Chapter 4: Negative Weights

Peter Hull

Applied Econometrics II
Brown University
Spring 2024

Outline

1. Negative Own-Treatment Weights
2. Contamination Bias
3. Application: Finkelstein et al. (2016)

Why So Negative?

The past few years has seen an explosion of applied ‘metrics work showing some conventional estimators don’t “play well” with effect heterogeneity

- Specifically, that they sometimes average together heterogeneous effects, with non-convex weights (seems bad!)

Recall the Angrist (1998) result on OLS + selection-on-observables:

- Convex weights, as long as you control flexibly enough for confounders
- But what if we’re in “parallel trends”-land, where we don’t assume the treatment is conditionally random?
- Angrist ‘98 only concerns a single treatment; what if they’re multiple?

We’ll tackle these problems in turn, before discussing some solutions

- Main takeaway: *Don’t Panic!* The jury is still out on how important these problems are empirically ...

Example: Staggered Adoption

Recall we previously studied the ATT interpretation of TWFE in two time periods, where treatment only flips on for some in $T = 2$

- Now suppose we have a panel with $t = 1, \dots, T$
- Units adopt a binary treatment at different dates $G_i \in \{1, \dots, T\} \cup \infty$, where $G_i = \infty$ means "never treated"

We continue to run a TWFE regression:

$$Y_{it} = \beta D_{it} + \alpha_i + \tau_t + v_{it}$$

where $D_{it} = \mathbf{1}[t \geq G_i]$ indicates treatment receipt

- If we start with a constant FX model, $Y_{it} = \beta D_{it} + \varepsilon_{it}$, we'd be done!
- But notice something a bit weird here: we can run this regression even if there are no never-treated units ...

Simple Staggered Adoption

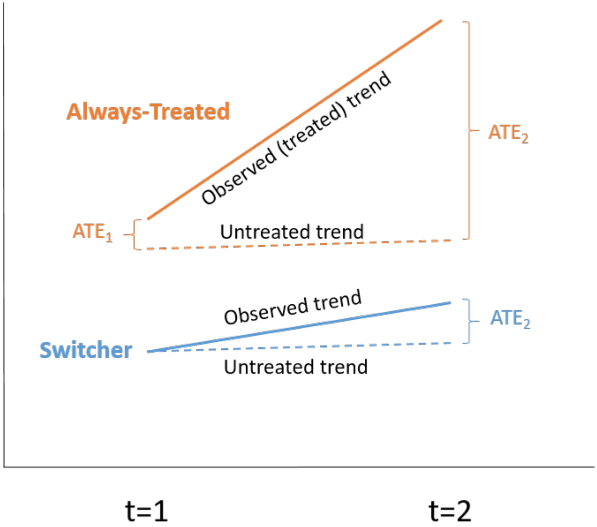
Consider $T = 2$ with two groups: *always-treated* units (with $G_i = 1$; $D_{i1} = D_{i2} = 1$) and *switchers* (with $G_i = 2$; $D_{i1} = 0$, $D_{i2} = 1$)

- We can use the usual two-period trick: $\Delta Y_i = \tau + \beta \Delta D_i + \Delta \varepsilon_i$, so $\beta = E[\Delta Y_i \mid G_i = 2] - E[\Delta Y_i \mid G_i = 1]$

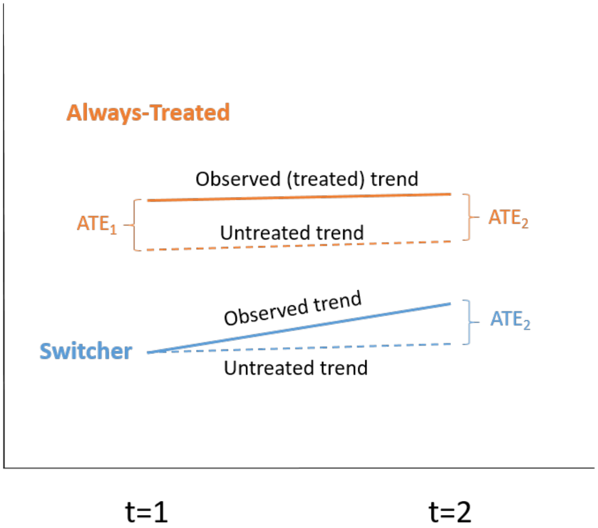
Assume PT holds: $E[Y_{i2}(0) - Y_{i1}(0) \mid G_i = 1] = E[Y_{i2}(0) - Y_{i1}(0) \mid G_i = 2]$

$$
\begin{aligned}
\beta =& E[Y_{i2}(1) - Y_{i1}(0) \mid G_i = 2] - E[Y_{i2}(1) - Y_{i1}(1) \mid G_i = 1] \\
=& E[Y_{i2}(1) - Y_{i2}(0) \mid G_i = 2] + E[Y_{i2}(0) - Y_{i1}(0) \mid G_i = 2] \\
& - E[Y_{i2}(1) - Y_{i2}(0) \mid G_i = 1] + E[Y_{i1}(1) - Y_{i1}(0) \mid G_i = 1] \\
& - E[Y_{i2}(0) - Y_{i1}(0) \mid G_i = 1] \\
=& \underbrace{E[Y_{i2}(1) - Y_{i2}(0) \mid G_i = 2]}_{\text{ATE for switchers}} \\
& - (\underbrace{E[Y_{i2}(1) - Y_{i2}(0) \mid G_i = 1] - E[Y_{i1}(1) - Y_{i1}(0) \mid G_i = 1]}_{\text{Change in ATE for always-treated}})
\end{aligned}
$$

“Forbidden Comparisons,” Illustrated



No Problem Under Constant Effects



General Problem

Suppose the causal model is $Y_{it} = \beta_{it} D_{it} + \varepsilon_{it}$ for heterogeneous β_{it}

- Linearity is without-loss for binary D_{it}; further complications arise with continuous treatments (see e.g. Calaway et al. (2021))
- Assuming static effects for simplicity (more on this soon)

Frisch-Waugh-Lovell: OLS yields

$$\hat{\beta} = \frac{\sum_{it} Y_{it} \tilde{D}_{it}}{\sum_{it} \tilde{D}_{it}^2} = \frac{\sum_{it} \beta_{it} D_{it} \tilde{D}_{it}}{\sum_{it} D_{it} \tilde{D}_{it}} + \frac{\sum_{it} \varepsilon_{it} \tilde{D}_{it}}{\sum_{it} D_{it} \tilde{D}_{it}}$$

Parallel trends implies $E[\sum_{it} \varepsilon_{it} \tilde{D}_{it} \mid D] = 0$, so $E[\hat{\beta}] = E[\sum_{it} \omega_{it} \beta_{it}]$

- The weights $\omega_{it} = \frac{D_{it} \tilde{D}_{it}}{\sum_{js} D_{js} \tilde{D}_{js}}$ aggregate to one ($E[\sum_{it} \omega_{it} = 1$). But they may not be convex: could have $\omega_{it} < 0$
- Unlike with Angrist '98, can't "average over" the random treatment

Is This A Problem?

In theory, negative weights could matter a lot: β_{it} could be zero or negative for all (it), but $E[\hat{\beta}] = E[\sum_{it} \omega_{it}\beta_{it}]$ could come out positive

- In practice, of course, the weighting scheme could matter little

The good news with $E[\hat{\beta}] = E[\sum_{it} \omega_{it}\beta_{it}]$ is that if treatment effects are "roughly constant" we have $E[\hat{\beta}] \approx E[\sum_{it} \omega_{it}]\beta = \beta$

- More generally, we could have a lot of variation in β_{it} as long as it's uncorrelated with ω_{it} (which we directly observe)

The recent literature contains some examples of negative weights mattering, but we should as always be aware of selection bias...

- 'Metrics papers are easier to write with compelling applications...
- ...but top applied papers already pass a lot of robustness checks
- Not clear which effect dominates (we need a comprehensive survey!)

Solutions: Use “Clean Comparisons”

Callaway & Sant’Anna (2020), Sun & Abraham (2021), and de Chaisemartin & D’Haultfœuille propose alternative estimators that aggregate simple “clean” comparisons

- E.g. only compare “switchers” in time t to never-treated units or units not treated until time t to identify switcher ATEs
- Can choose how to average ATEs (as before)
- See e.g. the *csdid* Stata package for Callaway-Sant’Anna

Careful sample + regressor choice can automate things with OLS. Recall

$$Y_{it} = \beta D_{it} + \alpha_i + \tau_t + W_i' \gamma_t + \nu_{it}$$

identifies a variance-weighted average of within-group DiDs when W_i contains group indicators and $T = 2$

- Can use this to “stack” groups containing clean two-period comparisons (just don’t forget to cluster by repeated observations!)

Regression-Based Solutions

Borusyak et al. (2021), Wooldridge (2021), and Gardner (2021) propose "imputation" estimators that estimate counterfactual $Y_{it}(0)$ directly

- E.g. regress Y_{it} on unit and time FE in $D_{it} = 0$ cells, then average $Y_{it}(1) - \hat{Y}_{it}(0)$ in $D_{it} = 1$ cells (sound familiar?)
- See e.g. the *did_imputation* Stata package for BJS '21

These use more variation (i.e. more pre-treatment periods), so are likely to yield more precise estimates than Callaway & Sant'Anna

- They also work for any approach based on a model for $Y(0)$, not just TWFE / parallel trends
- Sometimes they can also be automated with OLS (see Wooldridge)

In practice, people often try multiple solutions (in their appendix...)

Outline

1. Negative Own-Treatment Weights✓

2. Contamination Bias

3. Application: Finkelstein et al. (2016)

Multiple Treatments

We've seen how “model-based” identification strategies yield regressions with (possibly) negative own-treatment weights

- Contrast to “design-based” selection-on-observables regressions, where convex weights are ensured so long as we flexibly control
- I.e., "Negative Weights are no Concern in Design-Based Specifications” (Borusyak and Hull, 2024)

Alas, negative weighting becomes more general w/ multiple treatments:

- Both model-based & design-based regressions can suffer from “contamination bias,” incorporating effects from other treatments

This can again be a big deal in theory ... but in practice?

- People study multiple-arm RCTs with regression all the time. How come they hadn't noticed this problem until recently?

Example: Event Study Regressions

Sun and Abraham (2021) study TWFE regressions of the form:

$$Y_{it} = \alpha_i + \tau_t + \sum_{g \in \mathscr{G}} \mu_g \mathbf{1}[t - G_i \in g] + \nu_{it}$$

where $\mathscr{G}$ collects disjoint sets of relative periods $\ell \in [-T, T]$

- E.g. $\mathscr{G} = [-T, \ldots, -2, 0, \ldots, T]$ for a fully dynamic "event study" with a never-treated control group
- Without never-treateds, need to drop two periods (Borusyak et al '21)

They show (basically by Frisch-Waugh-Lovell) that the μ_g generally mix together comparisons from other periods $g' \neq g$

- Under PT, this means μ_g incorporates ATT's from other periods
- Note this holds even for pre-period μ_g! We can find $\mu_g \neq 0$ even when there are no "true" pre-trends...

General Problem

Goldsmith-Pinkham et al. '22 show the general form of contamination bias

- Consider a partially linear model: $Y_i = \sum_k D_{ik}\beta_k + g(W_i) + U_i$
- Assume "exogeneity": $E[Y_i(k) \mid D_i, W_i] = E[Y_i(k) \mid W_i]$ for all k
- Suppose $g(\cdot)$ is flexible enough to span $E[Y_i(0) \mid W_i]$ (e.g. parallel trends) or propensity scores $p_k = E[D_{ik} \mid W_i]$ for all k

We show each regression coefficient β_k can then be decomposed:

$$\beta_k = E[\lambda_{kk}(W_i)\tau_k(W_i)] + \sum_{\ell \neq k} E[\lambda_{k\ell}(W_i)\tau_\ell(W_i)]$$

where $\tau_k(W_i) = E[Y_i(k) - Y_i(0) \mid W_i]$, $\lambda_{kk} = \frac{E[\tilde{D}_{ik} D_{ik} | W_i]}{E[\tilde{D}_{ik}^2]}$, $\lambda_{k\ell} = \frac{E[\tilde{D}_{ik} D_{i\ell} | W_i]}{E[\tilde{D}_{ik}^2]}$, and $\tilde{D}_{ik}$ is the residual from regressing D_{ik} on $g(W_i)$ *and* all other $D_{i,-k}$

- $E[\lambda_{kk}(W_i)] = 1$, $E[\lambda_{k\ell}(W_i)] = 0$. Further $\lambda_{kk}(W_i) \geq 0$ if $g(\cdot)$ spans p_k

Unpacking The Result

$$\beta_k = \underbrace{E[\lambda_{kk}(W_i)\tau_k(W_i)]}_{\text{Own treatment effect}} + \sum_{\ell \neq k} \underbrace{E[\lambda_{k\ell}(W_i)\tau_\ell(W_i)]}_{\text{Contamination bias}}$$

$E[\lambda_{kk}(W_i)] = 1$, $E[\lambda_{k\ell}(W_i)] = 0$. Further $\lambda_{kk}(W_i) \geq 0$ if (*) $g(\cdot)$ spans p_k

- (*) corresponds to a "design-based" regression: No negative own-treatment weights (generalizing Angrist '98 further)
- Unless $\lambda_{k\ell} = 0$ identically, there's potential for contamination bias

Intuition: FWL partials both $g(W_i)$ and $D_{i,-k}$ out of D_{ik} to estimate β_k

- The trick to Angrist '98 was that this auxilliary regression identified a CEF (the p-score). But here $E[D_{ik} \mid W_i, D_{i,-k}]$ is likely nonlinear
- FWL residual $\tilde{D}_{ik}$ is thus likely not mean-zero given $(W_i, D_{i,-k})$, so it "picks up" effects of other treatments D_{ik} given W_i

Is This a Problem?

In principle, contamination bias can apply to a large number of settings:

1. RCTs with multiple treatments and randomization strata
2. Selection-on-obs with multiple treatments (e.g. value-added' models)
3. TWFE with multiple treatments (e.g. "mover" regressions)
4. IV with multiple instruments (e.g. "examiner/judge" IVs)
5. Descriptive regressions on multiple variables (e.g. disparity analyses)

But again, whether there is a big problem depends on the empirical weights

- Since the CB weights average to zero, if they're uncorrelated with effect heterogeneity there's no issue
- The weights are identified; we can estimate them to diagnose bias

Solutions

Contamination bias comes from the FWL auxilliary regression not controlling “flexibly enough” for $(W_i, D_{i,-k})$ ... but we can fix that:

$$Y_i = \sum_k D_{ik}\beta_k + g(W_i) + \sum_k D_{ik}(q_k(W_i) - E[q_k(W_i)]) + U_i$$

The blue term captures non-linearities in (W_i, D_i)

- When $D_i \mid W_i$ is as-good-as-randomly assigned, β_k identifies the ATE of treatment k (Imbens and Wooldridge, 2009)
- Sun and Abraham (2021) propose similar interacted regressions to solve contamination in event studies (where W_i is event time)
- See our *multe* Stata package for automating this + other CB checks

This works in principle, but in practice can fail / lead to noisy estimates

- Key challenge: limited overlap ($p_k(W_i)$ may be close to zero or one)
- If CB is limited, an uninteracted regression is likely more efficient...

Illustration: Project STAR

Krueger (1999) studies the STAR RCT, which randomized 12k students in 80 public elementary schools in Tennessee (!) to one of 3 classroom types:

1. Regular-sized (20-25 students) – Control
2. Small (13-17 students) – Treatment 1
3. Regular-sized with a teaching aide – Treatment 2

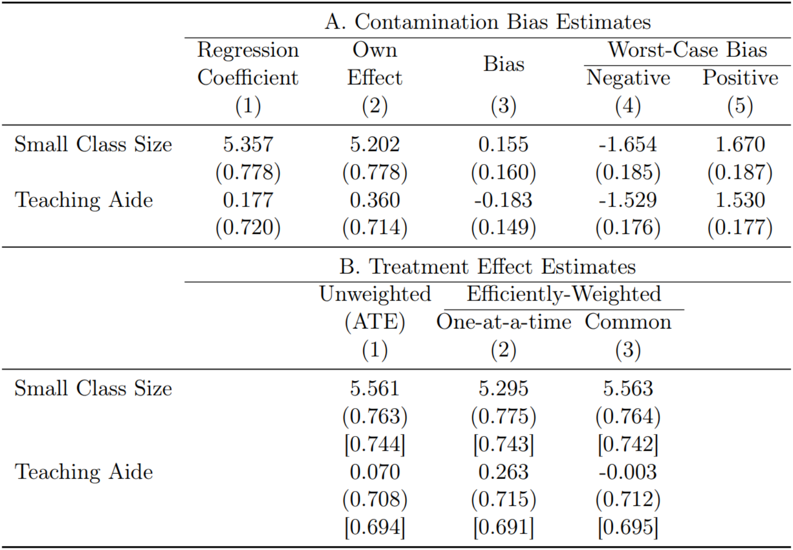
Kids were randomized within schools, so the propensity of assignment to each treatment varied by school

- Krueger thus estimates: $TestScore_i = \alpha_{school(i)} + \beta_1 D_{i1} + \beta_2 D_{i2} + \varepsilon_i$

We find significant *potential* for contamination bias: lots of treatment effect heterogeneity and variation in contamination weights

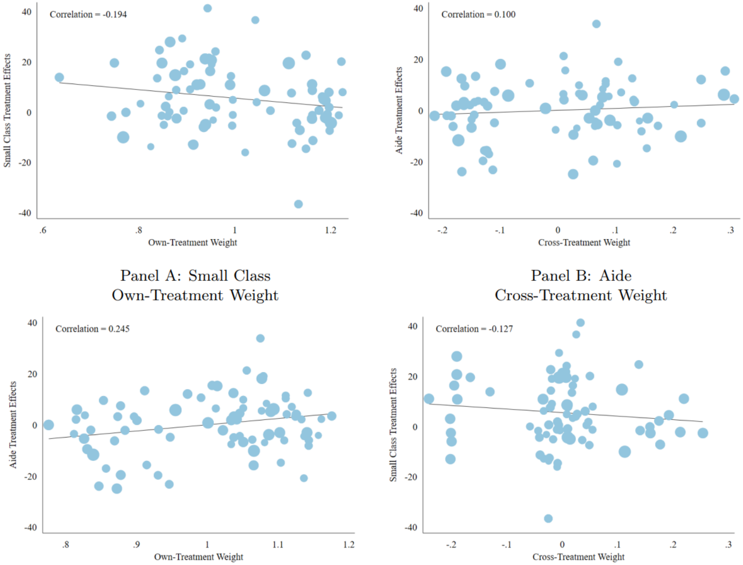
- But actual contamination bias is minimal: $Corr(effects, weights) \approx 0$

Project STAR, Revisited



	A. Contamination Bias Estimates				
	Regression	Own	Bias	Worst-Case Bias	
	Coefficient	Effect		Negative	Positive
	(1)	(2)	(3)	(4)	(5)
Small Class Size	5.357	5.202	0.155	-1.654	1.670
	(0.778)	(0.778)	(0.160)	(0.185)	(0.187)
Teaching Aide	0.177	0.360	-0.183	-1.529	1.530
	(0.720)	(0.714)	(0.149)	(0.176)	(0.177)
		B. Treatment Effect Estimates			
		Unweighted	Efficiently-Weighted		
		(ATE)	One-at-a-time	Common	
		(1)	(2)	(3)	
Small Class Size		5.561	5.295	5.563	
		(0.763)	(0.775)	(0.764)	
		[0.744]	[0.743]	[0.742]	
Teaching Aide		0.070	0.263	-0.003	
		(0.708)	(0.715)	(0.712)	
		[0.694]	[0.691]	[0.695]	

STAR Regression Weights vs. Treatment Effects



Panel A: Small Class
Own-Treatment Weight

Panel B: Aide
Cross-Treatment Weight

Panel C: Aide
Own-Treatment Weight

Panel D: Small Class
Cross-Treatment Weight

Outline

1. Negative Own-Treatment Weights✓
2. Contamination Bias✓
3. Application: Finkelstein et al. (2016)

Motivation: Geographic Variation in Healthcare Spending

A longstanding puzzle in health economics: why does utilization/spending differ so much across regions?

- In Medicare (65+), the highest-spending areas have twice the annual per-capita spending as the lowest spending areas (Austin et al 2020)
- Spending variation is not clearly correlated with health outcomes

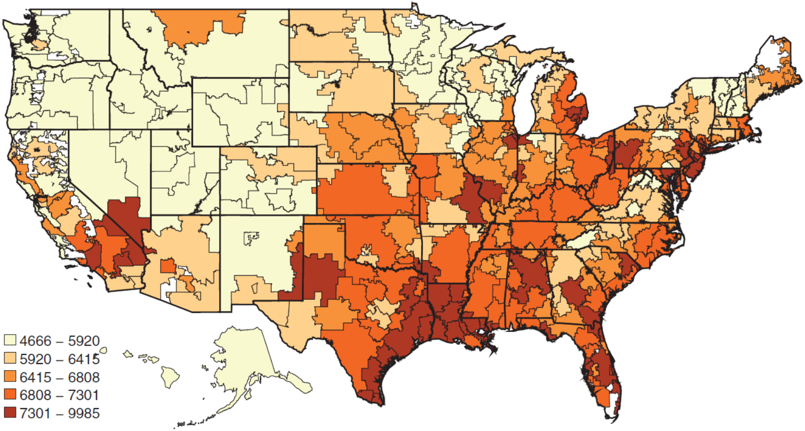
Two possible explanations: causal effects vs. selection bias

- Do regional conditions cause patients to spend more? ("supply")
- Or do high-spending patients sort to certain regions? ("demand")

If places drive meaningful spending differences with little to show for it, policies that standardize care can save several percentage points of GDP

- But if patients in high-utilization areas are sicker, or prefer more intensive care, such policies could be ineffective / counterproductive

Average Annual Per-Patient Medicare Spending ('98-'08)



Note: hospital referral regions (HRRs), defined by the Dartmouth Atlas

Identification Strategy: Patient Migration

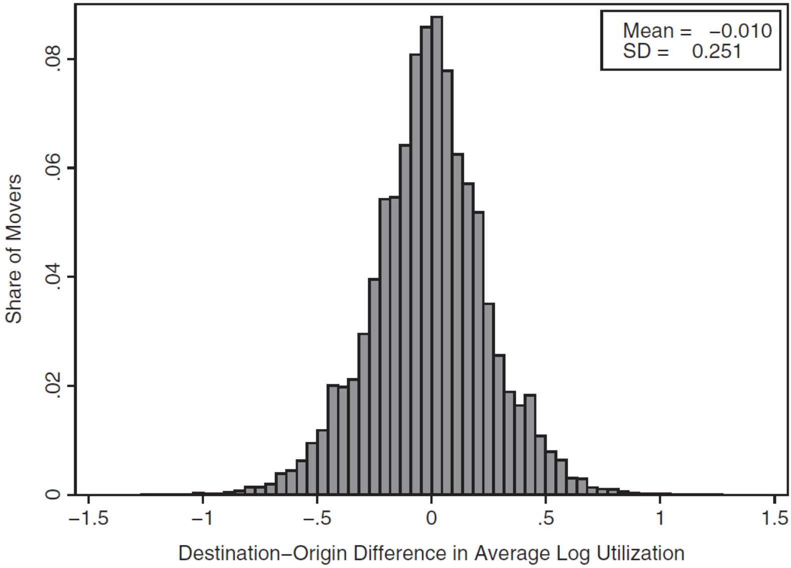
FGW's leverage the movement of Medicare beneficiaries across HRRs to disentangle place effects & patient sorting

- Thought experiment: if place effects are causal, patients moving from HRR *j* to HRR *k* should on avg see spending converge to region *k*'s
- Conversely, if regional variation is all due to sorting, patients should see no average change in spending following a move

It turns out beneficiaries move often & for arguably idiosyncratic reasons

- Most common (Health and Retirement Study): "to be near children/ relatives/friends" (41%) & "health problems or services" (13%)
- Importantly, FGW will leverage differential moves across HRRS with high/low spending – not directly compare movers and stayers
- Main concern: time-varying health shocks that lead to systematic moves towards/away from high-spending regions

HRR-Average Utilization Changes Across Movers



Causal Model and Event Study

FGW microfound a constant-effects causal model of (log) annual spending:

$$y_{it} = \alpha_i + \tau_t + \gamma_{j(it)} + x_{it}'\beta + \varepsilon_{it} \tag{1}$$

where $j(it)$ gives the HRR of patient i in year t

Main object of interest: avg share of utilization differences due to place fx:

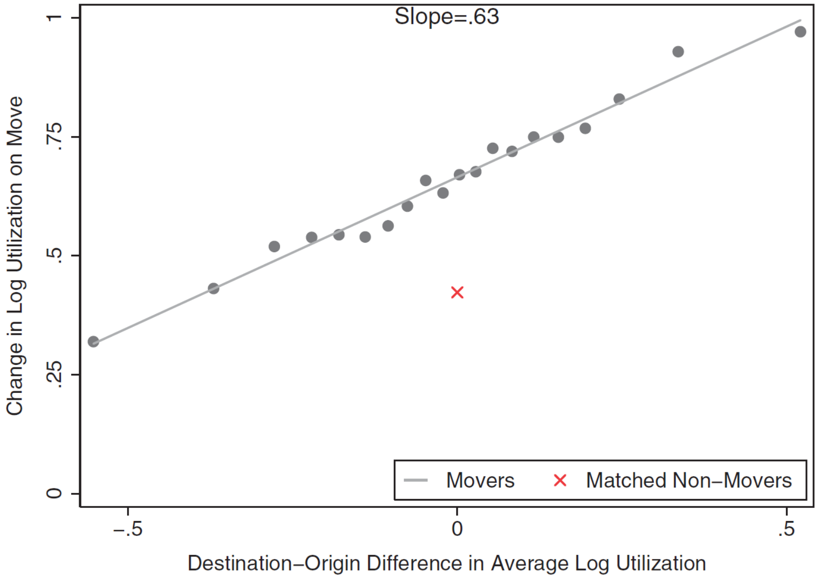
$$S = \sum_{j,j'} \omega_{j,j'} \left(\frac{\gamma_j - \gamma_j'}{\bar{y}_j - \bar{y}_j'} \right) \text{ for some weights } \omega_{j,j'}$$

Consider a patient i who moves from origin $o(i)$ to destination $d(i)$; let $r(i,t) = -T, \dots, 0, \dots, T$ index time relative to the move. Rewrite (1) as:

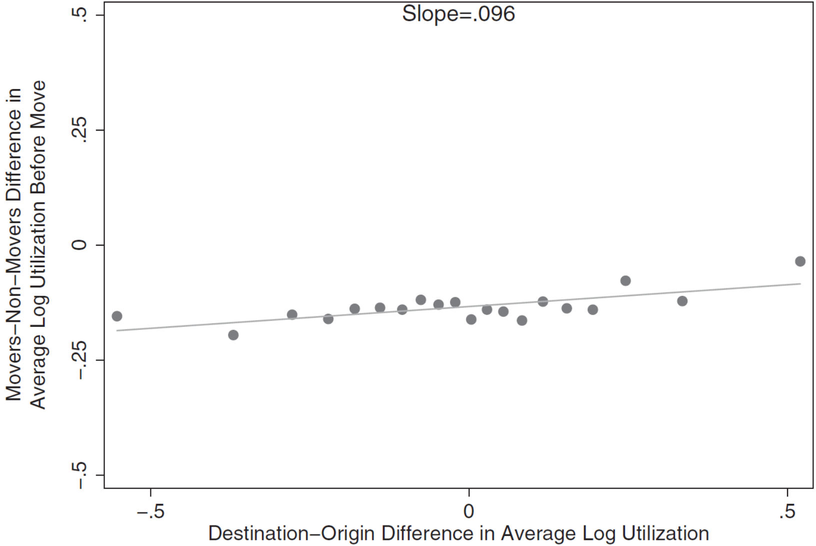
$$y_{it} = \underbrace{\alpha_i + \gamma_{o(i)}}_{\tilde{\alpha}_i} + \tau_t + \frac{\gamma_{d(i)} - \gamma_{o(i)}}{\bar{y}_{d(i)} - \bar{y}_{o(i)}} \mathbf{1}[r(i,t) > 0]\Delta_i + x_{it}'\beta + \varepsilon_{it}$$

for $\Delta_i = \bar{y}_{d(i)} - \bar{y}_{o(i)}$. This suggests a TWFE reg: of y_{it} on $\mathbf{1}[r(i,t) > 0]\Delta_i$

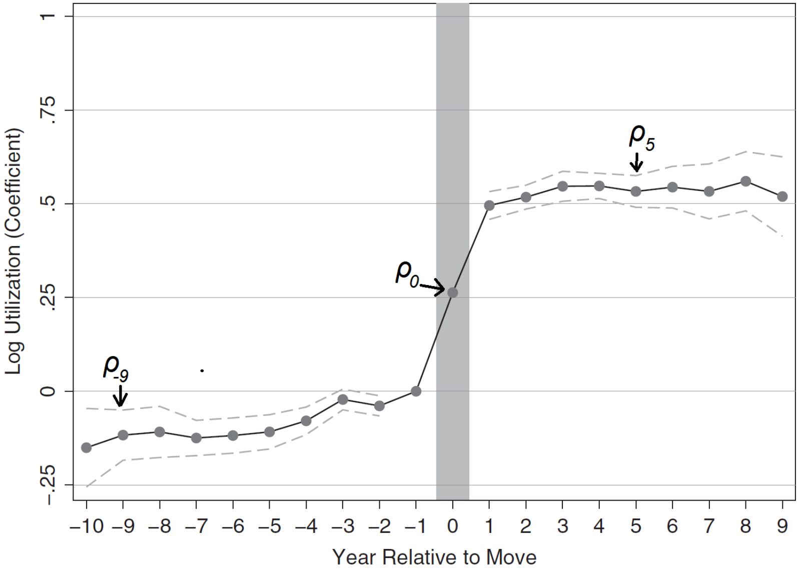
Motivating Diff-in-Diff



Motivating Pre-Trend Check



Main Event Study



$$y_{it} = \rho_{r(i,t)}\Delta_i + \phi_{r(i,t)} + \alpha_i + \tau_t + x_{it}'\beta + \varepsilon_{it}$$

Revisiting FGW '16, Post-Goodman-Bacon

The event study jump of 0.5 suggests around half of the observed variation in regional utilization $\bar{y}_j$ is causal / due to "supply-side" factors γ_j

- I.e. that $S_i = \frac{\gamma_{d(i)} - \gamma_{o(i)}}{\bar{y}_{d(i)} - \bar{y}_{o(i)}}$ is around 0.5, on average across movers i
- Pre-/post-trends look pretty good (though not perfect!)

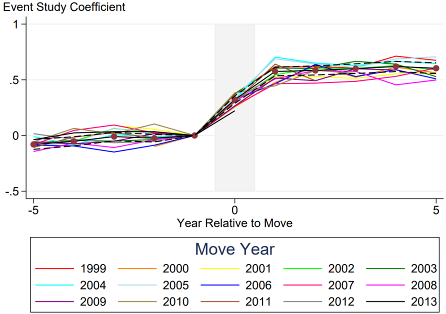
But as we now know, ρ_0 may identify a non-convex average of S_i

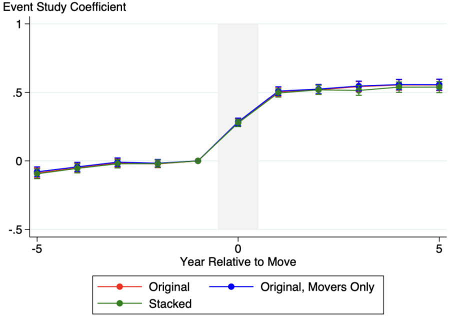
- "Staggered adoption" with no pure control group (non-movers)

Badinski et al. (2023), now older and wiser, check whether negative weights are actually an issue (as well as more substantive analyses!)

- Estimate the FGW event study separately by move year + stack
- Semi-pure control group: beneficiaries moving in any other year

Stacking Up Simpler Comparisons





Estimating Place Effects Themselves

FGW also directly estimate their constant-effects model (1):

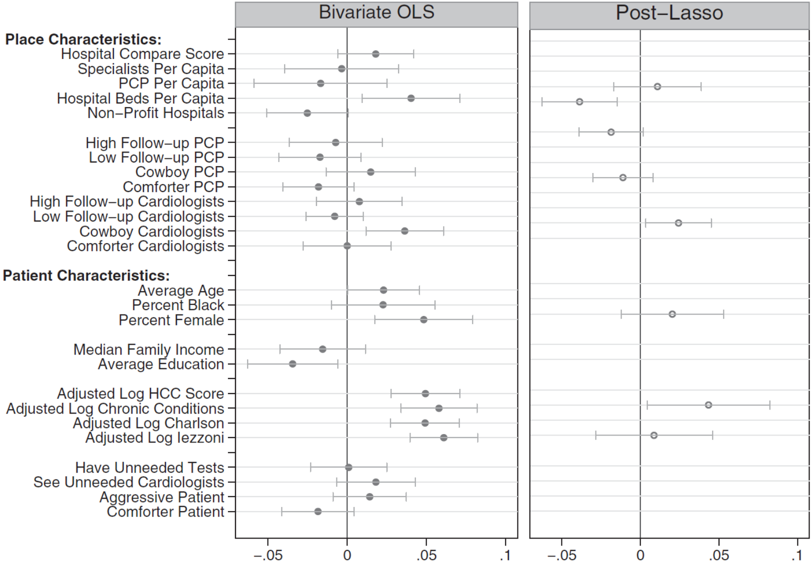
$$y_{it} = \alpha_i + \tau_t + \gamma_{j(it)} + x_{it}'\beta + \varepsilon_{it}$$

They correlate the estimates of γ_j with various place observables, and use them for certain partial-equilibrium counterfactuals

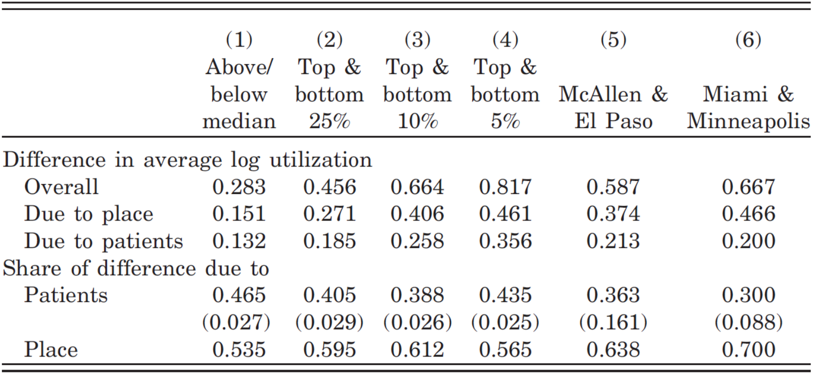
Here a concern is contamination bias: 306 HRR treatments + TWFE

- Hull (2018) formalizes this concern and proposes an alternative "mover average treatment effect" (MATE) estimator
- Similar to Callaway and Sant'Anna, but for multiple treatments; LMK if you'd ever like to work with a "beta" Stata package

Place Effect Correlates

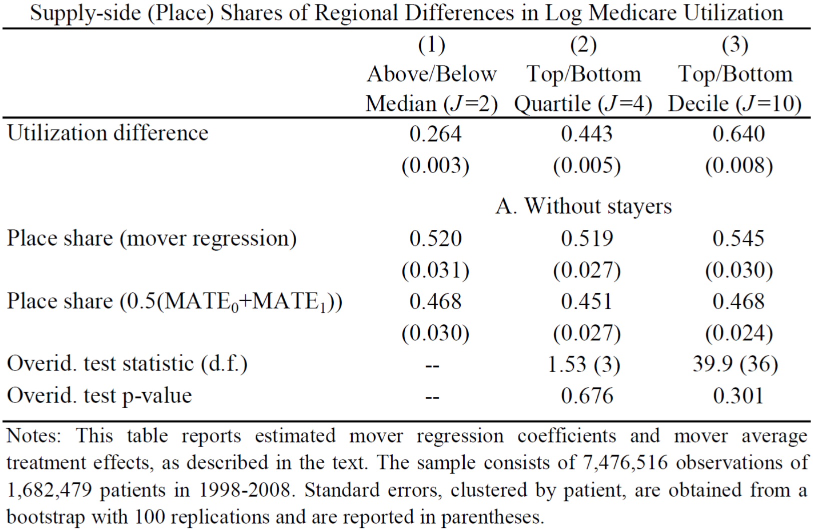


Decomposing Geographic Variation with Place Effects



	(1) Above/ below median	(2) Top & bottom 25%	(3) Top & bottom 10%	(4) Top & bottom 5%	(5) McAllen & El Paso	(6) Miami & Minneapolis
Difference in average log utilization						
Overall	0.283	0.456	0.664	0.817	0.587	0.667
Due to place	0.151	0.271	0.406	0.461	0.374	0.466
Due to patients	0.132	0.185	0.258	0.356	0.213	0.200
Share of difference due to						
Patients	0.465	0.405	0.388	0.435	0.363	0.300
	(0.027)	(0.029)	(0.026)	(0.025)	(0.161)	(0.088)
Place	0.535	0.595	0.612	0.565	0.638	0.700

Aggregating Simpler Comparisons



Supply-side (Place) Shares of Regional Differences in Log Medicare Utilization

	(1)	(2)	(3)
	Above/Below	Top/Bottom	Top/Bottom
	Median (J=2)	Quartile (J=4)	Decile (J=10)
Utilization difference	0.264	0.443	0.640
	(0.003)	(0.005)	(0.008)
		A. Without stayers	
Place share (mover regression)	0.520	0.519	0.545
	(0.031)	(0.027)	(0.030)
Place share (0.5(MATE$_0$+MATE$_1$))	0.468	0.451	0.468
	(0.030)	(0.027)	(0.024)
Overid. test statistic (d.f.)	--	1.53 (3)	39.9 (36)
Overid. test p-value	--	0.676	0.301

Notes: This table reports estimated mover regression coefficients and mover average treatment effects, as described in the text. The sample consists of 7,476,516 observations of 1,682,479 patients in 1998-2008. Standard errors, clustered by patient, are obtained from a bootstrap with 100 replications and are reported in parentheses.

My Takeaways

Finkelstein et al. (2016) is a great paper, in a few distinct ways:

1. Tackles a big + longstanding problem in a new + creative way
2. Carefully discusses model assumptions (e.g. constant effects)
3. Builds + illustrates intuition for identification w/simple comparisons
4. Results are super robust, even with respect to the neg. weight lit.

Their “mover” approach seems under-utilized, within and outside of health

- Cantoni & Pons ‘22 use it to study regional diffs in voting behavior
- Your colleague Mauricio Caceres Bravo is using it to study prison fx
- I suspect there are other arbitrage opportunities (happy to discuss!)