}\mathbf{R}_{T\times T}$$

- Difference in how $\boldsymbol{\Sigma}_{N\times T}$ is regularized

IFE

best subset

$$\begin{pmatrix}
\sigma_1 & 0 & 0 & \cdots & 0 \\
0 & \sigma_2 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0
\end{pmatrix}$$

MC

nuclear norm

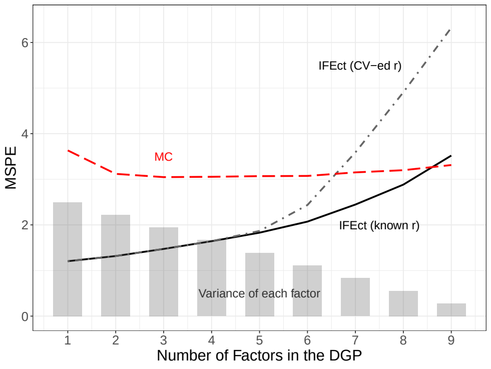
$$\begin{pmatrix}
|\sigma_1 - \lambda_L|_+ & 0 & 0 & \cdots & 0 \\
0 & |\sigma_2 - \lambda_L|_+ & 0 & \cdots & 0 \\
0 & 0 & |\sigma_3 - \lambda_L|_+ & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & 0 \\
 & & & & |\sigma_T - \lambda_L|_+ \\
0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0
\end{pmatrix}$$

in which $|a|_+ = \max(a, 0)$

Adapted from Athey et al. (2021)

IFEct vs. MC

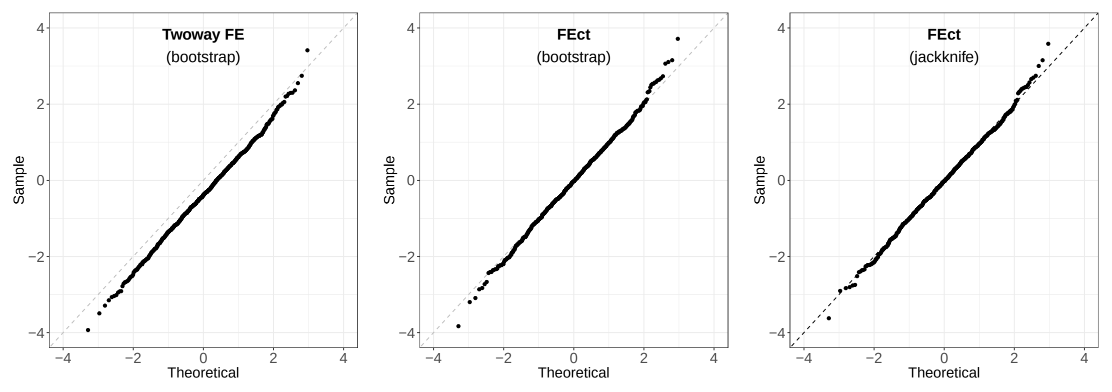
- IFEct works better with a small number of strong factors
- MC works better with a large number of weak factors



Inferential Methods

- Non-parametric block bootstrap
  - sample with replacement across units
  - valid when N is large, $\frac{N_{tr}}{N}$ is fixed
- Jackknife
  - dropping one treated unit a time
  - suitable when the number of treated units is small

QQ Plots: Theoretical vs. Empirical



Plan

1. Motivation
2. Estimators
   - FEct, IFEct, MC
   - Remarks on Properties and Inference
3. **Diagnostics**
   - A New Plot
   - Placebo Test
   - Test for No Carryover Effects
   - Tests for No Pre-trend
4. Empirical Examples
   - Hainmueller & Hangatner (2015)
   - Fouirnaies & Mutlu-Eren (2015)

Diagnostic Tests

A Simulated Example

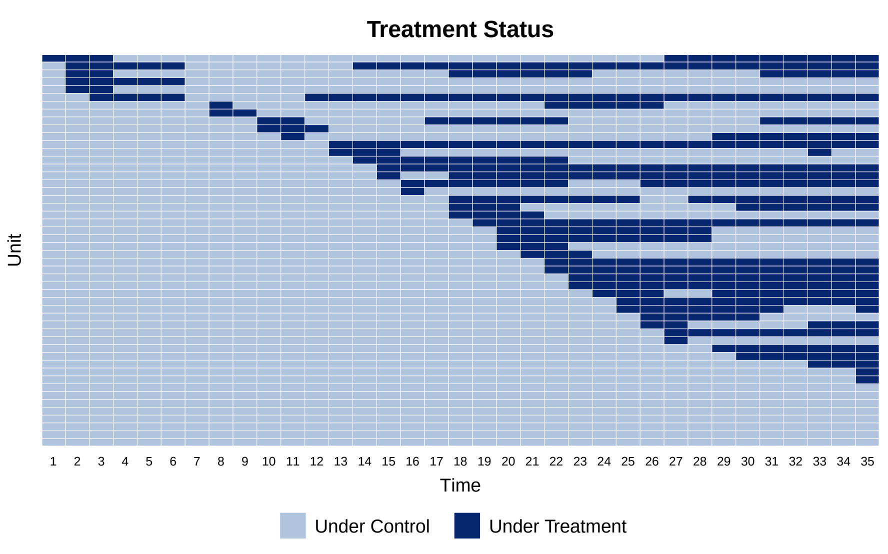
Data Generating Process:

- $T = 35$, $N = 200$
- **Outcome model**: a linear interactive fixed effect model with two factors: one drift process and one white noise.

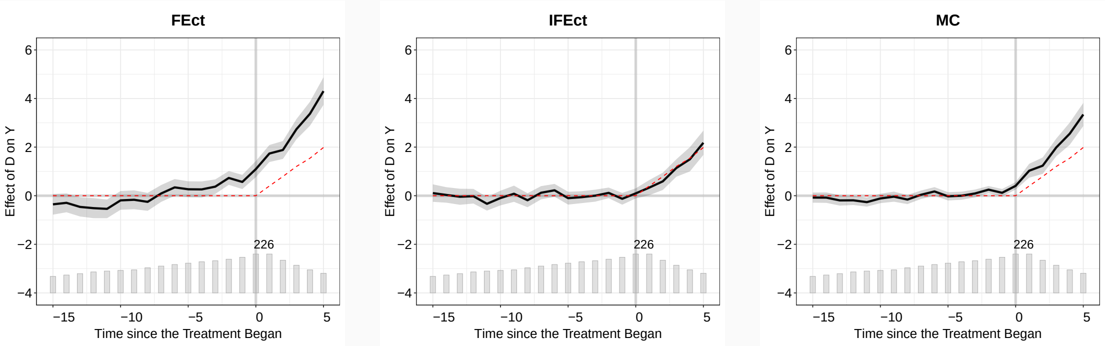
$$Y_{it} = \tau_{it} D_{it} + 5 + 1 \cdot X_{it,1} + 3 \cdot X_{it,2} + \lambda_{i1} \cdot f_{1t} + \lambda_{i2} \cdot f_{2t} + \alpha_i + \xi_t + \varepsilon_{it}$$

- **Treatment assignment**: general structure with the prob of getting treated correlated with additive and interactive fixed effect.
- **Treatment effects**: $\tau_{i,t>T_{0i}} = [0.4(t - T_{0i}) + e_{it}] * D_{it}$, hence, no carryover effects

Dynamic Treatment Effects



Dynamic Treatment Effects

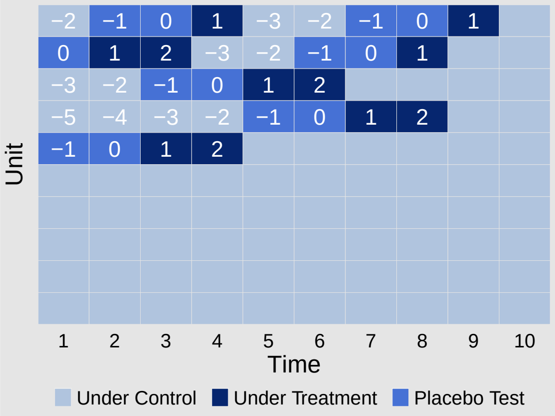

FEct

IFEct

MC

1. Placebo Test

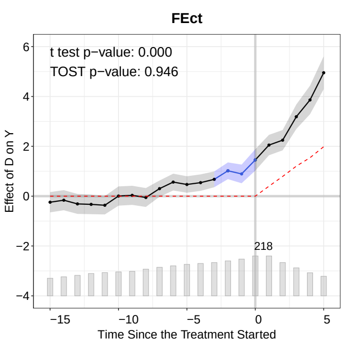
- Drop S periods before the treatment's onset, and estimate the "ATT" in these periods.

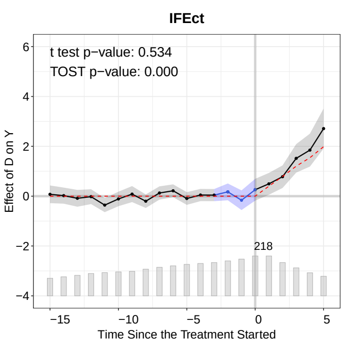


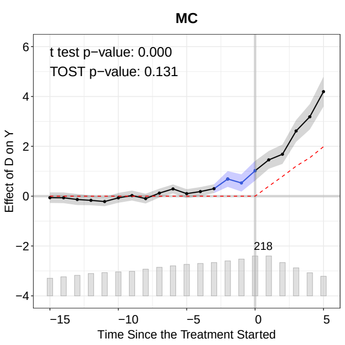
1	2	3	4	5	6	7	8	9	10
−2	−1	0	1	−3	−2	−1	0	1	
0	1	2	−3	−2	−1	0	1		
−3	−2	−1	0	1	2				
−5	−4	−3	−2	−1	0	1	2		
−1	0	1	2						

1. Placebo Test

- Drop S periods before the treatment's onset, and estimate the "ATT" in these periods.
- Benefits: intuitive and robust to model misspecification
- Accommodate both a difference-in-means (DIM) test or an equivalence test





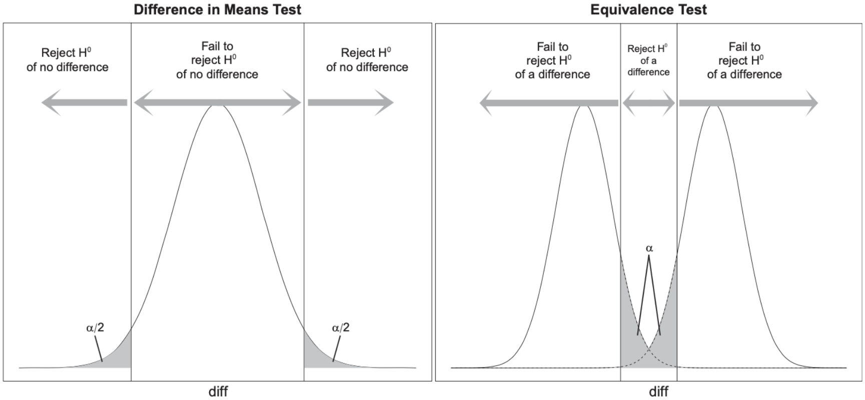


Why an Equivalence Test?

- A DIM test: H0: $|ATT^p| = 0$ vs. H1: $|ATT^p| > 0$
- An equivalence test: H0: $|ATT^p| > \theta$ vs. H1: $|ATT^p| \leq \theta$
- Compare with a DIM test,
  - it is conservative when the power is limited;
  - gains more power when the sample size (N) grows larger;
- Use a pre-specified threshold: $\theta = 0.36 * sd(\tilde{Y}_{it, D_{it}=0})$
- An extension to Hartman and Hidalgo (2018) in a TSCS setting

- **Drawback 1**: setting the threshold requires user discretion
- **Drawback 2**: use only limited information

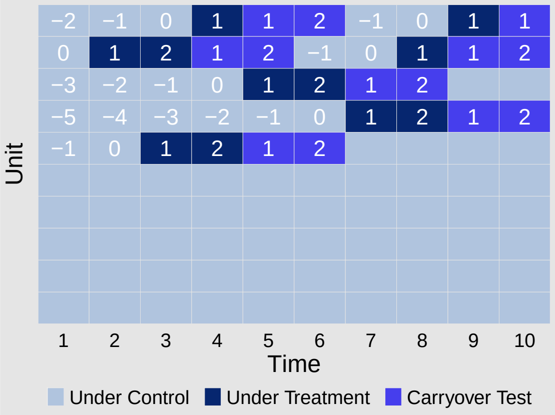
Why an Equivalence Test? (Hartman 2021)


Difference in Means Test
Reject H^0 of no difference
Fail to reject H^0 of no difference
Reject H^0 of no difference
α/2
α/2
diff
Equivalence Test
Fail to reject H^0 of a difference
Reject H^0 of a difference
Fail to reject H^0 of a difference
α
diff

Note: The left panel depicts the logic of tests of difference under the null hypothesis of no difference. The right panel depicts the logic of one type of equivalence test—the two one-sided t-test (TOST)—under the null hypothesis of difference.

2. Test for No Carryover Effects

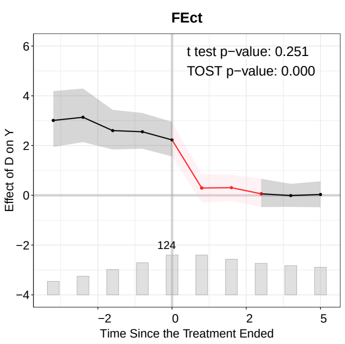
- Drop S periods after the treatment's ending, and estimate the average carryover effect (ACOE) in these periods.

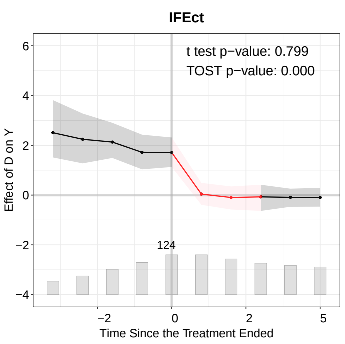


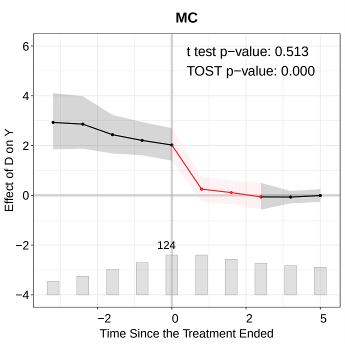
1	2	3	4	5	6	7	8	9	10
−2	−1	0	1	1	2	−1	0	1	1
0	1	2	1	2	−1	0	1	1	2
−3	−2	−1	0	1	2	1	2		
−5	−4	−3	−2	−1	0	1	2	1	2
−1	0	1	2	1	2				

2. Test for No Carryover Effects

- Drop S periods after the treatment's ending, and estimate the average carryover effect (ACOE) in these periods.





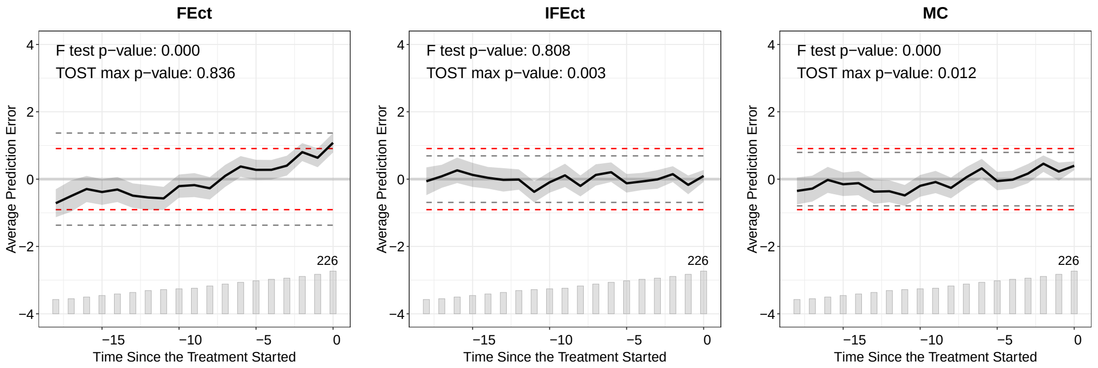


3. Test for No Pre-Trend

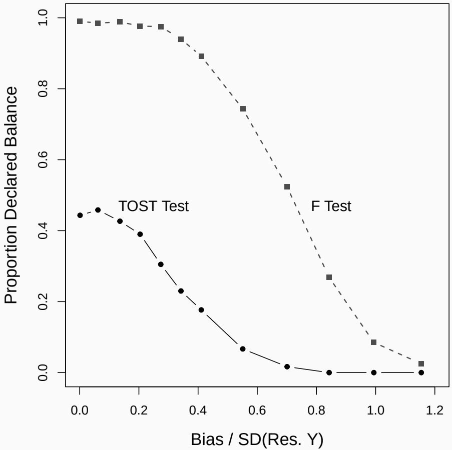
- One drawback of the placebo test is that it only uses limited information and may be under-powered
- We extend it to a test for no pre-trend by dropped one pre-treatment period a time (leave-one-period-out)
- H0: $|ATT_s| > \theta, \exists s \leq T_0$ vs. H1: $|ATT_s| \leq \theta, \forall s \leq T_0$

- **Drawback**: easy to pass when pre-treatment data are used to fit the model, e.g. IFEct and MC, because of serial correlation in data

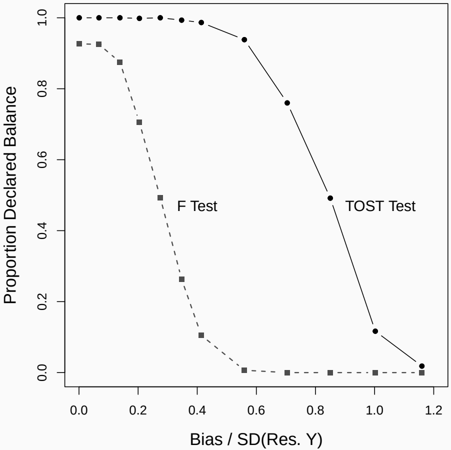
3. Tests for No Pre-Trend



Equivalence Test vs. the F Test?



$N = 100$



$N = 300$

Diagnostic Tests Summary

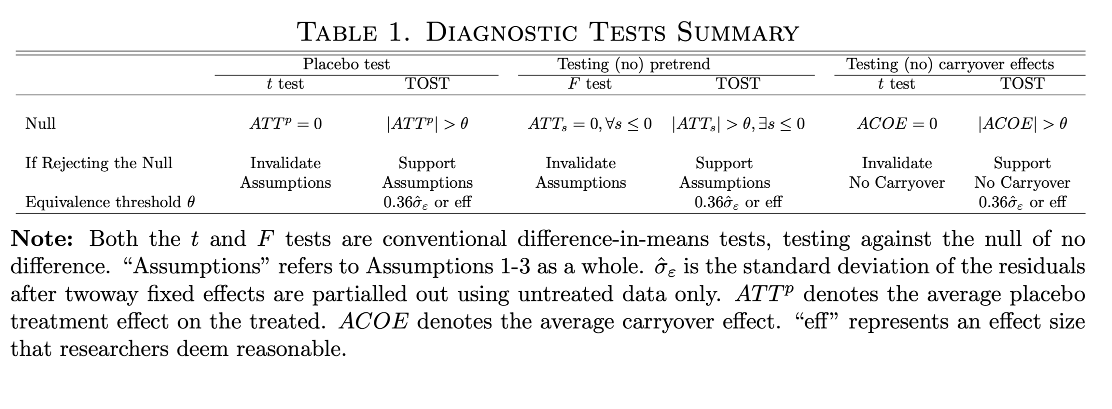


Table 1. Diagnostic Tests Summary

	Placebo test		Testing (no) pretrend		Testing (no) carryover effects	
	t test	TOST	F test	TOST	t test	TOST
Null	$ATT^p = 0$	$\|ATT^p\| > \theta$	$ATT_s = 0, \forall s \leq 0$	$\|ATT_s\| > \theta, \exists s \leq 0$	$ACOE = 0$	$\|ACOE\| > \theta$
If Rejecting the Null	Invalidate Assumptions	Support Assumptions	Invalidate Assumptions	Support Assumptions	Invalidate No Carryover	Support No Carryover
Equivalence threshold θ		$0.36\hat{\sigma}_\varepsilon$ or eff		$0.36\hat{\sigma}_\varepsilon$ or eff		$0.36\hat{\sigma}_\varepsilon$ or eff

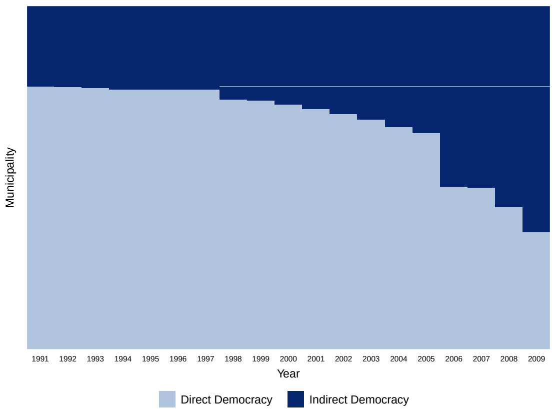
Note: Both the t and F tests are conventional difference-in-means tests, testing against the null of no difference. "Assumptions" refers to Assumptions 1-3 as a whole. $\hat{\sigma}_\varepsilon$ is the standard deviation of the residuals after twoway fixed effects are partialled out using untreated data only. ATT^p denotes the average placebo treatment effect on the treated. $ACOE$ denotes the average carryover effect. "eff" represents an effect size that researchers deem reasonable.

Empirical Examples

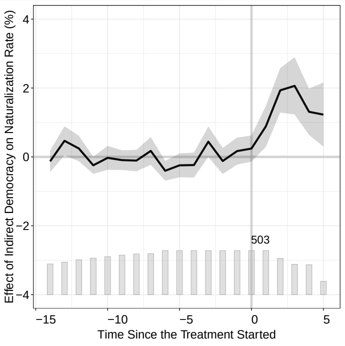
Hainmueller & Hangatner (2015)

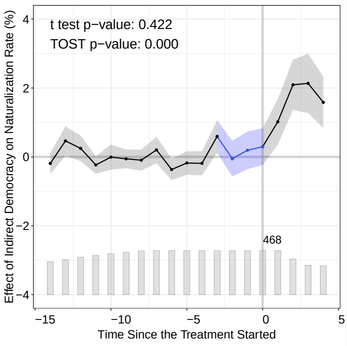
Does indirect democracy benefit immigrant minorities?

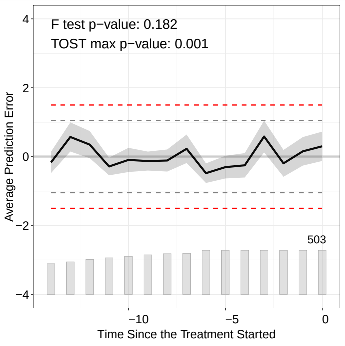
- Unit of analysis: 1400 Swiss municipalities from 1991 to 2009
- Treatment: Indirect (vs. direct) democracy
- Outcome: Naturalization rate



Hainmueller & Hangatner (2015) – FEct



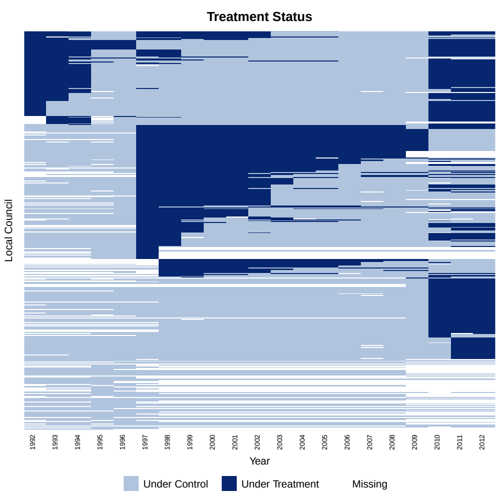




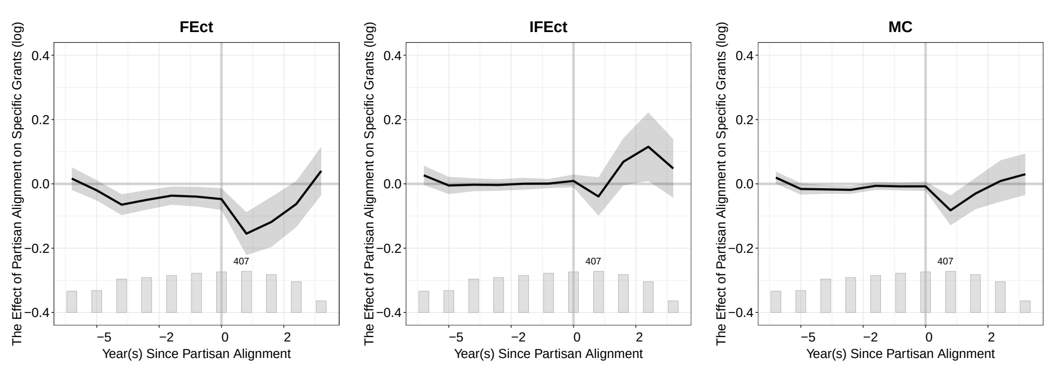
Fouirnaies and Mutlu-Eren (2015)

Does partisan alignment bring about central government grants in UK?

- Unit of analysis: 466 local councils from 1992 to 2012
- Treatment: Partisan alignment with the central government
- Outcome: Amount of specific grant

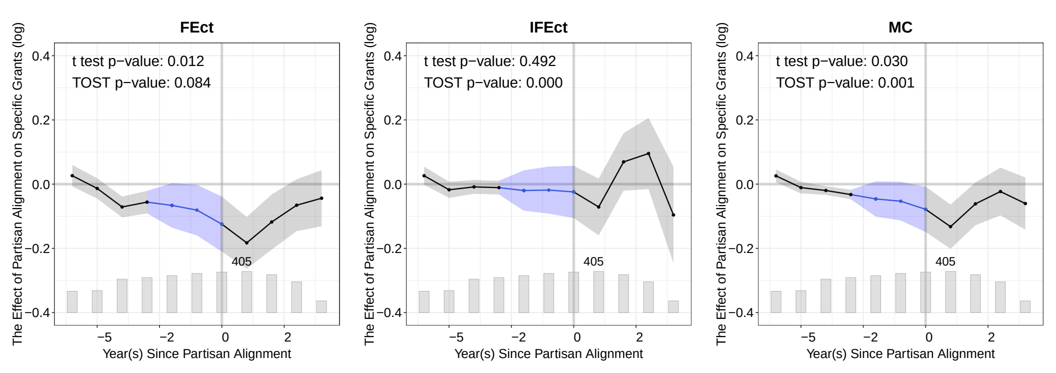


Fouirnaies and Mutlu-Eren (2015)

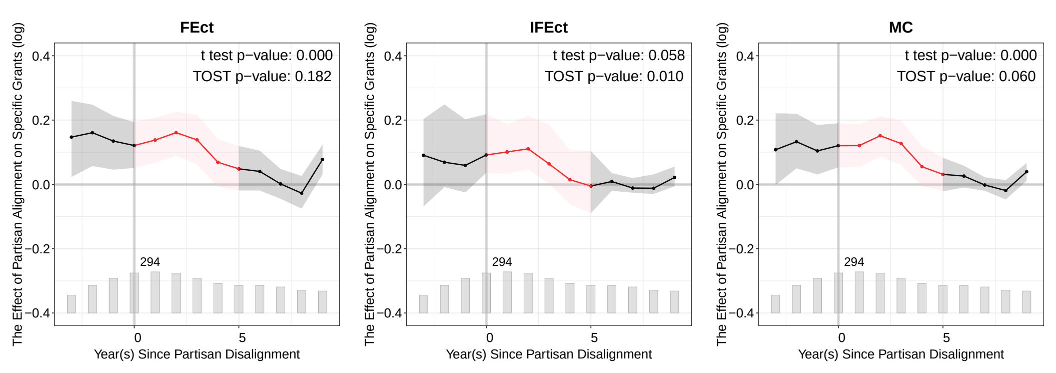


Note: The authors added unit-specific linear time trends to a TWFE model, whose results that are broadly consistent with those from IFEct.

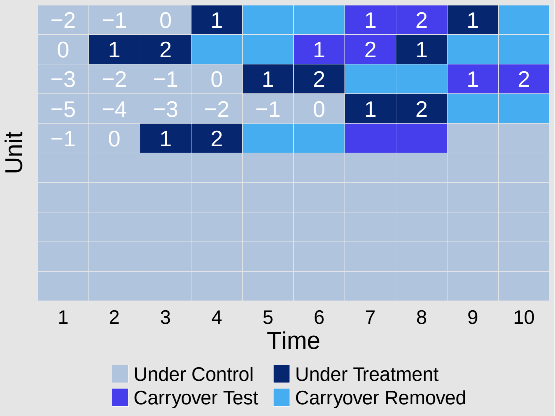
Fouirnaies and Mutlu-Eren (2015)



Fouirnaies and Mutlu-Eren (2015)

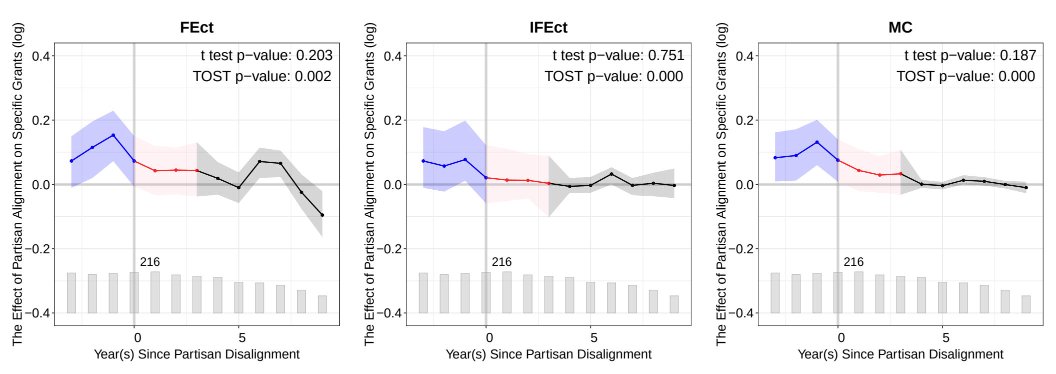


Addressing Limited Carryover (up to 2 periods)

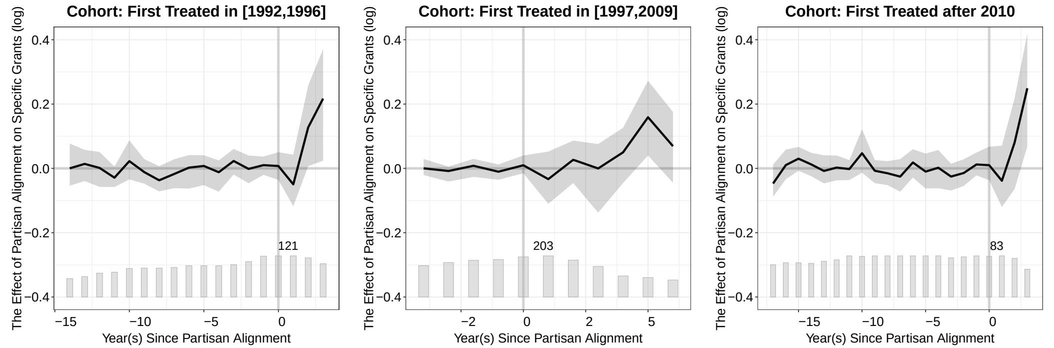


Fouirnaies and Mutlu-Eren (2015): Addressing Limited Carryover

Mark three periods after treatment ended as the “carryover” periods:



Fouirnaies and Mutlu-Eren (2015): Cohort Effect



Recommendations

- Plot your data (treatment and outcome) and ask whether strict exogeneity assumption is a plausible
- Start with FEct, draw the dynamic treatment effects plot and perform tests.
- If FEct fails the tests, apply more complex models, such as IFEct and MC, and perform diagnostics again.
- If the chosen method fails the test for no carryover effects, remove several periods after the treatment ends from the model-building stage, then re-apply the method and conduct diagnostics again.
- If a treatment effect is detected, perform subgroup analysis to understand which group(s) of units are driving the effect.
- Communicate your findings effectively, ideally with figures.

Concluding Remarks

1. We survey a group of counterfactual estimators that relax the homogeneity assumption (hence, no negative weighting issues) and account for decomposable time-varying confounders
2. We propose diagnostic tools to evaluate the key identification assumption and explore carryover effects
3. Open source package `panelView` and `fect` in R and Stata

   → transparency, transparency, transparency!
4. **Future work:** sequential assignment (w/ feedback); less parametric assumptions, interference; more complex structure; discrete outcomes...