The saturation dangers in multi-decade democracy studies

Ransi Clark
Caltech

Jonathan N. Katz
Caltech

October 20, 2025

Abstract

Political scientists empirically studying the short and long term impacts of democratization are constrained to observational cross-country data to make their inferences. To do so, they rely upon outcomes that can be measured across countries and time. However, overtime many of these variables reach their natural limits and outcomes of democracies and non-democracies converge. Sometimes, such as with school enrollment rates, the bounds are apparent. Yet at other times, such as with infant mortality rates, the outcomes never reach zero but can remain close to zero. Near these saturation points, the units behave as if they are resistant to the event's impact, inducing a heterogeneity in treatment effect that depends on the baseline. Indiscriminately aggregating these dynamic treatment effects into one overall effect will bias discovered treatment effects, and increase its variance. Under saturation, designs such as the differences-in-differences violate the essential parallel trends assumption causing long-term dynamic estimates to attenuate and even change sign. Our recommendation is to perform either baseline-weighted or baseline-matched differences-in-differences or the original synthetic control. A modular estimator we suggest helps diagnose saturation by reporting treatment effects conditional on their baseline. We use data from several studies to show how saturation can lead researchers to incorrectly conclude that democratization had an adverse impact on primary education enrollment and child mortality.

1 Introduction

The effects of democratic transition have always been of interest to political economists. Theoretical models suggest that democracies tend to invest more in public services than

autocracies as a way of "paying" off median voters (Acemoglu and Robinson 2006). But testing democracy theories empirically is challenging due to the small sample size [1]. Another challenge is measuring outcomes so that are comparable across space and time. The common approach is to analyze rates or indices that account for differences in population or territory. However, rates and indices are prone to *saturation*. For example, education enrolment can reach a 100 percent and childhood mortality rates can reach as low as 0. Childhood mortality is never zero but can be close to zero.

To understand the particular challenge of analyzing saturated data, consider a study done on a sample of teenagers where the treatment is a nutritional supplement and the measured outcome is height. Teenagers in the study are observed for a 10 year period after the treatment. Suppose that poor randomization results in the treated group being made of more late teens than the control group. Since late teens are generally past their growth spurts they are unlikely to change in height under any treatment status. In contrast, the control group made mostly of early teens will see large increases in height under any treatment status. A differences-in-mean comparison between these two groups will over-estimate the treatment effect, because late teens are taller than early teens. A differences-in-differences comparison will however underestimate the treatment effect on height because the early teens in the control group will grow in height faster than the late teens in the treatment group despite not receiving treatment. Unless the treatment increases treatment group heights by more than the average baseline difference in the two groups, a naive differences-in-differences will recover a negative estimate by the end of the 10-year period.

Many cross-country comparisons of data have these exact same forces impinging upon them. Many early democratizers such as the United States and Canada had high rates of school enrollment at time of democratization. So a differences-in-mean analysis is likely to overestimate the effect of democratization. But a differences-in-differences analysis will underestimate the effect of democratization because net school enrollments cannot grow beyond 100 percent. The effect of saturation under a differences-in-differences design can range from an attenuated treatment effect to recovering a wrong sign. For example we find that the naive differences-in-differences estimate calculated for Canada's primary enrollment in 2010 is -20. This is not because Canada's primary enrollment dropped 20 points, in fact, it was at a 100 percent. Rather, the average non-democracy grew its enrollment rate by 20 points because they started with a lower baseline.

Similar sign flips occur with childhood mortality. A case in point is the differences-in-differences estimate for the effect of child mortality due to democratization in Spain, which was found to be 40, implying that child mortality increased under Spanish democracy.

[1] Samples usually consist of about a 100 countries allowing for missingness

Such large unexpected positive or negative effects should not be interpreted as an effect of democracy because it is a failure in the experimental design rather than a feature of the treatment.

In the language of the differences-in-differences, saturation is a parallel trends violation. Such violations are prohibitive to identification unless a conditioning variable can predict when the violations occur and restrict the control group to be comparable to the treatment group. Harking back to the hypothetical nutritional study, one remedy for the bad randomization is to use the subject's age as a conditioning variable. If late teens are more numerous in the treatment group, then a weighting would increase the importance of late teens in the control group. If, however, there are only late teens in the treatment group and only early teens in the control group, there is perfect separation and therefore causal effects are not identified.

Analogously, in our setting, a predictor of saturation is the baseline (pre-treatment) outcome. Countries that have high levels of primary enrollment at democratisation are likely to reach full enrollment faster than those that have low levels. Similarly for countries with low levels of childhood mortality. Baseline outcomes can be used either to restrict the control set of non-democracies in the differences-in-differences to be within a coarsened outcome range of that of the democratizing country (coarsened exact matching). Or the baseline can be used to weight outcomes in the control set so that outcomes that are closest to the democracy's is weighted higher (inverse propensity score weighting) [2].

Another guard against saturation effects in our hypothetical study is to restrict dynamic estimates to shorter lengths. Since saturation biases usually arise because the outcomes of the two comparisons groups are converging but the control group is doing so faster, biases worsen in long term dynamic estimates. Therefore, the usual practice in cross-country studies of reporting a single overall average all dynamic effects, may be too optimistic for saturated data. Restricting aggregation to shorter time spans can protect against the worse impacts of saturation effects, although we cannot necessarily be rid of them.

A classical solution to the problem of restricted outcomes is to use a Tobit correction within a regression framework. Several peculiarities of the cross-country setting make this inadequate. For one, we generally do not have a known upper or lower bound. Childhood mortality is never zero, but remain close to zero in advanced countries. A bound of zero in a Tobit correction is as good as no correction. The second problem is that regressions (even two-way fixed effects) do not recover differences-in-differences estimates even under

[2]It is quite common for there to be no control units that are of comparable baseline for a treated unit, mostly for early democratizers and late democratizers. In such instances, coarsened exact matching completely removes these comparisons. However, inverse propensity score weighting can still produce an estimate, though the standard errors will be high.

non-saturated conditions if the treatment occurs at varying times (such is the case in democratization). Applying corrections such as Tobit to a regression can compound these many problems. Thirdly, regressions do not lend themselves easily to diagnostics. We cannot easily pull out the differences-in-differences estimate that relates to country A. Lastly, regressions will aggregate these component differences-in-differences estimates regardless of whether their comparison set of non-democracies have overlap or not.

A class of estimators that have recently being popularized as heterogeneity robust panel estimators seek to recover differences-in-differences in panel settings even when treatment occurs at varying times.[3] We follow the general multi-step architecture of many such estimators in this class, but adapt these to produce differences-in-differences estimates for each democracy instead of for the entire group of democracies that transition together.

In the next section, we demonstrate how ignoring saturation can result in contradictory estimates for seemingly correlated outcomes such as primary and secondary education enrollment. Estimates produced by both the two-way fixed effects regressions and the heterogeneity robust differences-in-differences estimator of Callaway and Sant'Anna (2021) produce effects that point in different directions for primary and secondary education, when advancement in one type of education should generally predict the advancement in the other type of education. Having demonstrated these puzzling results, we turn to a more formal description of the type of parallel trends violation that is perpetrated by saturation problems in Section 3. In the same section, we show why conditioning on baseline is a solution, and how aggregation to shorter spans can guard against the worst effects if conditioning fails due to a lack of baseline-comparable non-democracies.

Section 4 applies our corrections to real data. We find that once saturation problems are accounted for, democratization had a positively signed effect on education enrolment and a negatively signed effect on child mortality. These findings also explain the null effects for primary education enrollment under democratization discovered by Paglayan (2021), and the null effects for childhood mortality discovered by Ross (2006) and Ramos, Flores, and Ross (2020).

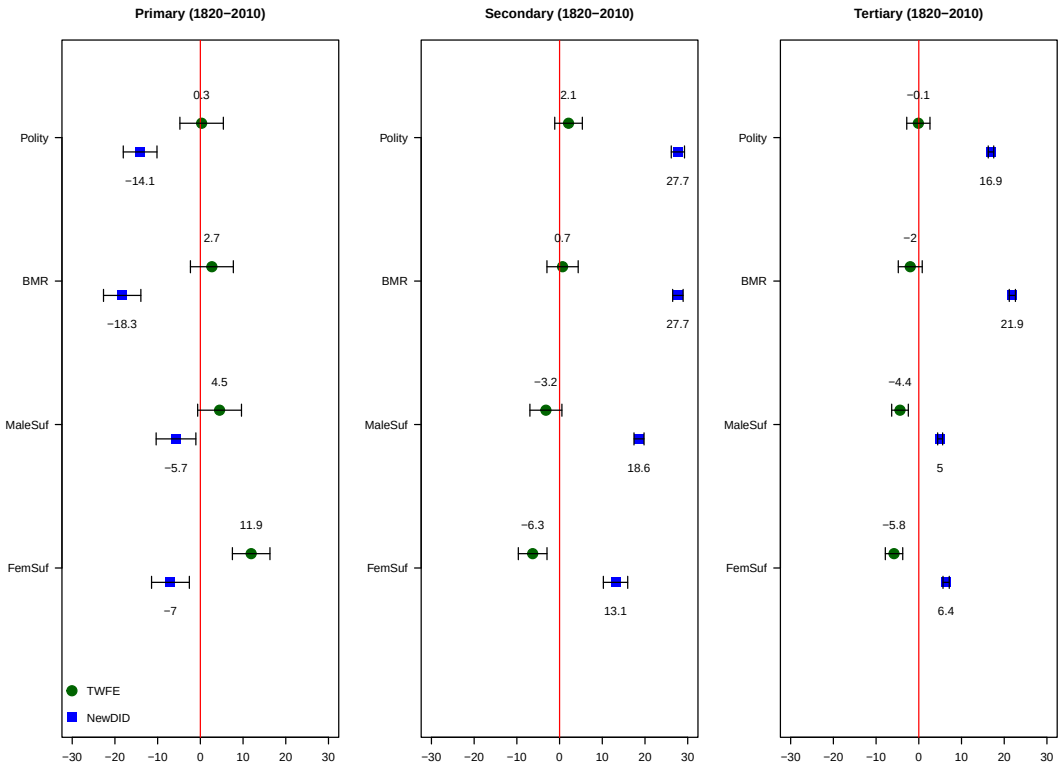
2 An education enrollment puzzle

The Figure 2.1 plots the differences-in-differences estimates for primary, secondary, and tertiary school enrollment rates (columns) implied by different democratization indicators (rows). The enrollment rates come from the data in J. Lee and H. Lee (2016). The sample

[3]The specific heterogeneity they are robust to is that of differences in treatment effects based on time of treatment.

of countries do not change across the types of enrollment. The top estimate in green circle are the estimates obtained from a two-way fixed effects regression (TWFE), and the bottom estimate from the heterogeneity-robust panel estimator of Callaway and Sant'Anna (2021) (NewDID).

Figure 2.1: Estimates of the democratization effect on net school enrollment rates



Several peculiarities are of note here. First, the two-way fixed effects regressions recover positively signed estimates for primary enrollment under all democratization indicators, but negative effects for tertiary education under all democratization indicators, and negative effects for secondary education for some democratization indicators. Seemingly, male and female suffrage expansion seems to have promoted primary education expansion, but led to either a stagnation (compared to the other non-democracies) or explicit downsizing of secondary and tertiary.

The analogous estimates produced by the differences-in-differences algorithm in Callaway and Sant'Anna (2021) (under label NewDID) are reverse of the estimates from the two-way fixed effects regression. The estimates for primary education turn out negative with some estimates as large as -18.3. But the estimates for secondary education turn out positive and as high as 27.7. Tertiary education is similar to secondary education.

These differences do not arise from changes in sample size across the type of education enrollments. So any differences must be methodological. There are two puzzles here. First is why does the heterogeneity robust differences-in-differences estimates completely reverse the sign of the two-way fixed effects estimates. The second puzzle is why do the signs between primary enrollment and secondary enrollment, and primary enrollment and tertiary enrollment differ.

To answer this, we start with setting down the imagined data-generating process. Denote the observed outcome of country i at time t be $Y_{i,t}$, where Y is either the primary, secondary, or tertiary enrollment rate. For our convenience, we normalize the first period of the sample to be $t = 1$. Each i has an associated treatment time denoted by g, so that $Y_{i,t}(g)$ is the outcome at time t of the unit i treated at time g. Applied to our setting, g is the year at which the country i first observed as a democracy. If the unit is never treated in the sample period, normalize $g = 0$.

At each time t, all countries have a Δ_t by which outcomes grow (or decline), regardless of their type of government. This is the secular trend.[4] Our target treatment effect is denoted, $\beta(g, l)$. This treatment effect can depend on g and the time since treatment l, $\beta(g, l)$.

Putting all these notational elements together, we can now explicitly state the data generating process as,

$$Y_{i,t}(g) = Y_{i,1} + \sum_{s=2}^{t} \Delta_s + \beta(g, l) * \mathbb{I}_{\{g>0 \ \& \ t \geq g\}} + \epsilon_{i,t} \tag{1}$$

Supposing our interest is in a β that is the summary of all such $\beta(g, l)$'s. Notationally,

$$\beta = \sum_{g} \sum_{l} w(g, l)\beta(g, l) \tag{2}$$

where $w(g, l)$ is some weighting scheme that add to 1 over all g's and l's, so that β is some weighted mean of component estimates.

A two-way fixed effects estimator recovers this β by running the following regression on the data,

$$Y_{i,t} = \alpha_i + \alpha_t + \beta * \mathbb{I}_{\{g>0 \ \& \ t \geq g\}} + \epsilon_{i,t} \tag{3}$$

This specification assumes that the treatment effect is constant across treatment time, so that $\beta(g, l) = \beta$ for all g and for all l. Although it is not apparent, the two-way fixed effects

[4]Sometimes also known as the common trend. The assumption that these Δ's are same in the treatment and control group is the parallel trends assumption.

calculates sub-estimates $\beta(g,l)$ and weights then using its own internal weighting scheme $w(g,l)$.

One criticism of two-way fixed effects is that these $\beta(g,l)$ are sometimes derived by comparing late democracies (as the treated) to early democracies (as the control), when early democracies should never enter a control set after their own g. This causes sign flips whenever early democracies have a different treatment effect to that of the later democracies, such that if $\beta(g',l) > \beta(g'',l)$ and $g' > g''$ for any l. These are referred to as forbidden comparisons. A preponderance of forbidden comparisons is a partial explanation of the negative signs observed for secondary education and tertiary education under two-way fixed effects.

Still another criticism is that the two-way fixed effects imposes arbitrary weights $w(g,l)$, when aggregating these estimates. Borusyak, Jaravel, and Spiess (2021) finds that for longer lags l the $w(g,l)$ can attenuate and sometimes turn negative. Although the negative weights are usually unlikely, since we are analyzing data over several decades (sometimes, centuries) the probability of negative weights increase in time.

The newer differences-in-differences estimators such as that of Callaway and Sant'Anna (2021) produces estimates using a more flexible procedure, where each $\beta(g,l)$ is calculated separately, instead of all at once as does the two-way fixed effect regression. In this way the algorithm can specifically impose that early democracies never enter a control set. Notationally, the estimate is calculated as,

$$\beta(g,l) = |\mathcal{I}_g|^{-1} \sum_{i \in \mathcal{I}_g} Y_{i,g+l}(g) - Y_{i,g-1}(g) - |\mathcal{C}_{g+l}|^{-1} \sum_{j \in \mathcal{C}_{g+l}} [Y_{j,g+l}(0) - Y_{j,g-1}(0)] \tag{4}$$

where $\mathcal{I}_g$ is the set of all countries democratizing at time g, and $\mathcal{C}_{g+l}$ are countries that are not yet democracies at time $g+l$.

These are then aggregated similarly to Eq. 2 above where $w(g,l)$ is always positive. Weights will differ based on number of democracies in g, but given g, do not differ in l. While these estimators guard against forbidden comparisons that plague two-way fixed effects estimates, they still require that parallel trends assumptions be satisfied. That is in the data generating process, Δ_s should be the same regardless of democracy status.

Saturation of primary enrollment would be a violation of this assumption; early democracies are generally close to full enrollment and grow more slower than their comparable non-democracies. The negative effects these estimates recover is an artefact of this violation. The effect of saturation is apparent when comparing the heterogeneity-robust differences-in-differences estimates across the types of enrollment. Secondary and tertiary enrollment rates are rarely at full enrollment even in the most developed countries. Therefore, saturation is

a minor concern as a parallel trend violation.

Saturation also explains why the two-way fixed effects estimates for primary enrollment were positive. Saturation problems are worst among the early democracies such as the Canada. When used as controls in a forbidden comparison, these early democracies produce a positive differences-in-differences estimate.

Parallel trend violations are usually prohibitive to causal identification under a differences-in-differences design. One solution to such violation is to find a conditioning variable that predict the violations. Such as age in our hypothetical nutritional experiment. The next section expands on the bias of saturation, and details an estimation strategy to alleviate the bias. Since this proposed estimator is modular — operates in steps — it lends well to saturation diagnostics.

3 Saturation in cross-country data

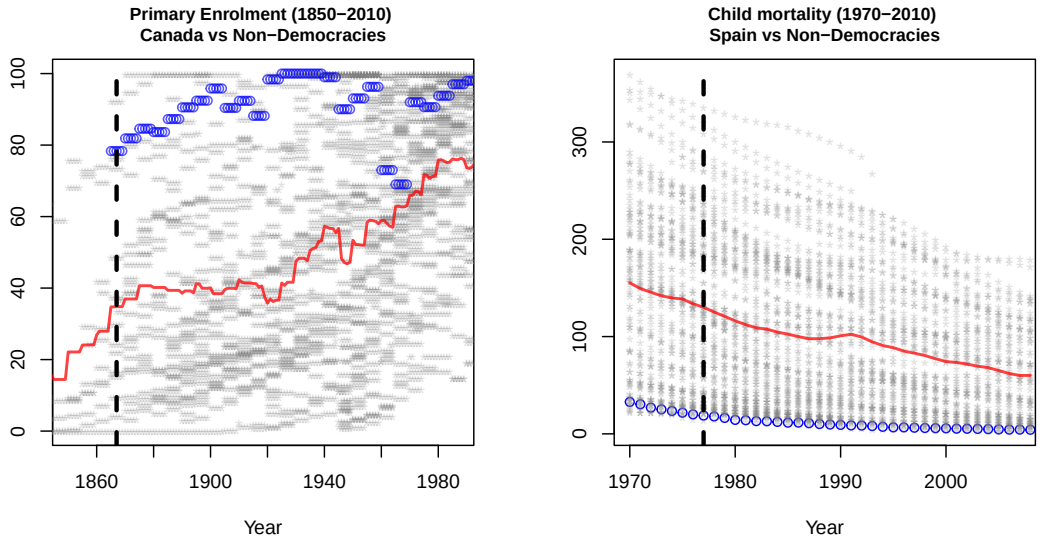
The Figure 3.2 demonstrates excerpts of data ripe for saturation biases. The left panel shows Canada's primary enrollment rate from the years 1866-2010 (in blue circles), along with the average primary enrollment rate for all other non-democracies at each year (in solid red line). Canada democratizes[5] in 1867 with a primary enrollment rate of about 80%. Canada's enrollment rate reaches full enrollment in 1920 and remains so until about 1940. It is again at full enrollment by 2010. The very low enrollment figures of 1960s (70%) is due to a missing value imputation done in UNESCO Statistical Yearbooks that J. Lee and H. Lee (2016) sources the enrollment raw figures from.

Because Canada's enrollment rate at democratization is higher than its contemporary non-democracies, a differences-in-mean calculation (differences of the blue circles and red dots) will attribute this baseline difference to the treatment effect. Differences-in-differences designs avoid this problem, since it recenters the trajectories at their respective baselines (at 80 for Canada and at 35 for the non-democracies). In turn, however, the design is exposed to the risks of saturation bias. Net primary enrollment cannot grow further beyond a 100 percent. Countries such as Canada that democratize with higher baselines are more restricted. Countries with lower baselines are less constrained. Therefore, the differences-in-differences estimate for Canada's democracy would be negative from 1940 onwards. In 2010, where it again reaches full enrollment the differences-in-differences estimate is -20. With an upward secular trend in non-democracies, once at full enrollment the differences-in-differences estimate for Canada's primary enrollment can only be negative.

In the right panel of Figure 3.2 is Spain's childhood mortality rate (in blue circles) and

[5]That is, according to binary indicators in Boix, Miller, and Rosato (2013).

Figure 3.2: Net primary enrolment rate (Canada) and child mortality rate (Spain)



Note: Left panel is net primary enrolment data with Canada in blue circles. Right panel is child mortality with Spain in the blue circles. The solid red lines in each panel demonstrate the yearly average of the non-democracies each year. The grey dots are trajectories that the yearly averages are calculated from. The dashed vertical line denotes the year of democratization of Canada and Spain respectively.

that of contemporary non-democracies (in solid red line). In 1977, Spain reemerges as a democracy after its almost 40-year Franco dictatorship. Spain's mortality rate is much lower than other non-democracies in 1977 at 10 deaths per 1000, and falls to 5 deaths per 1000 in 2010. But, all other non-democracies though with higher mortality rates than Spain see a steeper drop in this time period. Here, the differences-in-differences estimates for Spain will be positive from as early as 1970. Analogous to before, with a downward secular trend in non-democracies, the differences-in-differences estimate for Spain's child mortality can only be positive.

To recap, in the case of primary enrollment the differences-in-differences estimates were mostly negative. And in the case of child mortality differences-in-differences estimates were mostly positive. Both imply that democratization had adverse effects on the relevant outcomes. But the implications are not true. Rather, it is an artefact of the outcome analyzed and the assumed experimental design, rather than a feature of democratization.

To put this more formally, consider an amendment to the previous data generating process in Eq. 1 where the common trends Δ_s depend on the starting value $Y_{i,1}$, as do the treatment effects $\beta_{g,l}$.

$$Y_{i,t}(g) = Y_{i,1} + \sum_{s=2}^{t} \Delta_s(Y_{i,1}) + \beta_{g,l}(Y_{i,1}) * \mathbb{I}_{\{g>0 \ \& \ t \geq g\}} + \epsilon_{i,t} \tag{5}$$

Supposing we are considering Canada in the left panel and renormalize $Y_{i,1866}$ to be $Y_{i,1}$. Our new data generating process implies that the counterfactual common trends should depend on the starting value. This is denoted by the conditioning on $Y_{i,1}$, $\Delta_s(Y_{i,1})$ and $\beta_{g,l}(Y_{i,1})$. This implies that estimating Δ_s out of the entire control group, may induce bias. Instead, the control group must either be restricted to those countries with a similar baseline, or weighted in such a way that countries closer to the treated unit's baseline receive more importance.

Because of the small sample size and attrition[6], both restriction and weighting face practical challenges. We take these up in the next section after describing the estimation strategy.

3.1 Our proposed estimator

Similar to many heterogeneity-robust differences-in-differences estimators, we propose a modular estimation strategy. A first-step produces component differences-in-differences indexed by the country, and the year. A second step aggregates. We maintain an architecture similar to that of Callaway and Sant'Anna (2021), but differ in the granularity of the first-step. That estimator calculates its most granular first-step differences-in-differences at the level of the set of countries that democratize together, but we do so at the level of the country, so that each democracy has its own series of differences-in-differences estimates. This adaptation is necessary because countries that democratize together can have different baselines. Without separate estimates for these countries we cannot perform diagnostics.

Moreover, since countries can backslide from democracy at any time, treating all countries in the same group, biases dynamic estimates after one of the countries in the group exits (either the sample or becomes a non-democracy). Calculating a single estimate at the group level precludes any aggregation to attributes that differ within the treated time group.

In what follows,the index i will be a country. Years will be indexed by t. The year of democratization, is encoded in g. The index l denotes lag length, i.e. years since democratization.

As before, let $Y_{i,t}$ denote the outcome at year t for country i. Each country i is first observed as a democracy at year g. Under a potential outcome framework, $Y_{i,t}(g)$ should be meaningful for any i, g combination, though we mostly observe the outcomes for one g for each i. For ease of notation, we will assume that countries that democratize in a year g will remain a democracy until the end of the study period, T. This is without loss of generality, since new transitions can be recoded as new countries.

[6]Most attrition occurs in the control sample when non-democracies transition to democracies. Attrition changes the composition of the sets between dynamic lags.

The control set C_t are non-democracies at time t.[7] Countries that never transition to democracies have g normalized to 0. While a number of potential treatment effects can be calculated as $Y_{i,t}(g) - Y_{i,t}(g')$ for any pair of g, g', we restrict g' to be the set $\{g' : 0 \vee g' > g\}$. That is, the control set is either made of countries that never democratize by the end of the sample or countries that are yet to democratize. Therefore, g in $Y_{i,t}(g)$ denotes the true transition time of the associated i.

The estimator has a multi-step structure. The first step produces country-level 2×2 differences-in-differences estimates for each year of the sample period. These estimates are denoted $DID(i, t, g)$. The next step then aggregates these estimates. Aggregations can be within the lag l, group g, or even the baseline at time of democratization. Fixing a country i that democratizes at g, the first-step estimates of the treatment effect on the treated i calculated for all years t are of the form:

$$DID(i, t, g) = [Y_{i,t} - Y_{i,g-1}] - \sum_{j \in C_t} w_j(_{Y_{i,g-1}})[Y_{j,t} - Y_{j,g-1}] \tag{6}$$

Here w_j is a weighting that adds to 1. This weighting is positive for each j but can be very small for some of them. Weighting usually requires another estimation routine before calculating Eq. 6. The usual procedure is to estimate a propensity score using a logistic regression on the treatment indicator D_i and the baseline $Y_{i,g-1}$. These inverse propensity weights are obtained by taking a ratio of the propensity score over its differences from 1.

Weighting weakens the unconditional parallel trends assumption to the weaker conditional parallel trends assumption (that the trajectory of the democratizing country need not be parallel to every non-democracy but only those that are close to some baseline covariates). More specifically, countries in C_t whose $Y_{j,g-1}$ are closest to the $Y_{i,g-1}$ will be weighted higher than those that are further away. The synthetic control is another weighting algorithm. [8] Baseline matching is also possible, but requires some coarsening for practical use.[9].

[7] Two-way fixed effects regressions do not impose this restriction and admits earlier democratized countries even when they are democracies at time t, a major cause of its tendency to produce flipped signs.

[8] The synthetic control weighting algorithm is most useful if there is a longer pre-democratic period observable and there are no missing values in that period. Because countries enter and exit our sample at all times, we believe the synthetic control's sample balance is too demanding for a cross-country sample, unless imputed. We also caution against the use of synthetic controls that involve intercept shifts or negative weights for w_j. Shifting control country trajectory's in this way exacerbates the problem and will result in large sign flips.

[9] Another way to achieve baseline conditioning is to use coarsened matching, for example, by restricting the set C_t to units that are within some pre-specified region of the baseline covariates (Iacus, King, and Porro 2012). Under such a method w_j would have the form,

$$w_j = \frac{\mathbb{I}_{j \in R_i}}{|C_t \cap R_i|}$$

The first-step usually produces a large number if estimates. For a sample with 20 democracies observed over a century, 2000 such estimates will be produced. In the second-step these estimates are aggregated to summary statistics. Aggregation can be within the country i level, democratization year g level, or other custom aggregation such as the baseline of the democratizing country. The type of aggregation most relevant to us is that of time since democratization, or lag length l. Supposing that I_l is the set of all countries that democratize l periods before the end of the sample, the dynamic estimate of lag length l is,

$$ATT_l = |I_l|^{-1} \sum_{\{(i,g,t)|i \in \mathcal{I}_l \text{ and } t-g \equiv l\}} DID(i,g,t) \tag{7}$$

The lags l considered can also be negative. These are useful for testing for pre-treatment parallel trend violations. However, since saturation biases set in on long-term post-democratization estimates even when violations are not detected pre-treatment, they are a concern for post-treatment estimates.

We may also aggregate each democracy's post-democratization estimates into a specified lag-length. Supposing this lag length is L, for a democracy i,

$$ATT^L(i) = \frac{1}{L-g+1} \sum_{\{(i,g,t)|t-g \leq L\}} DID(i,g,t) \tag{8}$$

This can then be further aggregated to an overall average treatment effect on the treated. Supposing that $\mathcal{I}_L$ is the set of all democracies that are observed at least one period after democratization, this overall estimate is,

$$ATT^L = |\mathcal{I}_L|^{-1} \sum_{i \in \mathcal{I}_L} ATT^L(i) \tag{9}$$

When L is left unrestricted to be the longest period in the sample, this gives the overall treatment effect, ATT.

Other aggregates such as aggregating to the democratizing year, or to the baseline can be produced analogously.

Two issues warrant further discussion. One of this is what one might do if there is perfect baseline separation between the democratizing and control countries. The second is how one might calculate standard errors for the $DiD(i,g,t)$ with just a single treated unit.

where R_i is the region in the baseline covariate space where the unit i is.

3.1.1 Excising overlap violations

Weights w_j in definition Eq. 6 require that the baseline $Y_{i,g-1}(g)$ is "overlapped" by the control units' baselines (for weighting) or be within the coarsening (for matching). The practical requirement for overlap in inverse propensity weighting is that at least one of the control units' baseline outcomes should be above the treated unit's baseline, and at least one of the control units' baseline outcome are below the treated units' baseline.

When this is violated, baseline matching outputs a missing value for that comparison due to lack of controls, while the inverse propensity weights will be numerically unstable, and particularly problematic for standard error calculation.[10] Synthetic controls will also experience similar issues because an overlap violation usually implies that the treated unit's pre-treatment trajectory is outside the convex hull of control set's pre-treatment trajectory.[11]

One approach to overlap violation is to estimate effects only for the *feasible* comparison sets. This makes our interpretation be for the limited set rather than for the entire sample. However, useful information can be gleaned by those $DiD(i, g, t)$'s for whom overlap violations occur when conditioned on baseline.

When generating our results, we produce both overall estimates excising $DID(i, g, t)$s with overlap violations and overall estimates without excising those.

3.1.2 Inference

Inference for $DiD(i, g, t)$ and their second-step aggregates are done using techniques in conformal inference (Lei and Candès 2021). Because the first-step estimates only have a single treated unit, we cannot learn the treated group's variance. However, by assuming that the $\epsilon_{i,t}$'s in equation Eq. 1 and Eq. 5 share the same variance, we can produce a prediction interval for the untreated counterfactual of $Y_{i,t}(g)$ using the control set. This set is then inverted against the observed $Y_{i,t}(0)$. Algorithm 1 in the Appendix details how a jackknife+ algorithm can be used to produce these confidence intervals (Barber et al. 2021).

Briefly the algorithm operates by reestimating weights $w_{j'}$ leaving out each j in the set C_t. For each j left out, a new counterfactual rate change for i is estimated, $\Delta Y_{i,t}^{(-j)}$,

$$\Delta Y_{i,t}^{(-j)} = \sum_{j' \in C_t \setminus j} w_{j'}(Y_{j',t}(0) - Y_{j',g-1}(0))$$

[10] If all control baselines $Y_{j,g-1}(0)$ are above (below) $Y_{i,g-1}(g)$, then the control country with baseline just below (above) $Y_{i,g-1}(g)$ will be given a weight of 1. This inflates standard errors, because we now have effectively 2 units in our sample.

[11] The `Synth` package defaults to a uniform weight when the treated unit is outside the control sets outcomes' convex hulls.

and a new residual $r_i^{(-j)}$, where

$$r_i^{(-j)} = \Delta Y_{j,t} - \Delta Y_{i,t}^{(-j)}$$

Then the counterfactual's interval estimate is formed by taking α quantile of $\Delta Y_{i,t}^{(-j)} - |r_i^{(-j)}|$ and the $1-\alpha$ quantile of $\Delta Y_{i,t}^{(-j)} + |r_i^{(-j)}|$. The interval estimate for $DiD(i,g,t)$ is obtained by inverting these interval against the treated unit's realized outcome $Y_{i,t}(g) - Y_{i,g-1}(g)$. The interval estimates produced in this way have finite sample coverage guarantees instead of the usual asymptotic coverage guarantee.

To obtain the standard errors for aggregation, statistical independence among countries must be assumed. These intervals only have asymptotic coverage guarantees. If finite sample coverage guarantees are needed, Minkowski means of the interval estimates for $DiD(i,g,t)$ can be used. However, these are always too conservative. Algorithms 2 and 3 describe how intervals can be aggregated for the quantities describes in Eq. 8 and Eq. 9.

The assumption of exchangeability where the standard error of the residuals $\epsilon_{i,t}$, σ_ϵ is invariant to the unit i's treatment status ($\sigma_\epsilon(g > 0) = \sigma_\epsilon(g = 0)$) is stronger than that assumed by most multi-step differences-in-differences estimators. For example, Callaway and Sant'Anna (2021) allows for these residual variances to differ unconditionally ($\sigma_\epsilon(g > 0) \neq \sigma_\epsilon(g = 0)$). Our inferential approach under-covers if $\sigma_\epsilon(g > 0) > \sigma_\epsilon(g = 0)$, and recovers a conservative interval (over-covers) if $\sigma_\epsilon(g > 0) < \sigma_\epsilon(g = 0)$.

However, the inferential method we propose has better coverage when group (set of countries democratizing in the same year) sizes $|I_g|$ are (heuristically) smaller than 3 countries, even if this variance equality assumption is violated. This is more often the case, because in many sets I_g is only a single democracy.

3.2 Simulation

The Appendix includes results from a simulation which compares several regression estimators, the Callaway and Sant'Anna (2021) estimator and our proposed estimator under a data generating process where outcomes are unrestricted and a data generating process where outcomes are restricted. This latter process mimics a saturation situation such as in left panel of Figure 3.2. For visual depiction, we identify the Callaway and Sant'Anna (2021) estimator as `did`, the name of the package in `R` that implements that estimator. And identify our proposed estimator as `didunit`, also a package in `R`.

The estimators compared are country fixed effects regression (unitFE), year fixed effects regression (yearFE), two-way fixed effects regression (TWFE), two-way fixed effects estimator with unit-level time trends (u-TWFE), two-way fixed effects with Tobit correction

(TWFE Tobit), the unconditional estimates from (`did`), the baseline-weighted estimates of above (`did` Weighted), our unit-level adaptation without baseline-weighting (`didunit`) and with baseline weighting (`didunit` Weighted). The coarsened baseline outcome restricted version is named (`didunit` Restricted).

Under restricted outcomes, TWFE, Unweighted `did`, Weighted `did`, and Unweighted `did` recover the opposite sign. unitFE, yearFE, u-TWFE all recover the correct sign, but overestimates the treatment effect. TobitTWFE recovers the true treatment effect as long as the treatment effect is the same across all units (a strong assumption). Our proposed estimators recover the true treatment effect under all conditions.

4 Data section

This section demonstrates how saturation manifest in real data. Arranging first-step $DiD(i, g, t)$s on their pre-treatment baselines reveals tell-tale signs of saturation biases. Following these diagnostic demonstrations, we estimate aggregate effects.

As outcomes, we consider primary enrollment and childhood mortality. Experience tells us that neither outcome worsened considerable under democracies. Therefore, we should not expect large negative effects for primary enrollment or large positive effects for childhood mortality.

For effect of democratization on primary enrollment, we use data from Acemoglu, Naidu, et al. (2019) and Paglayan (2021). For the effect on child mortality, we use data from Ross (2006) and Ramos, Flores, and Ross (2020). Some datasets were processed further. Particularly, missing outcomes were imputed with past values not more than 6 years older.[12] Because our algorithm requires a binary indicator of democracy, even when the original data exercise used a non-binary indicator of democracy, our analysis used a binary indicator.[13]

We also recoded the original datasets to accommodate for backsliding. Countries that backslide for more than 5 years are re-introduced to the control set of non-democracies with a different code. The first 5 years of these reintroduced democracies is truncated, because these are usually times of war, or crisis. We chose to retain backsliding countries in the sample without eliminating them entirely because even typical autocracies can have short democracy stints.[14] Countries that democratize again are introduced back in with the same identification i.

[12]Imputation helps retain sample sizes when the immediate pre-treatment outcome is missing.

[13]Ross (2006) uses a continuous indicators of democracy PolityIV. We are restricted to binary indicators in our new estimator, so use the ACLP indicator of Cheibub, Gandhi, and Vreeland (2010) which we found in the replication folder of Ross (2006).

[14]Cuba, for example, had a short period of democratic governance.

4.1 Diagnosing saturation

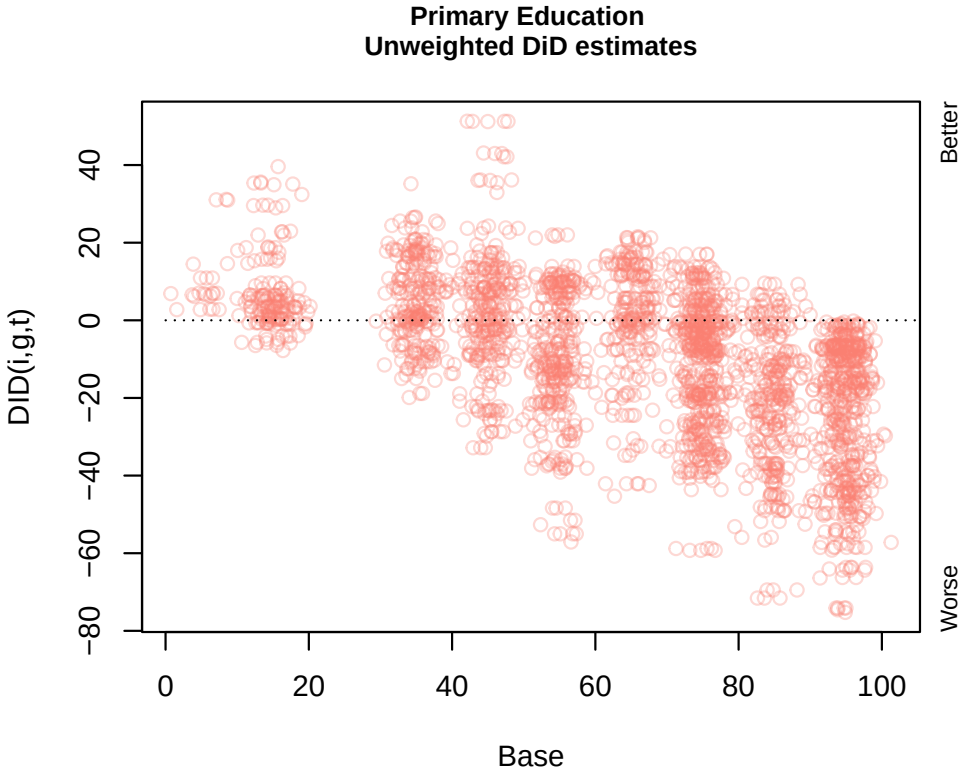
The worst effects of saturation manifest in sign flipping, since it can lead to incorrect conclusions. The component estimates $DiD(i, g, t)$s helps detect democracies that are most likely sign flipping. One diagnostic exercise is to arrange $DiD(i, g, t)$s based on their corresponding baseline outcomes $Y_{i,g-1}$. Closer the baseline is to a saturation point, worse are the saturation effects.

Figure 4.3 demonstrates possible saturation biases for primary enrollment data from 1820-2010. The plotted estimates are from $DiD(i, g, t)$s calculated as in Eq. 6 but where no covariates were provided for weighting. As a shorthand we call these estimates *unweighted*. The higher the baseline the more negative the estimate is. For those democracies whose baselines are close to 100, the estimates are overwhelmingly negative. Some estimated negative effects are as large as -80. On the other hand, for baselines around 15, the estimates are overwhelmingly positive.

Figure 4.4 reports those same estimates but when Eq. 6 is calculated by weighting on the vector of baseline outcomes. Some component $DID(i, g, t)$s did not have sufficient overlap and can be numerically unstable if the underlying propensity score estimation algorithm does not converge. Nevertheless, at most times some estimate is output. The grey crosses plot these. In contrast to the previous Figure 4.3, the worst negative estimates are ameliorated by the weighting. When weighting is not possible due to poor overlap, large negative estimates are still observed. We recommend dropping these estimates when aggregating so that they do contaminate overall treatment effects. Alternatively, as we do in the next subsection, report aggregates with and without excising the overlap-poor estimates.

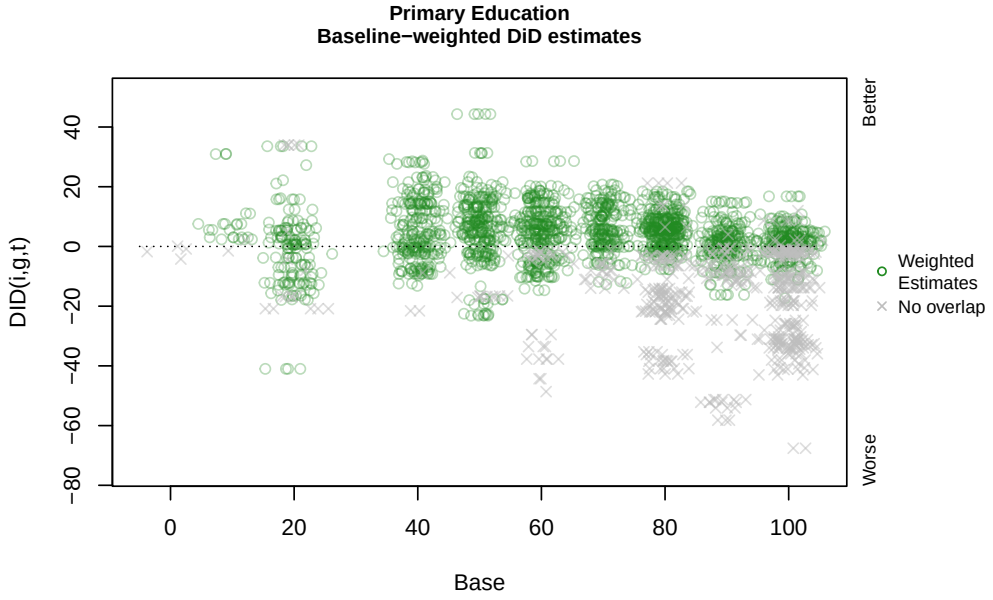
Estimates on childhood mortality data exhibit a similar pattern of violation, as in Figure 4.5. Since the lower child mortality is a better outcome, a negative estimate signifies improvement. Countries that democratize with a baseline close to zero produce a positive-signed effect on child mortality. Baseline-weighting deals with these positively biased estimates to a large extent. Compared to the previous example on primary enrollment much fewer $DiD(i, g, t)$s are dropped due to overlap violations.

Figure 4.3: Component estimates for effect of democratization on primary enrollment



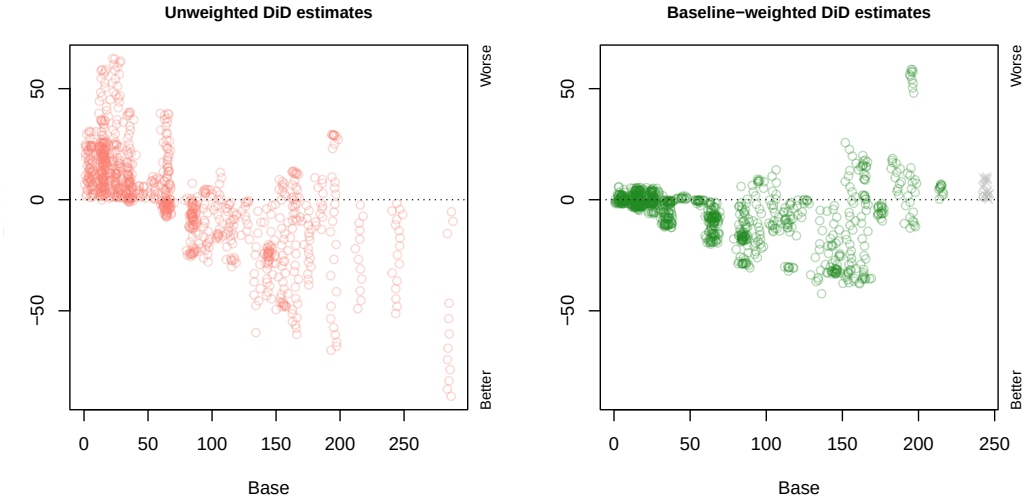
Note: These are the unweighted DiD first-step estimates $DID(i, g, t)$ as in Eq. 6 arranged based on the democratizing country's baseline $Y_{i,g-1}$. They are termed *unweighted* because all control countries are uniformly weighted. The enrollment data the estimates are based on is from J. Lee and H. Lee (2016) and the democracy indicator is from Boix, Miller, and Rosato (2013). Negative estimates imply an adverse effect for enrollment.

Figure 4.4: Component estimates for effect of democratization on primary enrollment



Note: These are the baseline weighted DiD first-step estimates $DID(i, g, t)$ as in Eq. 6 arranged based on the democratizing country's baseline $Y_{i,g-1}$. The enrollment data the estimates are based on is from J. Lee and H. Lee (2016) and the democracy indicator is from Boix, Miller, and Rosato (2013). Negative estimates imply an adverse effect for enrollment. Grey crosses show estimates that had no overlap, and should be dropped for aggregation.

Figure 4.5: Component estimates for effect of democratization on child mortality



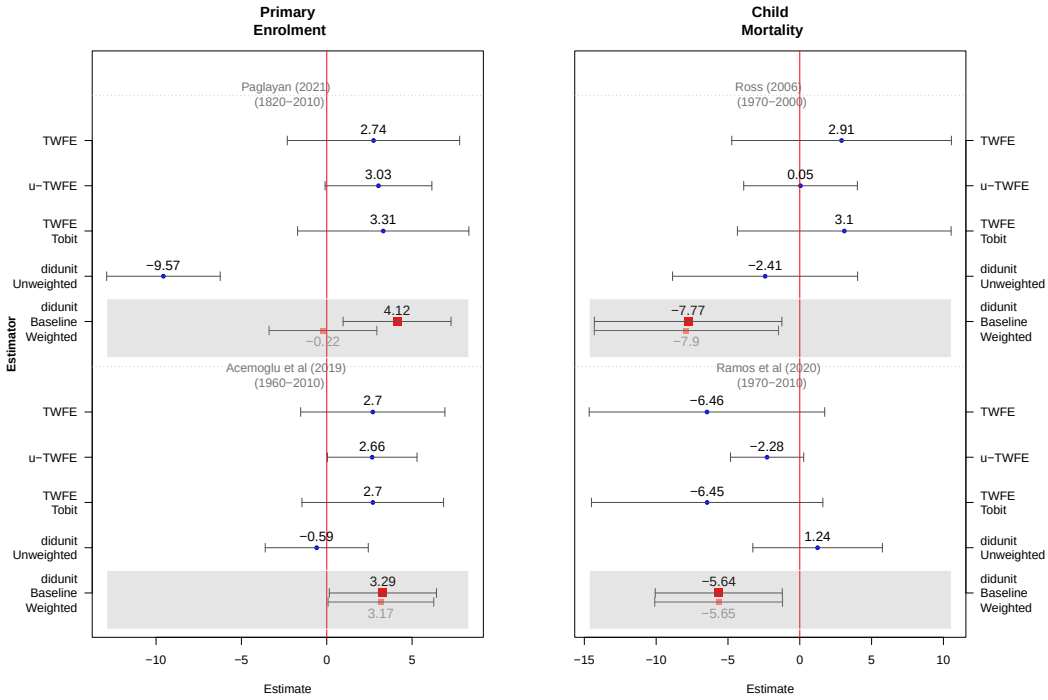
Note: The left panel are unweighted DiD first-step estimates as in (Eq. 6) arranged based on the democratizing country's baseline. The right panel are baseline-weighted counterparts of those on the left. The greyed out dots are those that violate overlap.

4.2 Overall estimates

Here we report the overall post-treatment effects that are recovered by our baseline-weighting procedure and compare them to estimate obtained by two-way fixed effects regressions.

While all original papers use some version of two-way fixed effects, they differ in the specifications they use. Since our intention here is not to verify the results of the original papers, we use a fixed set of regressions for comparison, even when the original authors may not have used the exact specifications.[15] These specifications are the two-way fixed effects regression (**TWFE**), the unit-level time trend two-way fixed effects regression (**u-TWFE**), Tobit-correction applied to two-way fixed effects (**Tobit-TWFE**), the unweighted aggregate of Eq. 9 (`didunit` **Unweighted**), and the baseline-weighted aggregate (`didunit` **Weighted**). Figure 4.6 demonstrates the above estimates recovered from the four datasets.

Figure 4.6: Overall estimates



Note: Demonstrates estimates from the relevant estimator labelled on the y-axis. The baseline-weighted version of our estimator is highlighted in gray (`didunit` Baseline Weighted). The bottom estimate in the gray region does not excise individual $DID(i, g, t)$s the violate overlap. All baseline-weighted estimates produced from comparison sets with sufficient overlap imply improvement of outcomes under democratization.

The baseline-weighted aggregates (highlighted in gray) ameliorate saturation biases, if

[15] Acemoglu, Naidu, et al. (2019) runs an additional two-stage least squares specification, which we do not compare here.

they matter. Two versions of these weighted estimates are plotted. The top estimate within the gray region drops overlap-poor $DID(i, g, t)$s, and the bottom estimate retains them. The two versions only differ substantively for the primary enrollment data (under Paglayan (2021)). This is because the time scale there is much longer and therefore has more opportunity for saturation biases to overwhelm the aggregates. In all four datasets, the weighted estimates imply that democratization had a positive effect on the relevant outcomes.[16]

Another tell-tale sign of saturation bias is a sign disagreement between the weighted and unweighted aggregates. This is most obvious in the primary enrollment data (in Paglayan (2021)), where the unweighted estimate is significantly negative, while the weighted estimate is significantly positive. A similar sign disagreement occurs for primary enrollment data in Acemoglu, Naidu, et al. (2019) and child mortality data in Ramos, Flores, and Ross (2020). Saturation biases are slightest in the child mortality data in Ross (2006). Yet, even in this instance they cannot be ruled out, because the weighted estimates differ.

Ironically, the two-way fixed effects and its versions are largely in sign-agreement with the weighted `didunit` estimates. The only exception to this is the child mortality data in Ross (2006) where all TWFE recover a positively signed effect. This coincidental agreement arises partly because of the additional forbidden comparisons made by two-way fixed effects. To recall, such a forbidden comparison would include early democracies in the control set for late democracies. If early democracies' primary enrollment outcomes are more likely to saturate because they started at a high baseline, when compared to late democracies as the resulting estimate is large and positive. These estimates cancels out the large negative effects produced by the remaining "legal" comparisons. Similar forces are at work with mortality data.

Despite the Tobit estimators' robustness to censoring, its use was limited here, because saturation occurred much earlier than the maximum (or minimum) possible value we imposed as the censoring point. This is confirmed by Figure 4.3 and left panel of Figure 4.5. Sign flipping occurs much before the highest (or lowest) possible value.

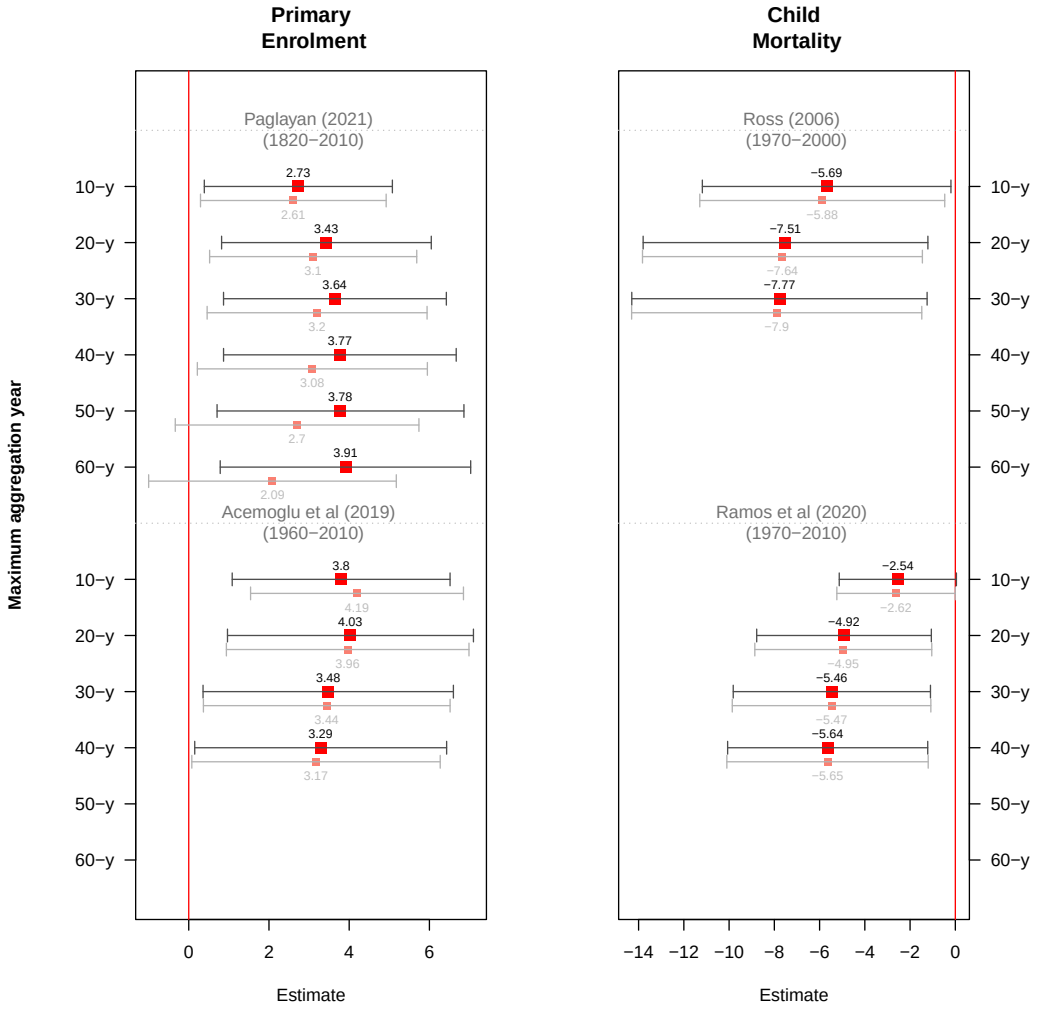
The estimates in the previous Figure 4.6 were aggregations to the highest possible lag length. Since saturation biases get worse over longer observation windows, another way to ameliorate these is to restrict the lag length in aggregation. For example, a restricted overall estimate can be produced by aggregating only first decades' estimates, i.e.

$$DiD(i, g, g), \cdots, DiD(i, g, g + 10)$$

. Figure 4.7 demonstrates aggregations for the first six decades using the baseline weighting.

[16]These estimates may change if other conditioning variables are added.

Figure 4.7: Overall estimates baseline weighted and decadal restrictions



Note: Variably aggregates baseline-weighted component estimates $DID(i, g, t)$. Top estimate in red square is the aggregate after dropping overlap-poor component estimates. The bottom estimate retains these.

For all datasets, increasing lag length does not substantively change the estimate though it seems to increase the variance. When aggregating with overlap-poor estimates, the aggregates for primary enrollment data in Paglayan (2021) show a greater tendency to sign-flip.

The results show that the methodological augmentations we propose are robust to saturation problems, although we cannot guarantee that they can be completely dealt with. The effectiveness of both weighting and restricting lag length depend on the availability of adequate controls.

5 Conclusion

Political scientists are no strangers to the problem of ceiling and floor effects caused by outcomes that can saturate. Page and Shapiro (1983) describes these issues in survey experiments as ceiling and floor effects. However, little attention is paid to the biases in time series cross-country studies. We conjecture that this is because researchers generally default to black box like estimators such as two-way fixed effects where saturation biases cannot be easily diagnosed.

While saturation biases can be slight in cross-sectional studies, as we have demonstrated, in long-term studies they can become the overwhelming effect. Under some conditions, these effects can cause sign changes, leading researchers to conclude against the truth. This is not a rare occurrence either, and exacerbates in the longer term. Some of the most widely studied outcomes such as school enrolment and childhood mortality can exhibit these biases. The problem is not limited to these variables. Growth rates, vote shares, and a variety of other outcomes can be similarly saturated.

Saturation is a type of parallel trend violations. Yet can be alleviated by conditioning on the baseline, a predictor of saturation as those with baselines close to saturation points attain those limits quicker than countries further away. Our proposed correction produces differences-in-differences estimates for each democratizing country, weighting on the baseline outcome. This allows diagnostics based on each country's baseline, and flexible aggregation to variable lag length. Aggregation can also be done excising estimates with poor baseline overlap. Applications of the method to real data recover treatment effects that imply that democracy has advanced public welfare.

References

Acemoglu, Daron, Suresh Naidu, et al. (2019). "Democracy does cause growth". In: *Journal of political economy* 127.1, pp. 47–100.

Acemoglu, Daron and James A Robinson (2006). *Economic origins of dictatorship and democracy*. Cambridge University Press.

Barber, Rina Foygel et al. (2021). "Predictive inference with the jackknife+". In: *The Annals of Statistics* 49.1, pp. 486–507.

Boix, Carles, Michael Miller, and Sebastian Rosato (2013). "A complete data set of political regimes, 1800–2007". In: *Comparative political studies* 46.12, pp. 1523–1554.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2021). "Revisiting event study designs: Robust and efficient estimation". In: *arXiv preprint arXiv:2108.12419*.

Callaway, Brantly and Pedro HC Sant'Anna (2021). "Difference-in-differences with multiple time periods". In: *Journal of Econometrics* 225.2, pp. 200–230.

Cheibub, José Antonio, Jennifer Gandhi, and James Raymond Vreeland (2010). "Democracy and dictatorship revisited". In: *Public choice* 143, pp. 67–101.

Goodman-Bacon, Andrew (2021). "Difference-in-differences with variation in treatment timing". In: *Journal of Econometrics* 225.2, pp. 254–277.

Iacus, Stefano M, Gary King, and Giuseppe Porro (2012). "Causal inference without balance checking: Coarsened exact matching". In: *Political analysis* 20.1, pp. 1–24.

Lee, Jong and Hanol Lee (2016). "Human capital in the long run". In: *Journal of Development Economics* 122, pp. 147–169.

Lei, Lihua and Emmanuel J Candès (2021). "Conformal inference of counterfactuals and individual treatment effects". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.5, pp. 911–938.

Page, Benjamin I and Robert Y Shapiro (1983). "Effects of public opinion on policy". In: *American political science review* 77.1, pp. 175–190.

Paglayan, Agustina S (2021). "The non-democratic roots of mass education: evidence from 200 years". In: *American Political Science Review* 115.1, pp. 179–198.

Ramos, Antonio P, Martin J Flores, and Michael Ross (2020). "Where has democracy helped the poor? Democratic transitions and early-life mortality at the country level". In: *Social Science & Medicine* 265, p. 113442.

Ross, Michael (2006). "Is democracy good for the poor?" In: *American journal of political science* 50.4, pp. 860–874.

6 Appendix

6.1 Algorithms for interval estimates

Algorithm 1: CV+/Jacknife+ Prediction Interval for $DiD(j,g,t)$ (two-sided, level $1-\alpha$)

---

1. **Inputs:** Data $\{(X_i, D_i, \Delta Y_{i,t})\}_{i=1}^{N+1}$, total miscoverage $\alpha \in (0,1)$, number of folds $K \geq 2$. This is for any arbitary t, which is suppressed in the rest of algorithm.

2. **Calculation of individual effect $DID(i,g,t)$:**

   (a) Calculate a propensity score $\hat{e}(X_i)$ using a logistic regression of D_i on X_i. If $D_i = 1$, set w_i to be 1. If $D_i = 1$, $w_i = \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}$.

   (b) Let $\hat{f}_0 = \sum_{i \in C} w_i \Delta Y_i$.

   (c) Now calculate the estimate as $\Delta Y_j - \hat{f}_0$.

3. **Calculation of conformal interval:**

   (a) Partition the untreated units $1, \cdots, N$ into K disjoint folds $F_1, \ldots, F_K$. Then for each fold $k = 1, \ldots, K$:

      i. Repeat 2(a)-2(c) on all but fold k, obtaining $f_0^{(-k)}$.

      ii. For each $i \in F_k$, compute the out-of-fold prediction for the treated unit's X_j,

$$\hat{f}_0^{(-k_i)}(X_j)$$

      and out-of-fold residual

$$r^{(k_i)} = Y_i - \hat{f}_0^{(-k)}(X_j).$$

   (b) Candidate bounds at treated unit's X_j, for each $i = 1, \cdots, N$ at each i's holdout set k_i,

$$L_i(X_j) = \hat{f}_0^{(-k_i)}(X_j) - |r^{(k_i)}|, \qquad U_i(X_j) = \hat{f}_0^{(-k_i)}(X_j) + |r^{(k_i)}|$$

   (c) Take interval across these bounds,

$$L(X_j) = Q_\alpha\big(\{L_i(X_j)\}_{i=1}^N\big), \qquad U(X_j) = Q_{1-\alpha}\big(\{U_i(X_j)\}_{i=1}^N\big).$$

(d) Take the $1-\alpha$ conformal interval to be

$$(\Delta Y_j - U(X_j), \Delta Y_j - L(X_j))$$

Algorithm 2: Aggregation of intervals for a single treated unit j treated at time g across time(two-sided, level $1-\alpha$)

---

1. **Inputs:** The counterfactual predictions for each post-treatment time period $t = g, \cdots, T$ $\{\{\hat{f}^{(k_i)}_{0\,[t]}(X_j)\}_{i=1}^N\}_t$ the residuals $\{\{r^{(k_i)}_{[t]}\}_{i=1}^N\}_t$ and total miscoverage $\alpha \in (0,1)$,

2. **Calculation of estimate $ATT(j)$:** The estimate is the overtime mean. Weighted means can be used, but is not indicated here to keep notation clean.

$$ATT(j) = \frac{1}{T-g+1}\sum_{t=g}^{T} DID(j,g,t)$$

3. **Calculation of the confidence interval**:

   (a) Take the across time means for $\hat{f}^{(k_i)}_{0\,[t]}(X_j)$ and $r_{[t]}$

$$\tilde{f}_0(X_j) = \frac{1}{T-g+1}\sum_{t=g}^{T} \hat{f}^{(k_i)}_{0\,[t]}(X_j)$$

$$\tilde{r}^{(k_i)}(X_i) = \frac{1}{T-g+1}\sum_{t=g}^{T} r^{(k_i)}_{[t]}(X_j)$$

   (b) Candidate bounds at treated unit's X_j, for each $i = 1, \cdots, N$ at each i's holdout set k_i,

$$L_i(X_j) = \tilde{f}^{(-k_i)}(X_j) - |\tilde{r}^{(k_i)}|, \qquad U_i(X_j) = \hat{\tilde{f}}^{(-k_i)}(X_j) + |\tilde{r}^{(k_i)}|$$

   (c) Take interval across these bounds,

$$L(X_j) = Q_\alpha\big(\{L_i(X_j)\}_{i=1}^N\big), \qquad U(X_j) = Q_{1-\alpha}\big(\{U_i(X_j)\}_{i=1}^N\big).$$

   (d) Take the $1-\alpha$ conformal interval to be

$$(\overline{\Delta Y_j} - U(X_j), \overline{\Delta Y_j} - L(X_j))$$

Algorithm 3: Aggregation of intervals for a several treated unit across time(two-sided, level $1-\alpha$)

1. **Inputs:** The counterfactual predictions for each post-treatment time period for each treated unit $j = N_1, \cdots, N_T$ $\{\{\tilde{f}_{[t]}^{(k_i)}(X_j)\}_{i=1}^N\}_j$ the residuals $\{\{\tilde{r}_{[t]}^{(k_i)}\}_{i=1}^N\}_j$ and total miscoverage $\alpha \in (0,1)$,

2. **Calculation of estimate ATT:** The estimate is the overtime mean. Weighted means can be used, but is not indicated here to keep notation clean.

$$ATT = \frac{1}{N_T - N_1 + 1} \sum_{j=N_1}^{N_T} ATT(j)$$

3. **Calculation of interval estimate: Minkowski mean**

   (a) Calculate the $1-\alpha/T$ conformal interval using the same procedures (a)-(d) above as previous algorithm.

   (b) Take the Minkowski mean o these confidence intervals,

$$(N_T - N_1 + 1)^{-1} \sum_j \left[\overline{\Delta Y_j} - U(X_j), \overline{\Delta Y_j} - L(X_j)\right]$$

4. **Calculation of interval estimate: Assuming independence**

   - For each $j = N_1, \cdots, N_T$, calculate standard errors of $\tilde{f}_{[t]}^{(k_i)}(X_j)_{i=1}^N$ and $\tilde{r}_{[t]}^{(k_i)}\}_{i=1}^N$ as σ_j^f and σ_j^r, respectively.
   - Calculate a normal approximated confidence interval where $c_{\alpha/2}$ is the critical value and σ^* is the associated

$$\sigma^* = (N_T - N_1 + 1)^{-1} \left(\sum_{j=N_1}^{N_T} (\sigma_j^f + \sigma_j^r)^2 \right)^{1/2}$$

   -

$$ATT - c_{\alpha/2} * \sigma^*, ATT + c_{\alpha/2} * \sigma^*$$

6.2 Simulation

This section demonstrate how estimators behave under saturation *ceteris paribus* other complications in real data. We test them under the simulated data generating process Eq. 1, repeated below. For all $t \geq 2$,

$$Y_{i,t}(g) = Y_{i,1} + \sum_{s=2}^{t} \Delta_s + \beta * \mathbb{I}_{\{g>0 \ \& \ t \geq g\}} + \epsilon_{i,t}$$

We fix a constant β at 2.5 as the *treatment effect*, which we hope to recover from simulations. A secular trend of Δ_s between any two consecutive periods. There are 20 treated countries, 4 treated in every period from year 1 to 5. The study period extends to year 21. There are additionally 100 control countries that never democratize. All countries, regardless of treatment status, have the same secular trend; this is necessary for identification. The countries treated at year 1 have no pre-treatment outcomes.

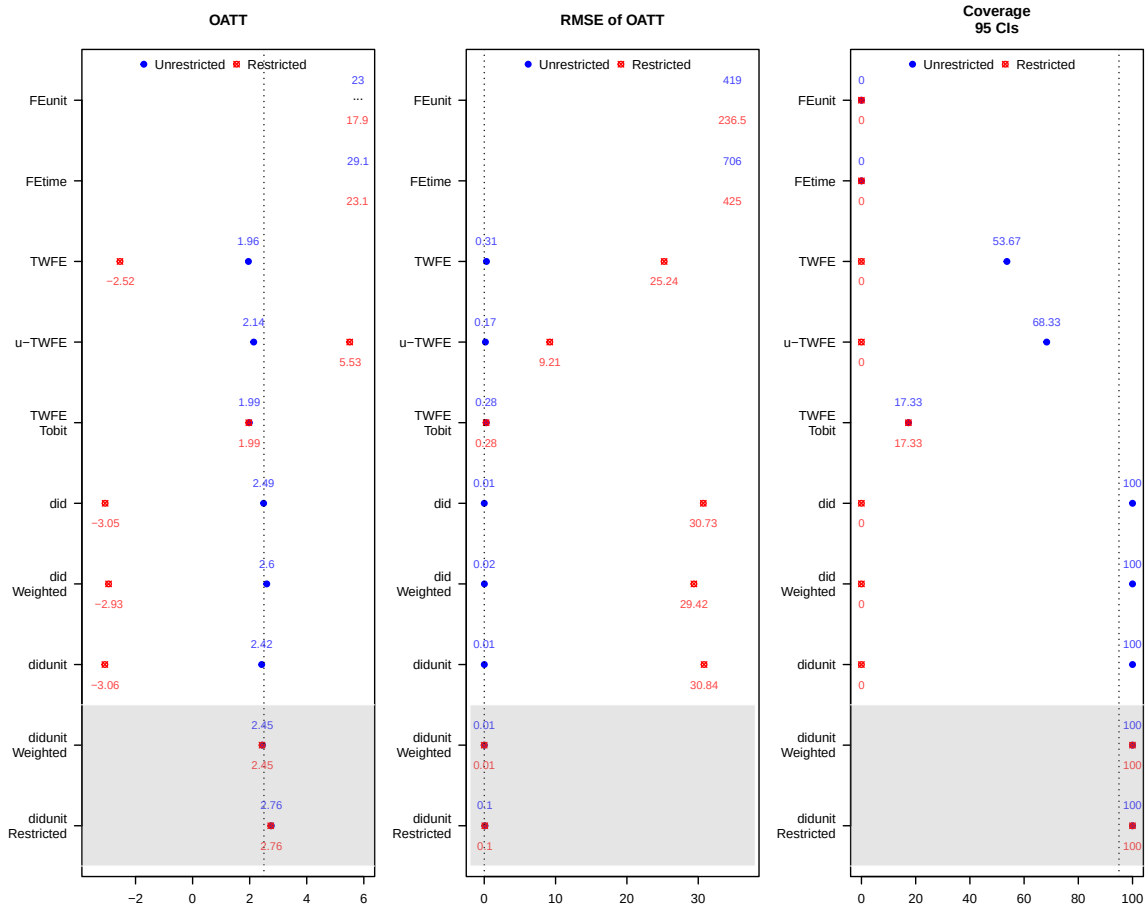
Baselines at year 1 for the 4 treated groups are two of 90 and two of 50. The control group has baselines between 10 to 60. Baselines are chosen to mimic the pattern we usually see with data, which is that several democratizers have more advanced outcomes even before democratization (therefore, closer to the saturation bound) than their comparable control countries. The resulting dataset should have 2520 records with 21 time periods.

We attempt to recover β from data generated under this unconstrained process using 8 different estimators: country fixed effects regression (unitFE), year fixed effects regression (yearFE), two-way fixed effects regression (TWFE), two-way fixed effects estimator with unit-level time trends (u-TWFE), two-way fixed effects with Tobit correction (TWFE Tobit), the unconditional estimator of Callaway and Sant'Anna (2021) (`did`), the baseline-weighted version of the estimator (`did` Weighted), our unit-level adaptation of the CS estimator without baseline-weighting (`didunit`) and with baseline weighting (`didunit` Weighted). The coarsened baseline outcome restricted version is named (`didunit` Restricted).

Then we constrain the outcomes at 100, so that any outcome that was above 100 will now just be 100. Even with a constant β as in process in Eq. 1, truncation can be interpreted as inducing a heterogeneous treatment effect as in Eq. 5. The overall effect that we expect should be the sum of these heterogeneous DiD estimates. For the purpose of this simulation demonstration, we compare the effects to 2.5, as before, interpreting the data actually coming from a process Eq. 1, but truncated at 100.

Figure 6.8 demonstrates these results graphically. Under a data generating process where the outcome is unrestricted the true effect of 2.5 is recovered by all except the unit fixed effect (FEunit) and time fixed effects (FEtime) regressions. These regressions overestimate the effect because they do not account for rising secular trends (in the case of FEunit), or

Figure 6.8: Overall estimates: uncensored data in blue and censored data in red



Note: The figure depicts the recovered estimates under simulated from a data generating process with unrestricted outcomes as in Eq. 1 in blue, and from the same process but where the outcome is restricted above an upper bound of 100 in red. The three panels correspond to their various properties. Left panel OATT is the estimated effect which should be compared to the true effect 2.5. The center panel calculates the root mean square error of the estimate in relation to the true effect of 2.5. The right panel calculates the percentage of times a 95% confidence interval crosses the true effect 2.5.

that treated units have higher baselines (in the case of FEtime).

Under the ceiling restriction, the TWFE estimates a negative effect. The unconditional version of the Callaway and Sant'Anna (2021) estimator, and its base level also report a negative estimate. Our unit-level adaptation estimates the exact same effect as the unconditional Callaway and Sant'Anna (2021) estimator. The unit-level time trends TWFE overestimates the effect.

The Tobit-TWFE recovers the true effect, and so does our baseline-weighted Callaway and Sant'Anna (2021) estimator. The reason our baseline-weighted estimator performs better than the baseline-weighted estimator of Callaway and Sant'Anna (2021) is due to its ability to

baseline-weight at a country level and excise countries with no baseline comparable controls. In this particular simulation, the countries with baseline at 90 will be dropped because there were no comparable controls. The original estimator only drops if the average of the entire cohort of democratizing countries lack overlap, a much stronger sort of overlap violation. The Tobit-TWFE has better coverage than our estimator.

It is no surprise that Tobit performs the best under these conditions. However, the Tobit's utility is contingent on whether the saturation bound is known or at least empirically discoverable. For example, had we incorrectly set the Tobit's censoring bound to be 120, its estimates would look quite similar to that of the two-way fixed effects.

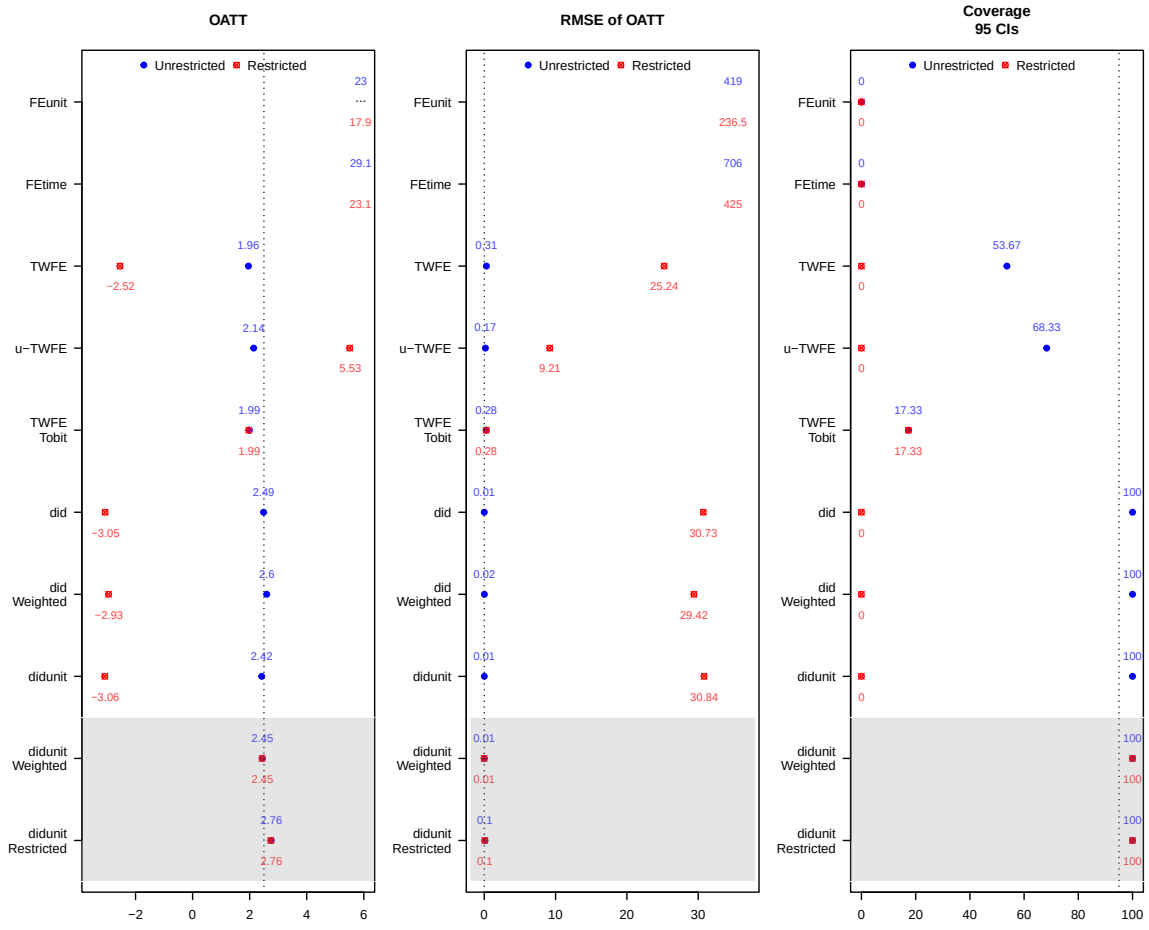
The other circumstance under which Tobit fails is if the two-way fixed effects model is biased even without the restriction in its range. This occurs if $\beta(g)$ differs based on g (Goodman-Bacon 2021). For the five groups, suppose that the treatment effect declines overtime starting from 4.93 to 0.93. Because the first g enter already treated a differences-in-differences desgin should drop this group. The overall effect works out to be about 2.5, once the different post-treatment lag lengths each period is observed is accounted for (i.e: $3.93 * 20 + 2.93 * 19 + 1.93 * 18 + 0.93 + 17$).

In Figure 6.9, Tobit-TWFE performs worse than our proposed unit-level adapted estimator.

In shorter horizons and provided sufficient richness in the set of control countries, our estimator would do just as well as TWFE-Tobit or better under many of these unfriendly conditions. If baseline-weighting, the closer the controls' baselines are to the democratizing country's, the less bias there will be. Coarsened exact matching on the baseline can further reduce these biases. Whether matching or weighting is appropriate is an empirical matter.

Over longer horizons, if the secular trajectory rises continuously, even the control set's outcomes saturate. Tobit performs better here as long as the saturation point is discoverable. Yet, the interpretation of such an estimate assumes the outcome has meaning beyond this saturation point.

Figure 6.9: Overall estimates when $\beta(g)$ differs: uncensored data in blue and censored data in red



Note: The figure depicts the recovered estimates under simulated from a data generating process with unrestricted outcomes but has varying $\beta(g)$ in blue, and from the same process but where the outcome is restricted above an upper bound of 100 in red. The three panels correspond to their various properties. Left panel OATT is the estimated effect which should be compared to the true effect 2.5. The center panel calculates the root mean square error of the estimate in relation to the true effect of 2.5. The right panel calculates the percentage of times a 95% confidence interval crosses the true effect 2.5.