Advanced Quantitative Methods

Difference-in-differences: Empirical practice & staggered designs

Instructor: Gregory Eady
Office: 18.2.10
Office hours: Fridays 13-15

Today

- Difference-in-differences in practice
- Staggered difference-in-differences

What do we need to look out for in difference-in-differences designs?

1. Parallel trends is a fundamental assumption to diff-in-diff designs
   - The legitimacy of your results relies on this assumption
   - Testing this rigorously is central to the legitimacy of the results
2. When treatment assignment varies (a "staggered" difference-in-differences design), two-way fixed effects can give biased estimates
   - Fortunately a recent paper by Chiu et al. (2023) shows in practice that corrections to these designs do not affect the substantive conclusions
   - However, we should aim to things correctly

Navigating Potential Pitfalls in Difference-in-Differences Designs: Reconciling Conflicting Findings on Mass Shootings' Effect on Electoral Outcomes

HANS J. G. HASSELL *Florida State University, United States*
JOHN B. HOLBEIN *University of Virginia, United States*

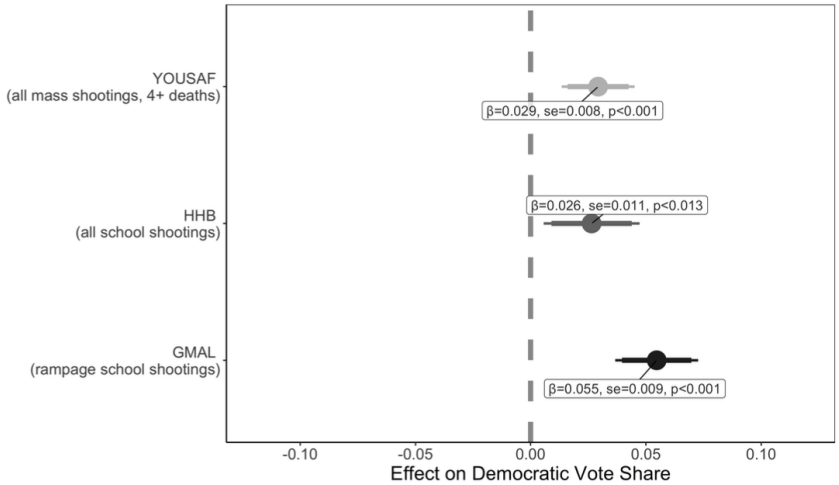
Work on the electoral effects of gun violence in the U.S. relying on difference-in-differences designs has produced findings ranging from null to substantively large effects. However, as difference-in-difference designs, on which this research relies, have exploded in popularity, scholars have documented several methodological issues including potential violations of parallel-trends and unaccounted for treatment effect heterogeneity. These pitfalls (and their solutions) have not been fully explored in political science. We apply these advancements to the unresolved debate on gun violence's effects on U.S. electoral outcomes. We show that studies finding a large positive effect of gun violence on Democratic vote shares are a product of a failure to properly specify difference-in-differences models when underlying assumptions are unlikely to hold. Once these biases are corrected, shootings show little evidence of sparking large electoral change. Our work clarifies an unresolved debate and provides a cautionary guide for scholars currently employing difference-in-differences designs.

Do mass shootings affect voting behavior?

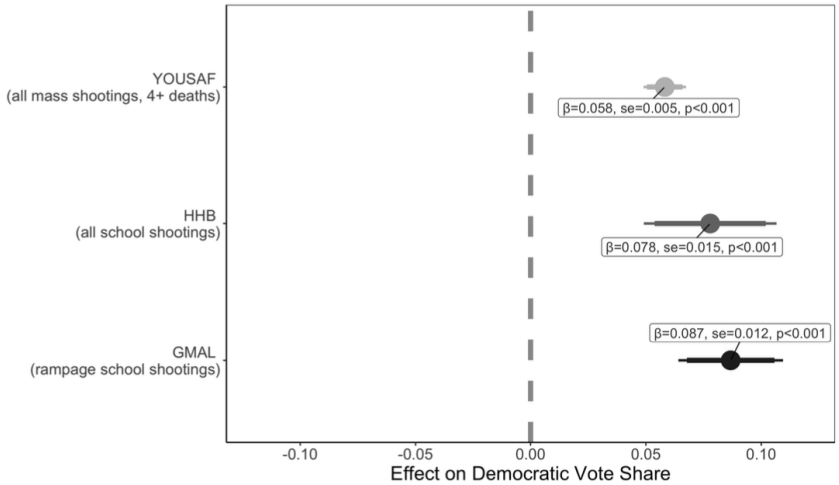
- Garcia-Montoya et al. 2022, and Yousaf 2021 find that mass shootings increase support for Democratic candidates
- The effects they find are large...

FIGURE 1. Differences in Previous Studies' Estimated Effect of Mass Shootings on Election Outcomes Are Not Driven by Data Choices

(a) Only Elections in Counties in Election Cycle When Shootings Occur are Treated



(b) All Elections in Counties After Shootings Occur are Treated



These effect sizes are large

- By comparison, a 6 standard deviation shift in television advertising leads to only a 0.5-point change in two-party vote share
- The effect sizes, in other words, are unrealistically large
- So what's going on?

In difference-in-differences designs we don't care about level differences, we care about parallel trends

- It is okay if treatment and control counties differ (e.g. in racial make-up, income, history of Democratic/Republican voting, occupational distribution, etc.)
- What matters is that treatment and control counties' Democratic vote share would counterfactually move in parallel were it not for a mass shooting

Authors often point this out:

Nonetheless, it is important to note that the event study design used to examine short-run effects does not require unobserved factors correlated with racism to be uncorrelated with treatment. The model instead relies on a parallel trends assumption and accounts for level differences between treatment and control areas by examining changes in acts of racial violence before and after the film's arrival. In particular, I estimate the following equation on weekly panel data from 1913 to 1922 for all counties in the continental United States:[13]

$$y_{c,t} = \delta_c + \lambda_{s,t} + \sum_{\tau=-6}^{6} \beta_\tau Show_\tau + \epsilon_{c,t}. \quad (1)$$

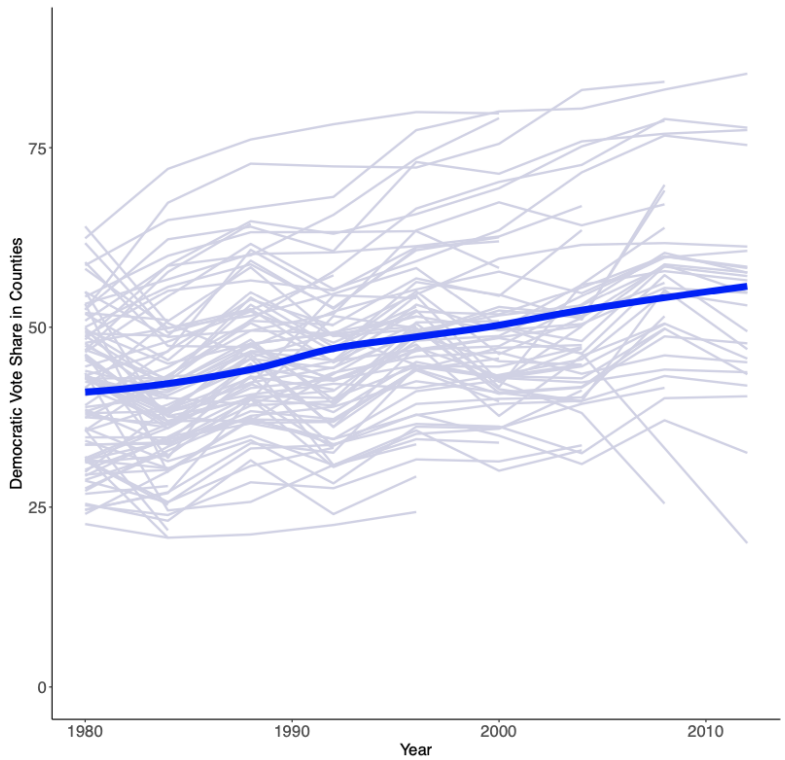
How can we justify the parallel trends assumption in practice?

1. Present a basic descriptive figure showing the pre-treatment trends for units in both the control and treatment groups...

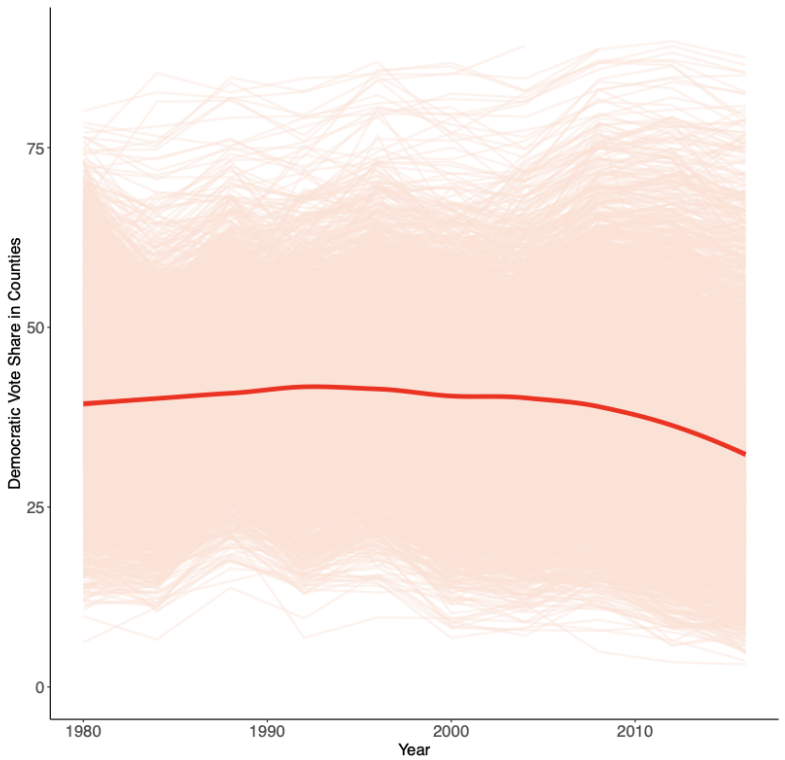
Democratic vote share in the years prior to a shooting for treated counties (left) and control counties (right):

FIGURE 2. Trends in Presidential Vote in Counties With Mass Shootings Prior to Shootings, Compared to Trends in Counties Without Shootings

(a) Pre-treatment Trends in Democratic Vote in Shooting Counties



(b) Trends in Democratic Vote in Non- Shooting Counties



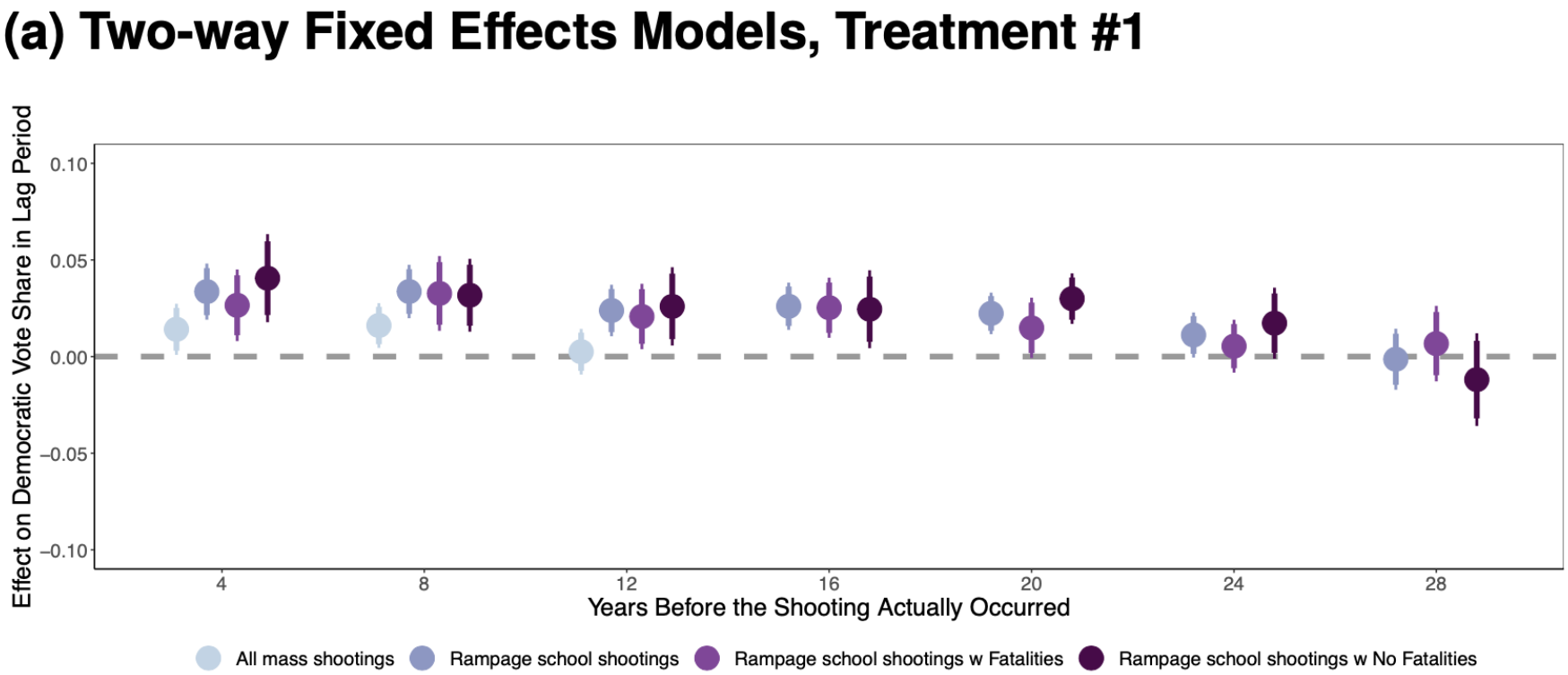
How can we justify the parallel trends assumption in practice?

1. Present a basic descriptive figure showing the pre-treatment trends for both units those that are in both the control and treatment groups
2. **Run your two-way fixed effects model on a lagged versions of your outcome. Essentially a placebo test: the future should not affect the past...**

If in 1984 there will be a mass shooting, this should not affect Democratic vote share in 1980...

county_id	year	treatment	treatment_lag1	treatment_lag2	...
1	1980	0	0	1	...
1	1984	0	1	0	...
1	1988	1	0	0	...
1	1992	0	0	1	...
1	1996	0	1	0	...
1	2000	1	0	0	...
1	2004	0	0	0	...
1	2008	0	0	0	...
1	2012	0	0	0	...
1	2016	0	0	0	...

The future shouldn't affect the past like this:

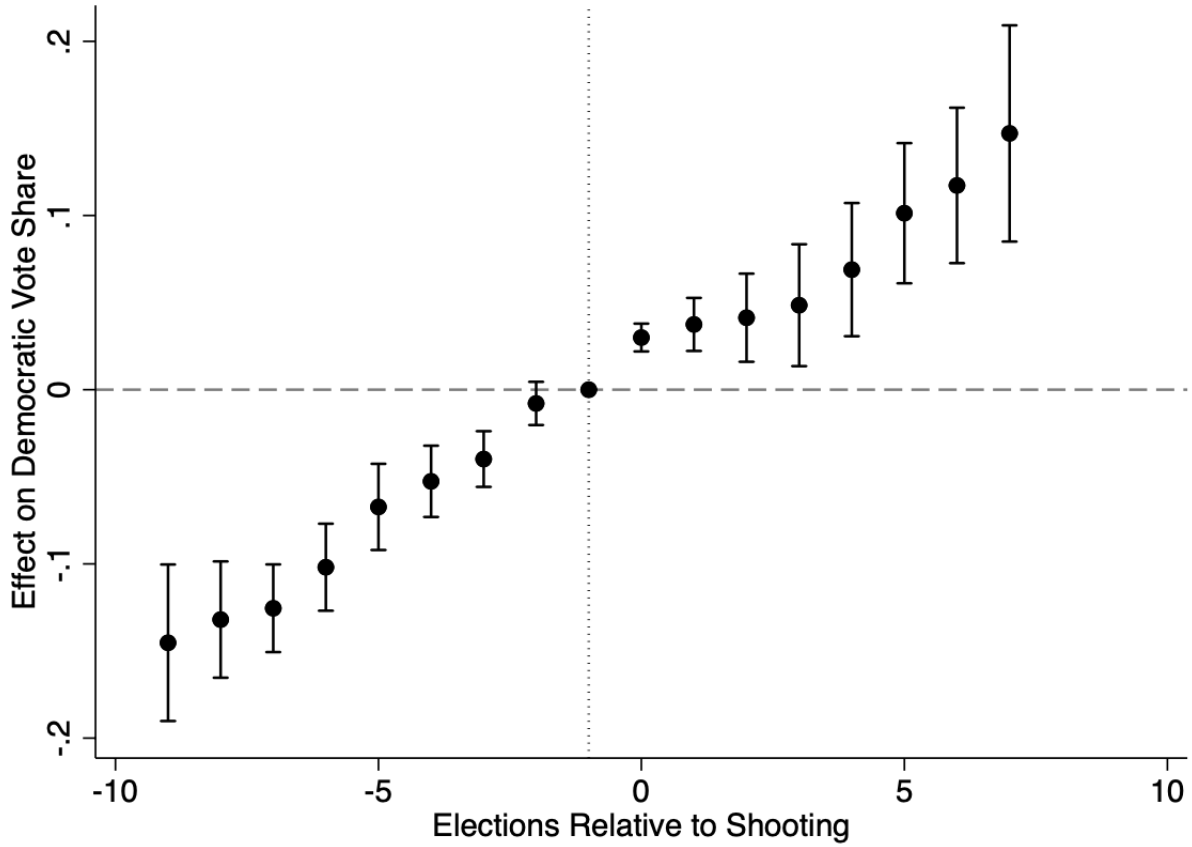


How can we justify the parallel trends assumption in practice?

1. Present a basic descriptive figure showing the pre-treatment trends for both units those that are in both the control and treatment groups
2. Run your two-way fixed effects model on a lagged versions of your outcome. Essentially a placebo test: the future should not affect the past...
3. **Check for parallel trends with an event study...**

Pre-treatment differences between the control and treatment units should not diverge like this:

FIGURE 4. Event-study Estimates Show that TWFE Fails to Account for Pre-Treatment Trends



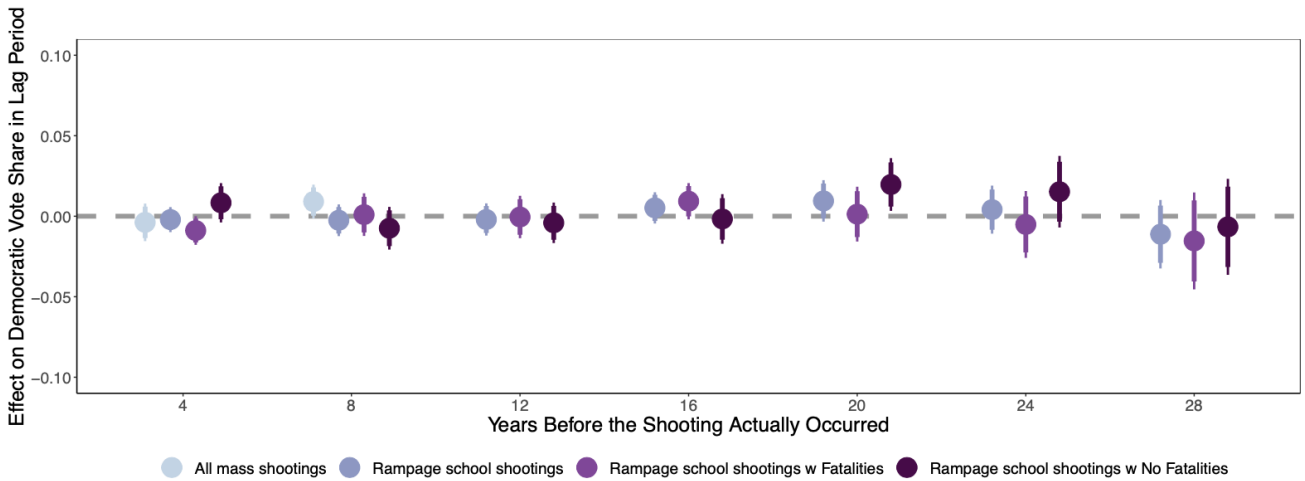
How can we justify the parallel trends assumption in practice?

1. Present a basic descriptive figure showing the pre-treatment trends for both units those that are in both the control and treatment groups
2. Run your two-way fixed effects model on a lagged versions of your outcome (essentially a placebo test—the future should not affect the past)
3. Check for parallel trends with an event study
4. **Test robustness with unit-level time trends...**

The future no longer affects the past when unit-level time trends are included:

With Time Trends

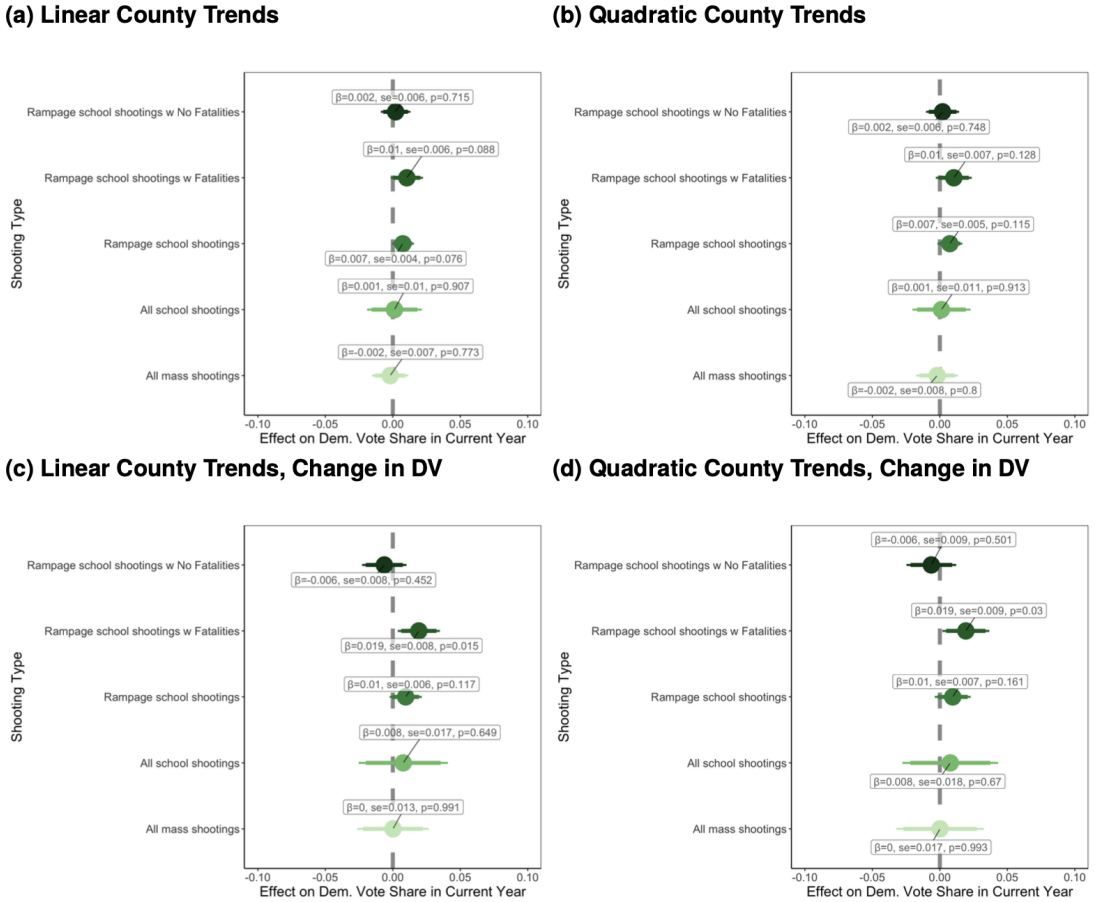
(c) Linear County Trends, Treatment #1



$$y_{it} = \alpha_i + \gamma_t + \beta D_{it} + \underbrace{\lambda_i t}_{\text{unit time trends}} + \epsilon_{it}$$

And in the main results, mass shootings are no longer estimated to affect Democratic vote share:

FIGURE 5. Effects of Mass Shootings on Presidential Elections After Absorbing County-Specific Trends



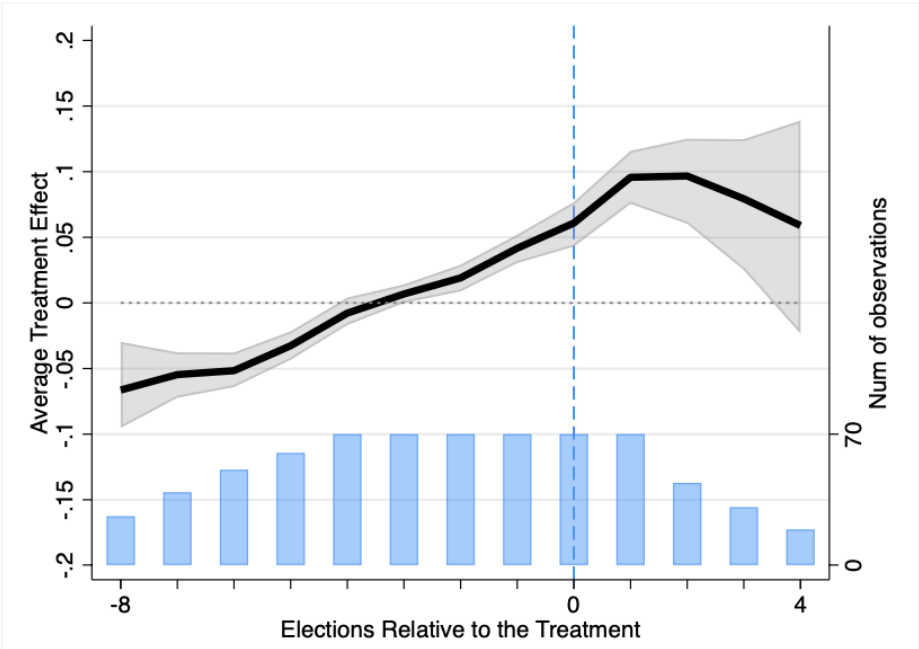
How can we justify the parallel trends assumption in practice?

1. Present a basic descriptive figure showing the pre-treatment trends for both units those that are in both the control and treatment groups
2. Run your two-way fixed effects model on a lagged versions of your outcome (essentially a placebo test—the future should not affect the past)
3. Check for parallel trends with an event study
4. Test robustness with unit-level time trends
5. **Newer procedures with the flavor of a synthetic control...**

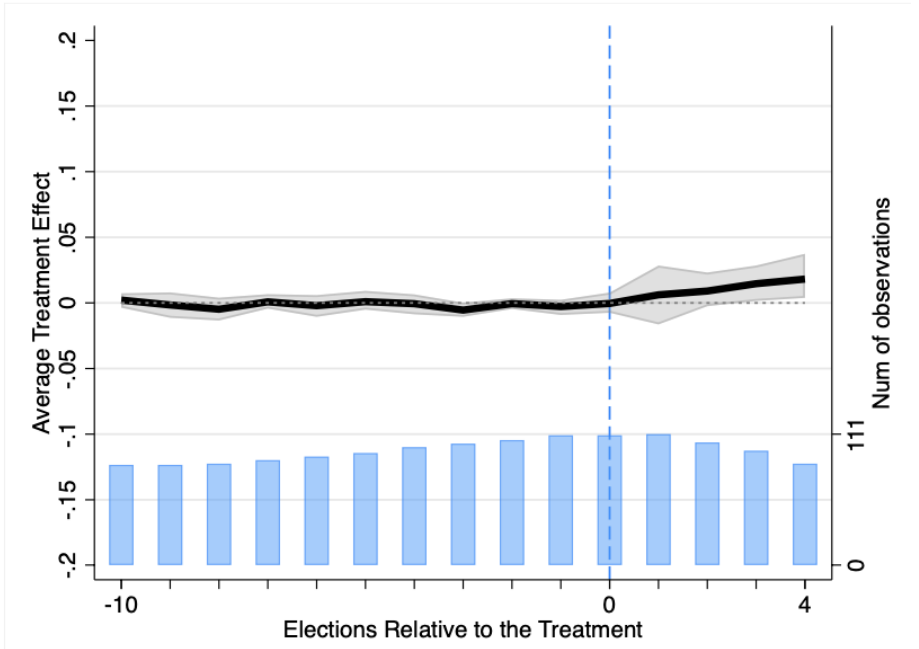
Interactive fixed effects/matrix completion methods impute the counterfactual for treated units

FIGURE 7. Liu et al. (2021) Interactive Fixed Effects Counterfactual Estimator

(a) Two-way Fixed Effects



(b) Interactive Fixed Effects (1)



In classical difference-in-differences all treated units are treated simultaneously:

id	year	treatment	outcome
1	2000	0	34
1	2001	0	25
1	2002	1	27
1	2003	1	30
1	2004	1	24
2	2000	0	78
2	2001	0	68
2	2002	1	68
2	2003	1	71
2	2004	1	89
3	2000	0	20
3	2001	0	13
3	2002	0	9
3	2003	0	30
3	2004	0	26

In this case, we can use our two-way fixed effects model without much troubles (aside from the usual concerns):

$$y_{it} = \alpha_i + \gamma_t + \beta \text{Treatment}_{it} + \epsilon_{it}, \tag{1}$$

where β is our estimate of the difference-in-differences comparing the pre- to the post-treatment period

A different case, however, is when treatment is staggered:

id	year	treatment	outcome
1	2000	0	34
1	2001	1	25
1	2002	1	27
1	2003	1	30
1	2004	1	24
2	2000	0	78
2	2001	0	68
2	2002	0	68
2	2003	1	71
2	2004	1	89
3	2000	0	78
3	2001	0	68
3	2002	0	68
3	2003	0	71
3	2004	0	89

First, for an event study, we need to create a new variable that captures time to treatment for each unit

id	year	treatment	outcome	time_to_treatment (τ)
1	2000	0	34	0
1	2001	1	25	1
1	2002	1	27	2
1	2003	1	30	3
1	2004	1	24	4
2	2000	0	78	-2
2	2001	0	68	-1
2	2002	0	68	0
2	2003	1	71	1
2	2004	1	89	2
3	2000	0	78	0
3	2001	0	68	0
3	2002	0	68	0
3	2003	0	71	0
3	2004	0	89	0

Second, unfortunately there is an additional problem related to estimation of staggered difference-in-differences models

- Goodman-Bacon (2021) shows that in a staggered design, a two-way fixed effects estimate is a weighted average of all possible 2×2 difference-in-differences in the panel data...

Imagine three units in a dataset:

1. Treated never, 2. Treated early, 2. Treated late

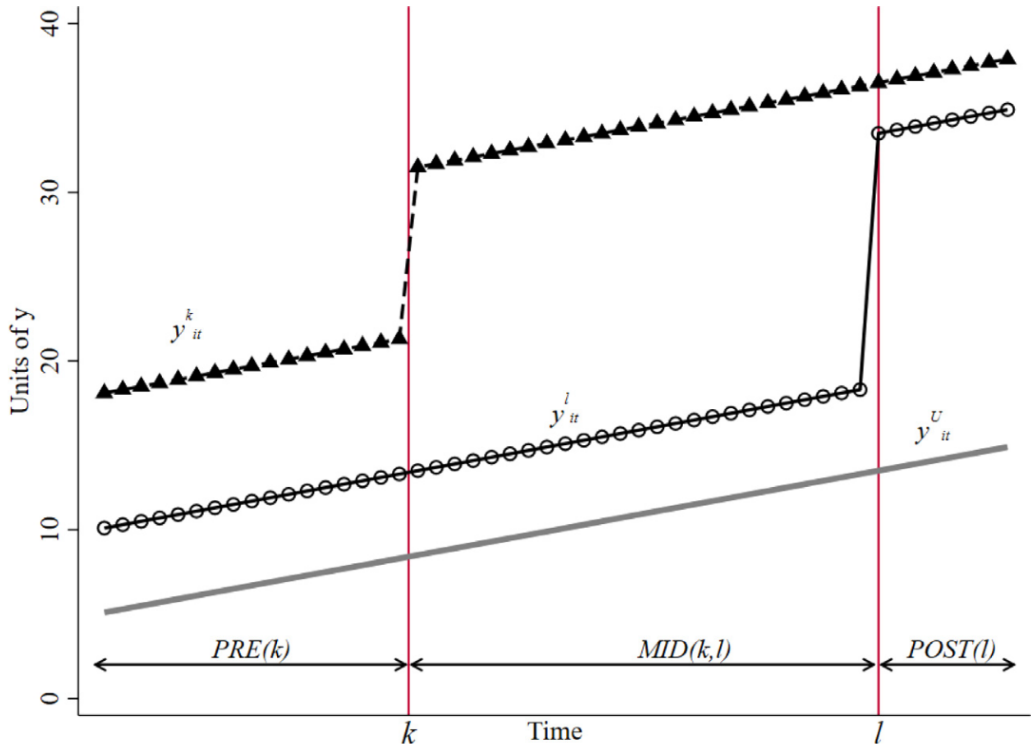


Fig. 1. Difference-in-Differences with variation in treatment Timing: Three groups. Notes: The figure plots outcomes in three timing groups: an untreated group, U; an early treatment group, k, which receives a binary treatment at $k = \frac{34}{100}T$; and a late treatment group, ℓ, which receives the binary treatment at $\ell = \frac{85}{100}T$. The x-axis notes the three sub-periods: the pre-period for timing group k, $[1, k-1]$, denoted by $PRE(k)$; the middle period when timing group k is treated and timing group ℓ is not, $[k, \ell-1]$, denoted by $MID(k, \ell)$; and the post-period for timing group ℓ, $[\ell, T]$, denoted by $POST(\ell)$. The treatment effect is 10 in timing group k and 15 in timing group ℓ.

From these, we can form four 2×2 difference-in-differences comparisons:

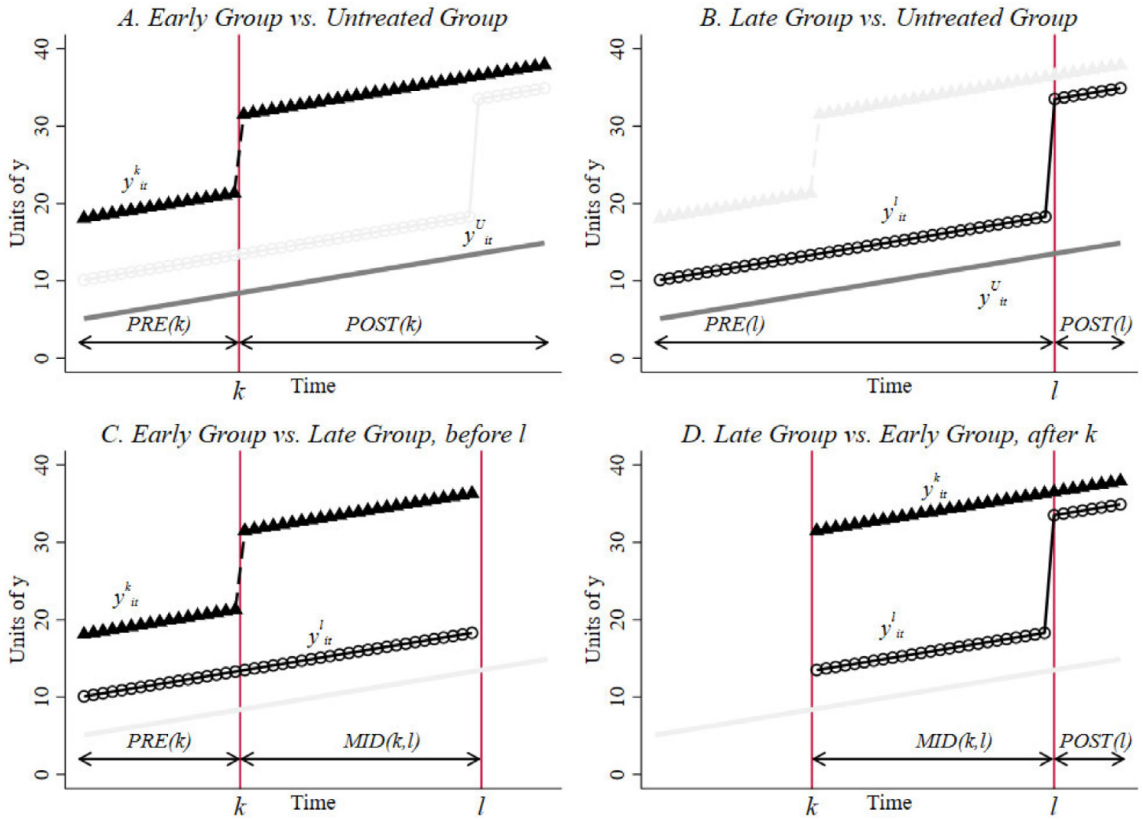


Fig. 2. The four simple (2x2) difference-in-differences estimates in the three group case. Notes: The figure plots outcomes for the subsamples that generate the four simple 2x2 difference-in-difference estimates in the three timing group case from Fig. 1. Each panel plots the data structure for one 2x2 DD. Panel A compares early treated units to untreated units ($\hat{\beta}^{DD}_{kU}$); panel B compares late treated units to untreated units ($\hat{\beta}^{DD}_{\ell U}$); panel C compares early treated units to late treated units during the late timing group's pre-period ($\hat{\beta}^{DD,k}_{k\ell}$); panel D compares late treated units to early treated units during the early timing group's post-period ($\hat{\beta}^{DD,\ell}_{k\ell}$). The treatment times mean that $\overline{D}_k = 0.67$ and $\overline{D}_\ell = 0.16$, so with equal group sizes, the decomposition weights on the 2x2 estimate from each panel are 0.365 for panel A, 0.222 for panel B, 0.278 for panel C, and 0.135 for panel D.

Imagine calculating all difference-in-differences estimates from all possible 2×2 comparisons:

- Goodman-Bacon (2021) shows that a two-way fixed effects estimate is a weighted average of all of these estimates
- He shows that this weighted average recovers the ATT only when the treatment effect is:
  1. Equivalent for units
  2. Does not vary within-unit over time
- Unfortunately, these are rather strong assumptions
- Goodman-Bacon (2021) also shows that certain units gets weighted more as treatment units than others...

Units treatment in the middle of the time period are weighted more than those treated early or late:

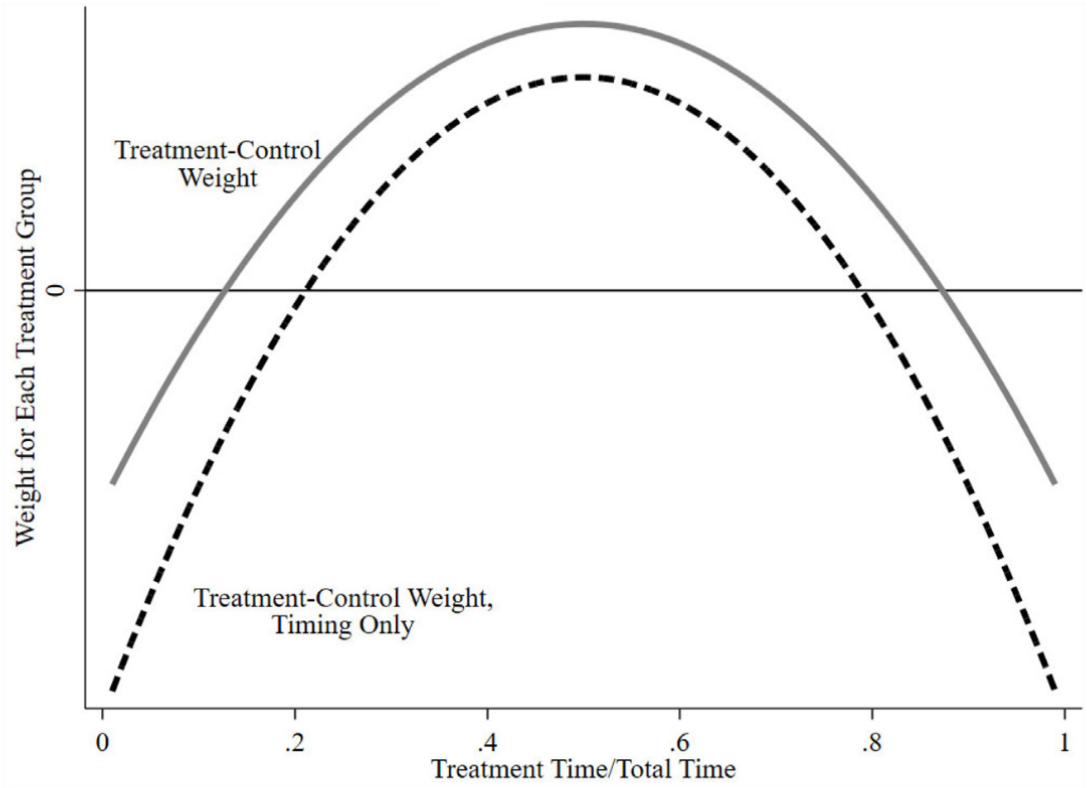


Fig. 4. Weighted common trends: The treatment/control weights as a function of the share of time spent under treatment. Notes: The figure plots the weights that determine each timing group's importance in the weighted common trends expression in Eqs. (16) and (17).

Note: "Timing only" refers to designs where all units are eventually treated

Fortunately, there many new estimators to fix these issues:

- Liyang Sun and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics*. **R library**: `fixest`
- Brantly Callaway and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics*. **R library**: `did`
- Kirill Borusyak, Xavier Jaravel, and Jann Spiess. 2021. "Revisiting Event Study Designs: Robust and Efficient Estimation." Unpublished manuscript. R library: `didimputation`
- Eli Ben-Michael, Avi Feller, and Jesse Rothstein. 2021. "The Augmented Synthetic Control Method." *Journal of the American Statistical Association*. **R library**: `augsynth`
- Licheng Liu, Ye Wang, and Yiqing Xu. 2021. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data." Unpublished manuscript. **R library**: `fect`
- Kosuke Imai, In Song Kim, and Erik H. Wang. 2023. "Matching Methods for Causal Inference with Time-Series Cross-Sectional Data." *American Journal of Political Science*. **R library**: `PanelMatch`

For implementation and comparison, see:
`https://github.com/fhollenbach/did_compare/blob/main/ComparingDiD.md`

Are lots of ways to correct for these issues. The most intuitive might be Callaway and Sant'Anna's (2021):

- Separate data into cohorts
  - i.e. Those units treated at the same point in time
- For each cohort, compare them to only the as-yet-untreated units
  - i.e. Compare treatment units in a cohort to "clean" controls
  - e.g. Clean controls look like Panels A, B and C on Slide 28
- Calculate a set of event study estimates separately for each cohort
  - i.e. We obtain a set of event study estimates *per cohort*
- Average over these estimates to calculate a single set of event study estimates (aggregated across all cohorts), or to calculate an overall ATT

Exercise is a replication of Grumbach (2023)

American Political Science Review (2023) 117, 3, 967–984
doi:10.1017/S0003055422000934

Laboratories of Democratic Backsliding

JACOB M. GRUMBACH *University of Washington, United States*

The Trump presidency generated concern about democratic backsliding and renewed interest in measuring the national democratic performance of the United States. However, the US has a decentralized form of federalism that administers democratic institutions at the state level. Using 51 indicators of electoral democracy from 2000 to 2018, I develop a measure of subnational democratic performance, the State Democracy Index. I then test theories of democratic expansion and backsliding based in party competition, polarization, demographic change, and the group interests of national party coalitions. Difference-in-differences results suggest a minimal role for all factors except Republican control of state government, which dramatically reduces states' democratic performance during this period. This result calls into question theories focused on changes within states. The racial, geographic, and economic incentives of groups in national *party coalitions may instead determine the health of democracy in the states.*

Main results are standard two-way fixed effects models (Table 1)

TABLE 1. Explaining Dynamics in State-Level Democracy

	Outcome: State Democracy Score						
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Competition	0.200			0.170	0.194	0.169	0.134
	(0.107)			(0.099)	(0.099)	(0.106)	(0.114)
Polarization		0.017		0.024	0.037	0.027	0.028
		(0.131)		(0.119)	(0.111)	(0.126)	(0.121)
Republican			−0.462**	−0.444**	−0.435**	−0.443**	−0.475**
			(0.162)	(0.159)	(0.162)	(0.154)	(0.183)
Competition × Polarization					0.082		
					(0.066)		
Polarization × Republican						−0.013	
						(0.198)	
Competition × Republican							0.110
							(0.206)
Constant	−0.707***	−0.683***	−0.532***	−0.535***	−0.544***	−0.533***	−0.532***
	(0.068)	(0.116)	(0.093)	(0.134)	(0.136)	(0.139)	(0.135)
State FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	833	833	833	833	833	833	833
R^2	0.683	0.676	0.699	0.704	0.706	0.704	0.705
Adj. R^2	0.656	0.648	0.673	0.679	0.680	0.678	0.679

Note: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Main results are standard two-way fixed effects models (Table 2)

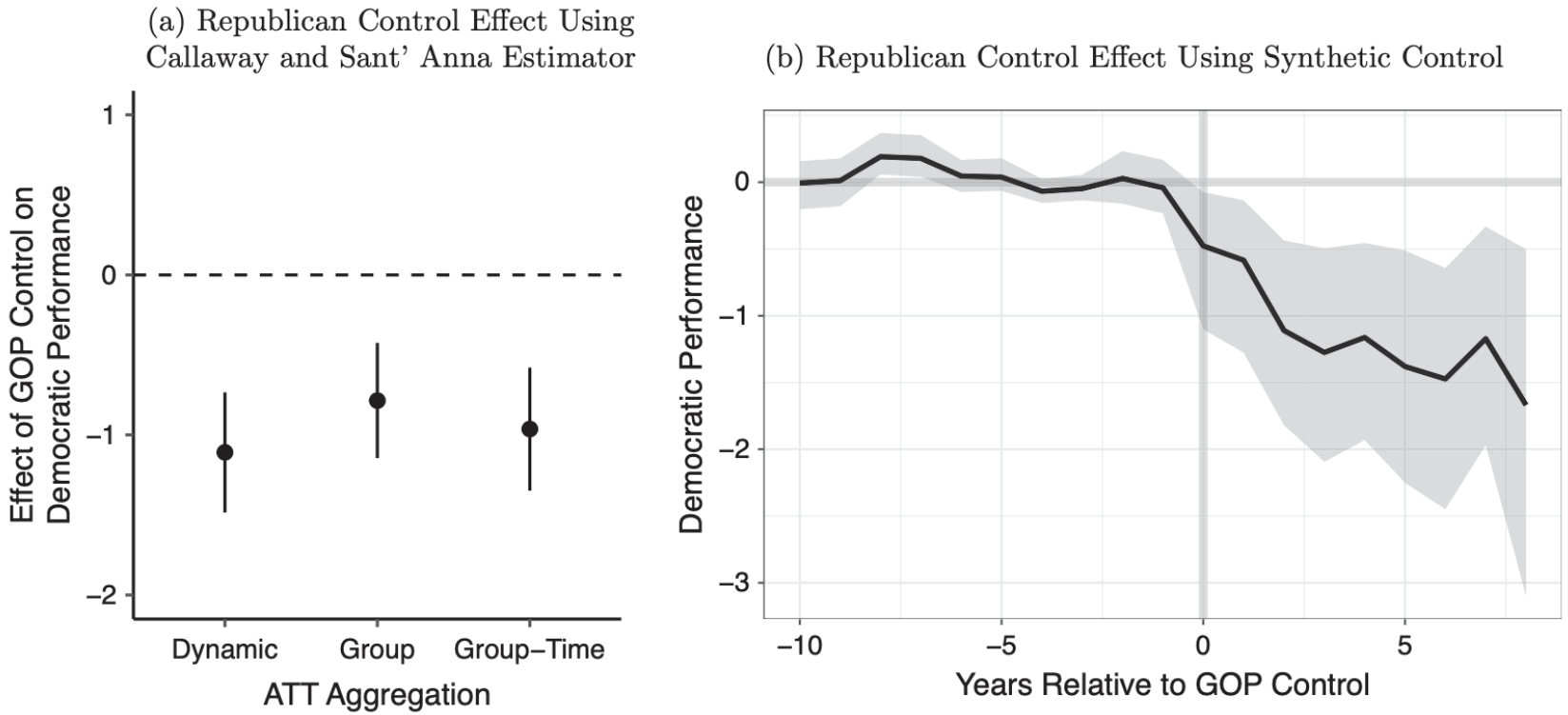
TABLE 2. Racial Demographic Change and State Democracy

	Outcome: State Democracy Score			
	Model 1	Model 2	Model 3	Model 4
Δ % Black	−0.012	−0.105	0.058	0.071
	(0.249)	(0.266)	(0.374)	(0.253)
Δ % Latino	−0.019	0.020	−0.010	−0.174
	(0.202)	(0.189)	(0.207)	(0.186)
Competition		0.317		
		(0.165)		
Polarization			0.007	
			(0.199)	
Republican				−0.726**
				(0.252)
Δ % Black × Competition		0.014		
		(0.280)		
Δ % Latino × Competition		−0.140		
		(0.095)		
Δ % Black × Polarization			0.094	
			(0.226)	
Δ % Latino × Polarization			−0.029	
			(0.130)	
Δ % Black × Republican				−0.140
				(0.280)
Δ % Latino × Republican				−0.325*
				(0.156)
Constant	−0.673***	−0.670***	−0.694***	−0.358*
	(0.166)	(0.166)	(0.169)	(0.177)
State FEs	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes
N	833	833	833	833
R^2	0.676	0.685	0.676	0.705
Adj. R^2	0.648	0.657	0.647	0.678

Note: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Time-varying treatment, so also implements Callaway & Sant'Anna, and generalized synthetic control (Figure 5)

FIGURE 5. Effect of Republican Control on Democratic Performance



Note: Panel (a) shows results using the Callaway and Sant'Anna estimator alternative ATT aggregation methods. Panel (b) shows the results of a generalized synthetic control analysis.

Exercise

- Please work through the code in the R file and data from the course website.
- There is nothing you need to "complete" in the exercise today, because it's rather involved.
- Instead, read the commented code and look through the code, data, and models to see how it all works