

Unintended Consequences of Overlap Violations in Covariate-adjusted Causal Regressions

Ransi Clark

September 8, 2026

Abstract

Lack of covariate overlap is a perennial challenge in empirical social science. When covariate overlap is poor, even if the covariates are relevant in the regression estimation, the imbalance can induce a bias that can have consequences such as sign swapping. This is because the regression's internal propensity weighting of outcomes within the treatment and untreated samples become negative. If those negative weights are associated with sufficiently extreme outcomes, the estimated treatment effect is biased and possibly of a different sign. Yet the exclusion of covariates in the regression is not the solution. Instead I propose a detect-and-correct approach where the observations associated with the negative weights due to poor overlap are identified and given zero weight. I recommend a normalized overlap weighting algorithm over the more classic inverse probability weighting. I demonstrate the extent of this concern in five studies within the electoral accountability literature. Negative weights are ubiquitous in all applications. Therefore, the recovered treatment effect differs with the overlap-weighted correction in most instances. The consequence of the correction varies. In some studies, the sign of the estimate reverses after correcting for negative weights. In others, the estimate becomes stronger. This tendency of regressions can make regression-derived causal estimates vastly different from those estimated using synthetic controls and post-2020 differences-in-differences methods which explicitly choose sets that have overlap. This paper helps researchers diagnose when those differences are caused by overlap violations.

Covariate overlap is a precondition in causal interpretation. But covariate imbalance is endemic in observational settings. Consider such instances as calculating the effect of hurricanes on incumbent vote share, the effect of wildfires on voting for climate change mitigation policies, or

the effect of fatal school shootings on Democratic vote share. The type of voting districts that are affected by hurricanes, wildfires, and school shootings are quite different from those that are not for characteristics other than the effect of these events.

Researchers use a variety of methods to algorithmically achieve this balance between treated and untreated sets without achieving exact covariate balance. Matching methods choose a untreated comparison set to mimic the characteristics of the treated set, but discards many untreated observations that are sufficiently different from the treatment set (Iacus et al., 2012; Rosenbaum & Rubin, 1983).

Weighting methods attempts to use all untreated observations, but gives more importance to those that are similar to the treatment set, where similarity is determined using a propensity score calculation (Li et al., 2019; Robins, 1986) or a convex optimization such as that used by synthetic controls (Abadie, 2021). Discontinuity designs attempt covariate overlap by making comparisons only within a narrow set of treated and untreated units (Hahn et al., 2001).

Regression specifications fall under these weighting methods. In a regression setting, when the user specifies a set of covariates that account for the differences in characteristics between a treated and untreated set, the regression weights these sets to achieve balance using the distribution of those covariates. However, important differences in the weighting algorithms can create large discrepancies between the estimates recovered by matching and other weighting methods previously mentioned. This happens because regressions manage overlap violations by extrapolating into negative propensity scores that then become negative weights when calculating the weighted sum of treated units and weighted sum of untreated units.

Despite a burgeoning menu of matching and weighting techniques, regressions remain the most common implementation of recovering average treatment effects, particularly when panel data is involved, and researchers want to control for unobserved unit and time effects in addition to the observed covariates. While the covariates must be included in the specification to avoid omitted variable biases, their inclusion can also aggravate a certain type of bias, with sufficient lack of covariate overlap. The first objective of this paper is to demonstrate the conditions under which this bias arises.

The second objective is to show that a simple post-regression diagnostic test can reveal how much of the average treatment effect is due to the negative weighting. The diagnostic test recasts

the regression estimate of the average treatment effect as a difference of two sums with the first sum being the weighted sum of treated units and the second sum being the weighted sum of untreated units. The weights of the sums are a function of the linear propensity score model internally calculated by the regression. When there is poor covariate overlap, this propensity score model calculates treatment propensities above 1 or below 0, leading to negative weights. The presence of negative weights, by itself, does not lead to bias. For instance, the outcomes associated with the negative weights could be close to zero.

To see whether the negative weights are at risk of causing bias, each of the treated and untreated weighted sums are further decomposed to the parts of the sum associated with positive weights and the sum associated with the negative weights. This allows the user to see how much of the recovered treatment effect is due to negative weights. If the outcome sum associated with negative weights are negligible in context of the scale of the outcome variable, the overlap violations are of little concern.

The third objective of the paper is to recalculate the treatment effect by eliminating the impact of the overlap-violation induced negative weights. I consider two routines, inverse propensity score weighting and the more recent overlap weighting. In simulation exercises, overlap weighting performs better in reducing bias and achieving confidence interval coverage as it avoids the large weights the the inverse probability weight tends to impose on a few units.

Finally, I apply this detect-and-correct approach to five studies in the electoral accountability literature. Four of five applications use two-way fixed effects regressions with covariate adjustment to estimate average treatment effects. In all specifications, regressions produce substantial negative weights, almost always in the untreated sum. The percentage of negative weights in the untreated sum range from 25% to 70%. The outcomes associated with the negative weights are also substantial and in some cases, overwhelm the outcomes sum associated with the positive weights where there is common support, i.e. the region of the covariate space where there is sufficient overlap of treated and untreated units.

Recalculating treatment effects with overlap weighting shows the negative weighting matters. In the case of Hazlett and Mildemberger (2020), which explores the effect of exposure to wildfire on support for pro-climate policies, regression’s negative weighting created sufficient bias to swap the sign of the estimated effect, affecting the conclusions of the study. In the case of García-Montoya et

al. (2022)’s study on the effect of shooting on Democratic vote share, the negative weighting of the regression had attenuated the regression’s recovered effect. When negative weights were relatively few, as in Markarian and Newman (2025)’s study of a mass shooting on gun control policy voting, the recovered effects were mostly in agreement. In still other cases, such as in Gasper and Reeves (2011), which studies effect of natural disaster declarations on incumbent vote share, reweighting affected both the treatment and outcome sums in a similar direction, which left the estimated effect unchanged by coincidence.

The issue I raise here is not endemic to fixed effects regressions, but I demonstrate the problems in that context because of the ubiquity of such regressions in applied research. In recent years, there have been other concerns about the use of two-way fixed effects regressions because of the possibility of negative weights on dynamic treatment effects (De Chaisemartin & d’Haultfoeuille, 2023; Imai & Kim, 2021). I abstract from these other critiques of regressions and deal explicitly with the negative weights introduced by the regression’s covariate adjustment algorithm.

The rest of the paper proceeds with formal definitions of overlap violation, the derivation of the diagnostic, and then introduces the competing methods. Then I move to applications.

1 Detecting covariate overlap violations

This section formally introduces covariate overlap violations and how it manifests in a treatment effect calculated by regressions. This analysis is the basis for detecting the extent of bias overlap violations induce in regression-calculated treatment effects. I begin by defining average treatment effects, propensity scores, and covariate overlap to pin down the exact objects of discussion. Then I introduce the mechanics of the diagnostic decomposition, which is based on Frisch-Waugh-Lovell theorem that gives an algebraic representation of the coefficient of any covariate in a multivariate regression (Frisch & Waugh, 1933; Lovell, 1963).

1.1 Definitions

1.1.1 Average treatment effect

The average treatment effect of any binary event D_1 is estimable as the difference between two weighted sums of the outcome Y . Letting the set of treated units be represented by T and the set

of untreated units be represented as C , the average treatment effect of an event D_1 would take the following general form.

$$A\hat{T}E = \sum_T w_i Y_i - \sum_C w_i Y_i \quad (1)$$

In most methods intended for experimental analysis such as inverse probability weighting the weights w_i are imposed to be positive, but this is not true of the implied regression weights. Although not generally the case, for the purpose of this paper I also impose that w_i 's within each of the sets sum to 1.

1.1.2 Propensity score

The propensity score is the probability of being treated given a vector of covariates X .

$$e(X) = \mathbb{E}[D_1 = 1|X] \quad (2)$$

Though the probability of treatment cannot be above 1 or negative, if a linear probability model is used to estimate $e(X)$, the quantity can exceed 1 or be negative. Incidentally, this is how regressions estimate propensity scores internally and where the concern for negative weights arise from.

1.1.3 Overlap

Closely related to the propensity score is that of the overlap assumption, the property also known as common support or positivity. This says that the probability of treatment in a causal study for any given profile of covariates X must be interior to $[0, 1]$. Notationally, this means that,

$$0 < e(X) < 1. \quad (3)$$

When this is not true, overlap is violated for that X . Theoretically, the overlap condition is the requirement that any untreated unit has some positive chance of being treated, and any treated unit has some positive change of not being treated. Under a strictly theoretical view, even if the analysis includes covariate profiles that are never treated, the user proceeds as if a chance of

treatment exists.

Practically, however, propensity scores outside of this unit interval or even too close to 0 or 1 create problems. In regressions, an overstretched propensity score can become a negative weight within the component sums of 1. In inverse probability weighting, a propensity score of 0 or 1 becomes an inverse probability weight of positive or negative infinity. Sufficiently large weights lead to large standard errors when calculating the average treatment effects. There are several common techniques to deal with overlap violations caused by propensity scores close to bounds. One is trimming, where observations associated with covariate profiles that are close to the extreme propensities are removed (Crump et al., 2009). The other is clipping or weight truncation, where propensity scores that come close to bounds are changed to a value further interior in $[0, 1]$ (Cole & Hernán, 2008). The original synthetic control uses an extreme version of trimming, whereby if the pre-treatment covariate profile of the treated unit lives outside the convex hull of its comparison units then the estimation cannot proceed for that treated unit (Abadie, 2021). In the multi-period differences-in-differences estimator of Callaway and Sant’Anna, 2021, sets of treatment units that have no comparable untreated units are removed from the analysis¹.

1.2 Decomposition of the regression estimate for diagnosing overlap failure

The regression coefficient τ associated with the binary D_1 in the following regression model can be recasted in the form of Equation 1 where the weights are some function of 2. The generic regression model is,

$$Y = FE + \tau D_1 + \gamma X. \quad (4)$$

Here, Y is the outcome of interest, D_1 is the binary treatment indicator, and X is the covariate adjustment. FE represents any type of fixed effects: unit and time, unit only, or even just the intercept. I assume that the model always has an intercept or fixed effects, i.e., it cannot have neither.

Using the theorem of Frisch-Waugh-Lovell (FWL, for short)², the quantity recovered for τ by implementing Regression 4 in any statistical software can also be recasted as the orthogonal

¹Trimming is the algorithm’s default setting, but users can change to retain any estimates

²The theorem is restated in Appendix A

projection off of the combined matrix $\tilde{X} = [X \quad FE]$. Notate that orthogonal projection as $M_{\tilde{X}}$. Then from the FWL theorem, isolate the representation of $\hat{\tau}$ as

$$\hat{\tau} = (\widetilde{D_1}^\top \widetilde{D_1})^{-1} \widetilde{D_1}^\top Y \quad (5)$$

where $\widetilde{D_1} = M_{\tilde{X}} D_1$.

Since D_1 is a single vector, the matrix notation can be written as a summation equivalent to Equation 5.

$$\hat{\tau} = \frac{\sum_i \widetilde{D_{1i}} Y_i}{\sum_i \widetilde{D_{1i}}^2}$$

Even more succinctly,

$$\hat{\tau} = \sum_i \omega_i Y_i \quad \text{where} \quad \omega_i = \frac{\widetilde{D_{1i}}}{\sum_i \widetilde{D_{1i}}^2}$$

The ω_i across all i sum to zero (See corollary A.2 in Appendix). But because D_1 is also binary, ω_i also sum to +1 and -1 within the set of treated units and untreated units (See corollary A.3 in Appendix), which I previously denoted T and C in Equation 1.

Therefore, letting,

$$T = \{i : D_{1i} = 1\}$$

$$C = \{i : D_{1i} = 0\}$$

$$\sum_{i \in T} \omega_i = \sum_{i \in T} w_i^{(1)} = 1 \quad \text{and} \quad \sum_{i \in C} \omega_i = \sum_{i \in T} -w_i^{(0)} = -1 \quad (6)$$

Hereafter, the weights denoted $w_i^{(1)}$ and $w_i^{(0)}$ will be referred to as FWL weights. Using the FWL weights, the estimated treatment effect $\hat{\tau}$ can be written as a difference of two sums where the weights add to 1.

$$\hat{\tau} = \underset{\text{treated sum}}{Y^T} - \underset{\text{untreated sum}}{Y^C} \quad (7)$$

where

$$Y^T = \sum_{i \in T} w_i^{(1)} Y_i \quad \text{and} \quad Y^C = \sum_{i \in C} w_i^{(0)} Y_i$$

Though the weights add to 1 inside of each sum, they are not necessarily positive within each sum. To see this, consider that $\widetilde{D}_{1i}$ is the residual of the regression of D_1 on $\tilde{X}$. Because regression residuals must add to 0, they cannot be all positive. The fitted value of this regression of D_1 on $\tilde{X}$ is the regression's calculation of the propensity score in Equation 2. But it does not constrain that the fitted value is within $(0, 1)$, so that when the difference between D_1 and that fitted value is taken some treated units will have a fitted value above 1 and therefore a negative $\widetilde{D}_{1i}$ and a negative $w_i^{(1)}$. Similarly, some untreated units will have fitted value below zero, so that when a difference between 0 and a negative fitted value is taken they will have a positive $\widetilde{D}_{1i}$, but a negative $w_i^{(0)}$. In terms of the definition of 3, the negative FWL weights arise when overlap is violated.

Notationally, further sets can be defined depending on whether the sum carries a positive or negative FWL weight. I call these sets T^+, T^-, C^+ , and C^- .

$$T^+ = \{i : D_{1i} = 1 \quad \text{and} \quad w_i^{(1)} \geq 0\}$$

$$T^- = \{i : D_{1i} = 1 \quad \text{and} \quad w_i^{(1)} < 0\}$$

$$C^+ = \{i : D_{1i} = 0 \quad \text{and} \quad w_i^{(0)} \geq 0\}$$

$$C^- = \{i : D_{1i} = 0 \quad \text{and} \quad w_i^{(0)} < 0\}$$

The treatment effect $\hat{\tau}$ estimated from the regression can be further expressed as,

$$\hat{\tau} = \underbrace{Y^{T^+}}_{\text{treated sum w. +ve}} + \underbrace{Y^{T^-}}_{\text{treated sum w. -ve}} - \underbrace{Y^{C^+}}_{\text{untreated sum w. +ve}} - \underbrace{Y^{C^-}}_{\text{untreated sum w. -ve}} \quad (8)$$

where

$$\begin{aligned}
Y^{T+} &= \sum_{i \in T^+} w_i^{(1)} Y_i \\
Y^{T-} &= \sum_{i \in T^-} w_i^{(1)} Y_i \\
Y^{C+} &= \sum_{i \in C^+} w_i^{(0)} Y_i \\
Y^{C-} &= \sum_{i \in C^-} w_i^{(0)} Y_i
\end{aligned}$$

Ideally, the Y^{T-} (treated sum with negative FWL weights) and Y^{C-} (untreated sum with negative FWL weights) are small in relation to other sums. The sums can be small because there are fewer overlap violations, or the Y_i associated with the negative weights are close to zero. Or both the negative-weight sums are of similar size and cancel each other out. In those ideal cases, the treatment effect is not biased by a lack of covariate balance.

But when one or both of Y^{T-} and Y^{C-} sums is overwhelmingly large, the resulting $\hat{\tau}$ can be a different sign to the effect calculated only from units where overlap condition was met. Sign-changes are not the only problem. The negative-weight sum or sums can overestimate $\hat{\tau}$ making the effect seem statistically significant, when it is not, even if the sign is the same. Or it can underestimate $\hat{\tau}$ making the effect seem statistically insignificant.

2 Simulating covariate overlap violations

This section uses decomposition of 8 on simulated data that is explicitly generated to demonstrate the problem of overlap violations. I generate two scenarios of overlap violations in a simulation using a single covariate. The first scenario involves a quadratic non-linearity in the tail of the distribution of $Y|X$. The second scenario involves a censoring-type non-linearity in the tail of the distribution of $Y|X$.

The panel data with a unit and time indicator is generated as follows³. For a fixed number of

³The data need not be a panel dataset, but I choose a panel dataset because it reflects the empirical applications I discuss later. Despite being a panel setting, however, it is not a staggered treatment setting where treatment is

units G indexed by g whose latent value of the covariate is X_g , the single covariate X_{gt} is drawn from $\mathcal{N}(X_g, \sigma_X^2)$. The overlap violation is imposed by a logit model. The index of the logit is

$$\eta_{gt} = \alpha + \beta X_{gt},$$

where β controls for the extent of the overlap violation as it increases the probability of being assigned to treatment as X increases. Simulated treatment status is derived from

$$\mathbf{P}(D_{1,gt} = 1 | X_{gt}) = \frac{\exp(\eta_{gt})}{1 + \exp(\eta_{gt})}$$

If this probability is greater than 0, the treatment status of the associated X_{gt} is 1, otherwise 0. A multiple-treatment scenario can also be generated with a multinomial logit index.

The quadratic tail bias of the outcome is imposed by a user-specified bias cutoff c and direction $\{\text{left, right}\}$. Let this additional bias term be $h(X_{gt})$ which for a left tail quadratic bias is

$$h(X_{gt}) = (c - X_{gt})^2 \{X_{gt} < c\}.$$

.

Along with a user-specified additive treatment effect τ , a user-specified latent Y_g for the units in group g , a user-specified bias extent b that controls the extent and direction of quadratic bias, a user-specified coefficient on X_{gt} γ , the outcome Y_{gt} associated with X_{gt} is

$$Y_{gt} = Y_g + \gamma X_{gt} + \tau D_{1,gt} + \underset{\text{bias}}{bh(X_{gt})} + \epsilon_{gt}.$$

.

In the above equation, if $b = 0$, there is no quadratic bias. And if, $b > 0$, the outcome has a quadratically rising tail after the cutoff c , and conversely for $b < 0$ the outcome has a quadratically falling tail. The left panels of Figure 1 demonstrates data generated with a quadratic bias parameter of -11 at a cutoff of -0.5 . On the top-left, the data is generated with $\beta = 1$, so that in region of X left of -2 , no treated units are observed, and in the region of X right of 0.5 , no untreated units are once-received-always-received. Instead, a unit of group g treated in one period can be untreated the next.

observed. On the top-right, the data is generated with $\beta = 0$, where treatment status is distributed across the range of X .

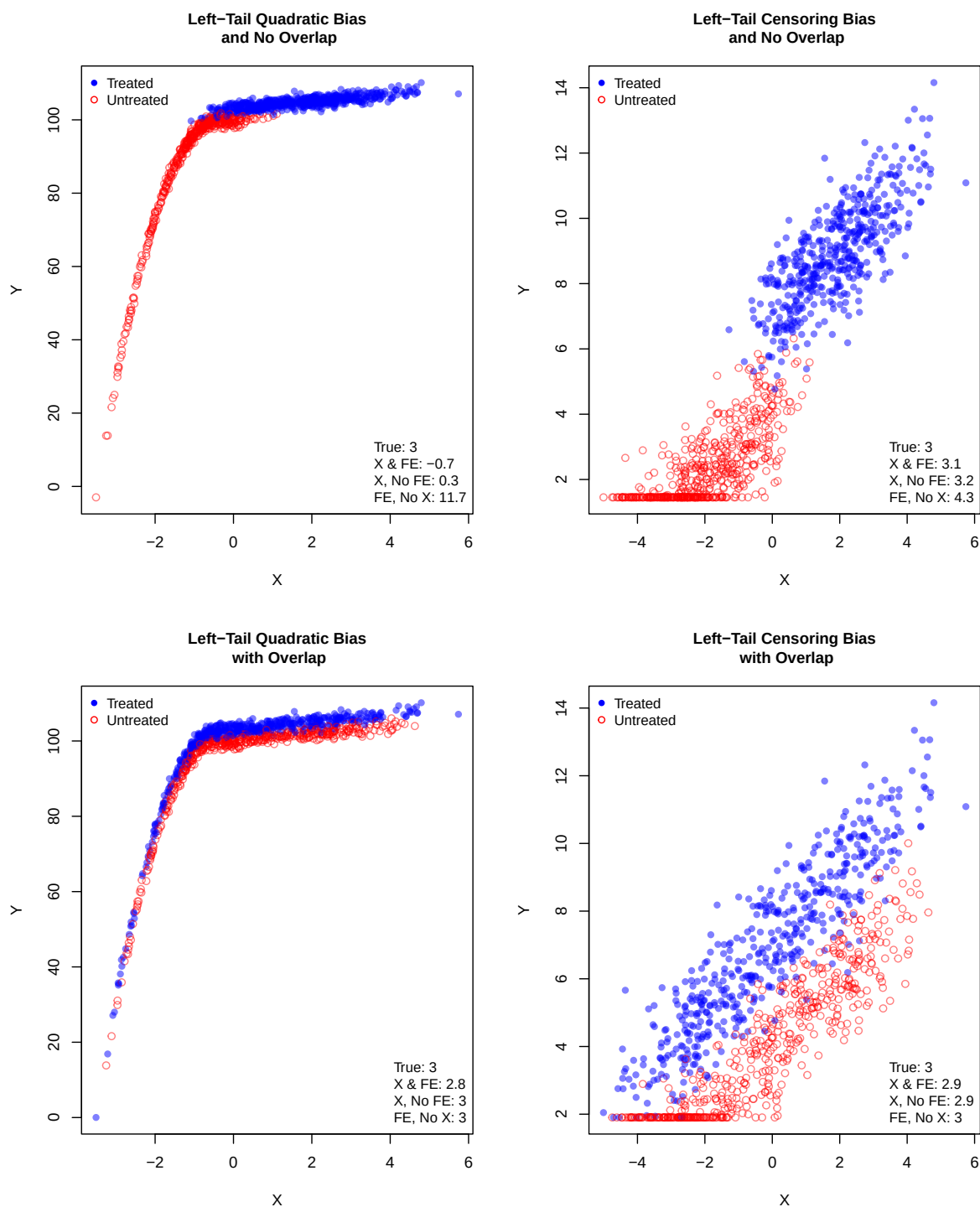


Figure 1: Regression estimates under two types of left-tail bias (quadratic and lower censoring) with and without sufficient covariate overlap

To impose a censoring bias the user specifies the cutoff, the direction, and if the censoring is lower or upper censoring. The top-right panel of Figure 1 demonstrates such a bias where overlap is also violated. The censoring is on the left tail, with lower censoring to the left of -0.5. The bottom-right panel of Figure 1 illustrates data where overlap is satisfied but the outcome is still censored.

For all the four simulations in Figure 1 the true additive treatment effect is 3, $\tau = 3$. On the bottom-right corner of each of the four panels, I derive the regression estimate of τ that will be estimated by the illustrated data under the three regression models below. I use FE as a shortening for both time and unit fixed effects.

$$X \text{ \& FE: } Y = FE + \tau D_1 + \gamma X$$

$$X, \text{ No FE: } Y = \text{Intercept} + \tau D_1 + \gamma X$$

$$\text{FE, No X: } Y = FE + \tau D_1$$

The printed numbers are for the average of those estimates under 500 simulations under each scenario, though the illustrated point data is from a single representative simulation. With the large quadratic bias and insufficient overlap (top-left panel), the regression estimates with fixed effects and covariate adjustment recovers a negative effect of -0.6, signed opposite to the true effect of 3.

The other models do not recover the true effect either. Dropping the fixed effects still severely underestimates, and dropping the covariate adjustment severely overestimates. In the bottom-left panel where despite the large quadratic bias, overlap is satisfied, the fixed effects and covariate adjustment recovers the true effect, and so does the model with only fixed effects. Exclusion of fixed effects underestimates, but only slightly.

The bias effected on τ by left censoring without sufficient overlap is less egregious (top-right panel). Yet, it slightly overestimates and if the true τ is close to zero, sign swaps are still possible. The exclusion of the covariate adjustment worsens the bias. With overlap, the recovered effects are at or close to the true τ of 3 under all regression models.

Figure 2 demonstrates the FWL weights from 6 under the four scenarios of the previous Figure 1. The top panels have poor overlap. As predicted by the derivation of the weights, the FWL weights

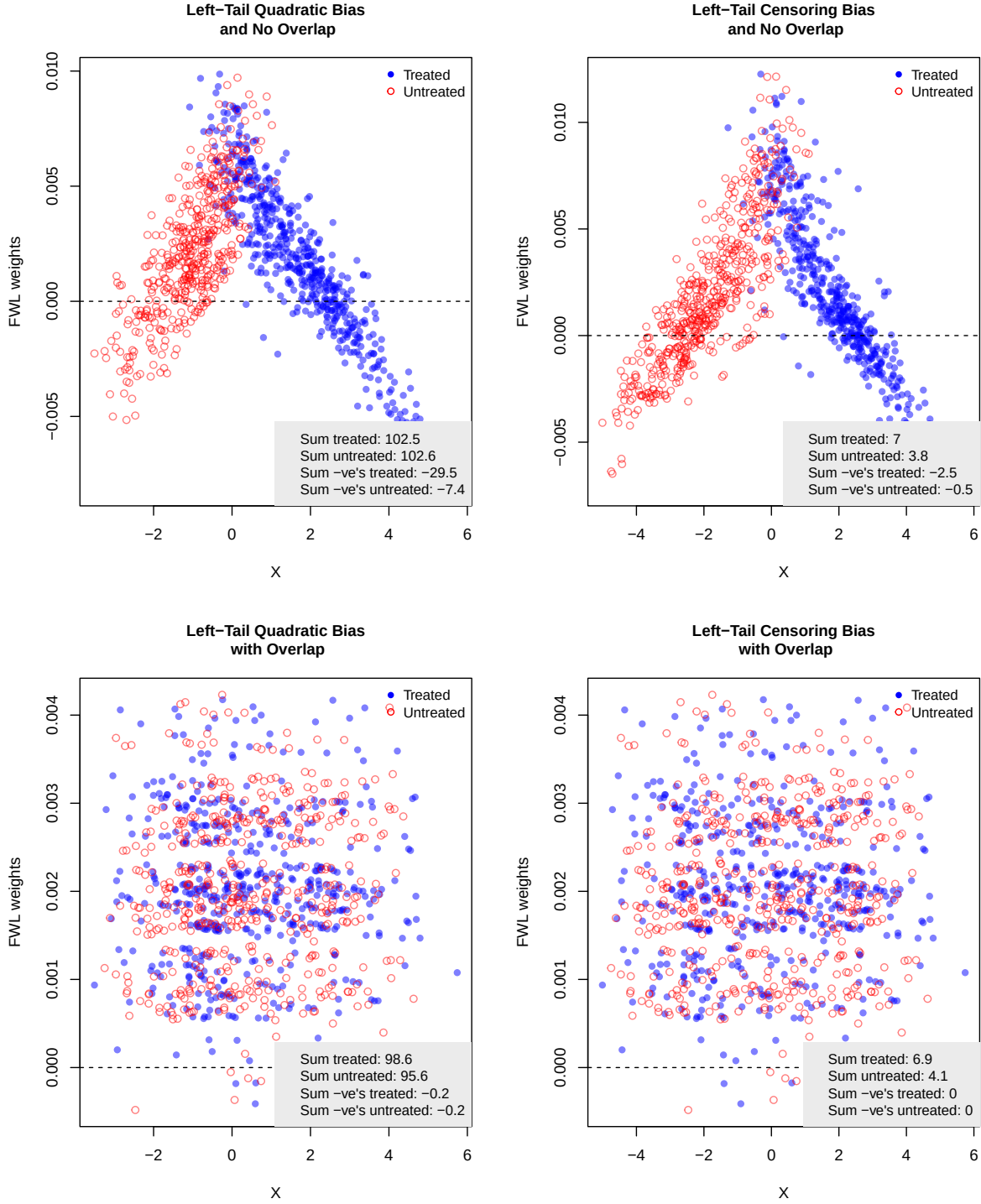


Figure 2: FWL weights under two types of left-tail bias (quadratic and lower censoring) with and without sufficient covariate overlap

$w_i^{(1)}$ (solid blue points) are negative in the region of X with poor overlap. Similarly, the FWL weights $w_i^{(0)}$ (open red points) are also negative in the region of X with poor overlap. Both bottom

panels of Figure 2 have sufficient overlap, and therefore very few negative weights. When negative weights are present they are also very small compared to the larger positive weights.

For overlap problems to create biases, however, they must be associated with large or extreme Y values. On the bottom-right corner of each panel is the average of four quantities, across 500 iterations, of the data generating processes of the four panels in Figure 1. These quantities are

- Y^T from Eq 7 as ‘Sum treated’
- Y^C from Eq 7 as ‘Sum untreated’
- Y^{T-} from Eq 8 as ‘Sum -ve’s treated’
- Y^{C-} from Eq 8 as ‘Sum -ve’s untreated’

Recalling that the outcomes of the left panels of Figure 1 had a large negative quadratic decline left of -0.5 , the associated Y^{C-} and Y^{T-} from Eq. 8 are non-negligible. A sign-swap arises out of these conditions. With left-tail censoring the outcomes associated with the negative weights on the left-tail are smaller than with the quadratic bias, so the components associated with negative weights are comparatively smaller, and the bias is also small. This is even if the sum of negative weights $\sum_{i \in T-} w_i^{(1)}$ and $\sum_{i \in C-} w_i^{(0)}$ are similar across the two left panels.

When overlap is met, as in the bottom panels of Figure 1 and Figure 2, the Y^{T-} and Y^{C-} are zero or close to zero. The estimated $\hat{\tau}$ is composed mostly out of Y^{T+} and Y^{C+} in the region where overlap is satisfied.

3 Correcting for covariate overlap violations

While some number of negative weights $w_i^{(1)}$ and $w_i^{(0)}$ is harmless and unavoidable with fixed effects, when such weights are overwhelmingly associated with large Y_i ’s the average treatment effect estimates are biased. Simulation results accompanying 1 show that excluding the covariate in the regression model and simply calculating a regression with fixed effects does not recover the true effect under lack of overlap. Excluding the fixed effect does not improve bias either. Even once a user is aware of violations in covariate overlap, little can be done by altering the regression model that is estimated.

A crude correction would be to restrict any negative $w_i^{(1)}$ and $w_i^{(0)}$ to be zero and rescale the remaining positive weights so that they add to 1 within the treated set T and untreated set C . However, many established weighting methods do a similar function to such a correction, their key functionality being that the propensity score is restricted within the unit interval. I compare two weighting methods to the FWL weighting internal to the regression. The first is the familiar inverse probability weighting (IPW), and the second is the more recently developed overlap weighting (OW). Even normalized (Hajek-type) inverse probability weighting tends to place high weight on a small subset of observations resulting in higher instability. Overlap weighting mitigates this problem without requiring special intervention such as trimming (Li et al., 2018).

I continue with the panel data setting, but the process is similar with just unit fixed effects and the intercept. The starting point for both weighting methods (IPW and OW) is propensity score estimation via logit regression⁴. I first estimate the propensity score $e_i = \mathbb{E}(D_1|X_{gt}, g, t)$ given X_{gt} and the group id g and time id t using the logit regression.

$$\text{logit}(e_i) = \alpha_g + \alpha_t + X_{gt}^\top \beta$$

The fitted values $\hat{e}_i$ derived from this regression are always between 0 and 1, unlike the fitted values derived from the linear projection of D_1 into the combined matrix of covariates and fixed effects $\tilde{X}$. Even if certain covariate profiles perfectly predict D_1 , these fitted values never go outside the unit interval.

For perfectly *separated* groups, i.e. groups that are either treated in all t or untreated in all t , the logit likelihood is optimized by setting α_g to $-\infty$ or $+\infty$, respectively. So that the predicted $\hat{e}_i$ for these units are 0 and 1. The linear projection of D_1 into $\tilde{X}$, in contrast, gives $\hat{\alpha}_g$ a finite value, and because the sum of these fitted values $\widehat{D}_{1gt}$ over t must add to 0, some of these are negative and their corresponding Y_{gt} are associated with negative FWL weights. This is not necessarily a consequence of overlap violation. However, the occurrence of overlap violations in the X -space is correlated with the occurrence of units that are either always-treated or always-untreated.

When the logit likelihood optimization forces the α_g of units that are always-treated or always-

⁴Other discrete variable methods could work in place of logistic regression, but I do not implement them in the package

untreated into $-\infty$ or $+\infty$, the optimization algorithm may fail to converge⁵ Even if it converges, the extreme propensity scores result in inverse probability weights close to $-\infty$ or $+\infty$.

One solution to this latter issue is clipping. Supposing the user chooses a clipping parameter of $\epsilon > 0$, a clipped propensity score will be forced to be within $[\epsilon, 1 - \epsilon]$ instead of $[0, 1]$. So that if the estimated propensity score is less than ϵ , then $\hat{e}_i$ is replaced by ϵ . Clipping ϵ can be reduced to something very close to zero for overlap weighting. Practically this makes little difference for small values of ϵ if there are enough units in T and C .

The two alternative weighting routines diverge here, but the goal is to obtain weights in lieu of FWL weights $w_i^{(0)}$ and $w_i^{(1)}$.

3.1 Normalized Inverse Probability Weighting

The first weighting routine is the normalized version of the familiar inverse probability weighting (Robins, 1986). The relevant weights for the treated and untreated sets T and C are

$$w_i^{IPW,(1)} = \frac{\frac{1}{\hat{e}_i}}{\sum_{j \in T} \frac{1}{\hat{e}_j}}$$

$$w_i^{IPW,(0)} = \frac{\frac{1}{1-\hat{e}_i}}{\sum_{j \in T} \frac{1}{1-\hat{e}_j}}$$

Then the inverse probability weighted equivalent to Eq. 1 is

$$A\hat{T}E^{IPW} = \sum_T w_i^{IPW,(1)} Y_i - \sum_C w_i^{IPW,(0)} Y_i \quad (9)$$

Instability is a concern in inverse probability weighting because calculated propensity scores $\hat{e}_i$ close to 0 or 1 can make the weights large. Clipping mitigates this somewhat. Normalizing inverse probability weights further mitigates large weights. However, simulation studies show that under worsening covariate imbalance $A\hat{T}E^{IPW}$ can worsen bias over even regression adjustment (Li et al., 2019).

⁵An alternative to this is to impose only a random effect.

3.2 Normalized Overlap Weighting

The second weighting routine is the normalized version of the overlap weights (Li et al., 2018). The relevant weights are

$$w_i^{OW,(1)} = \frac{1 - \hat{e}_i}{\sum_{j \in T} 1 - \hat{e}_j}$$

$$w_i^{OW,(0)} = \frac{\hat{e}_i}{\sum_{j \in T} \hat{e}_j}$$

Then the inverse probability weighted equivalent to Eq. 1 is

$$A\hat{T}E^{OW} = \sum_T w_i^{OW,(1)} Y_i - \sum_C w_i^{OW,(0)} Y_i \quad (10)$$

Overlap weighting puts the highest weight on units in the common support region and is more stable because it smoothly reduces the influence of those units whose propensity scores are extreme.

3.3 Standard errors

The standard method of deriving treatment effect variance in regressions is to isolate the top-left element of the matrix

$$\text{Var}([\hat{\tau} \ \hat{\gamma} \ \dots]) = \hat{\sigma}^2 ([D_1 \ \tilde{X}]^\top [D_1 \ \tilde{X}])^{-1} \quad (11)$$

where $\hat{\sigma}^2$ is variance of the residuals between the observed outcome Y and that implied by the regression outcome model in 4. IPW or OW routines don't assume an outcome model, so standard methods of variance calculation for those associated ATEs involve the weights directly. To compare the routines under similar algorithms, I calculate the variance of the expressions in 1, 9, and 10 using the weights directly. To keep matters simple, I also assume the weights are known, but the implementation provides a bootstrap option that accounts for the uncertainty of calculation of weights. For any of those ATEs, the algorithm assuming known weights is calculated as

$$\text{Var}(\widehat{ATE}) = \sum_{i \in T} (w_i^{(1)})^2 (Y_i - \hat{\mu}^T)^2 - \sum_{i \in C} w_i^{(0)^2} (Y_i - \hat{\mu}^C)^2 \quad (12)$$

where $\hat{\mu}^T$ and $\hat{\mu}^C$ are treated and untreated sums respectively.

In the case of the linear regression, the outcome of 12 differs from the variance of $\hat{\tau}$ obtained from 11. In many overlap-violated applications, the variance from 12 will be larger than $\text{Var}(\hat{\tau})$ 11 since larger deviations from the treated or untreated sums are also likely to be associated with larger absolute values of weights $w_i^{(1)}$ or $w_i^{(0)}$. But cluster-robust standard errors can exceed 12, even if the homoskedastic-assumed 11 is smaller.

3.4 Applying to previous simulations

Figure 3 demonstrates the weights from each routine. The top row of this figure is the same as that of Figure 2 and demonstrates the FWL weights $w_i^{(0)}$ in open points and $w_i^{(1)}$ in solid points. The center row shows the weights inverse probability weights $w_i^{IPW,(0)}$ and $w_i^{IPW,(1)}$. And the bottom row shows overlap weights $w_i^{OW,(0)}$ and $w_i^{OW,(1)}$. All the weights add to 1 within their respective set T or C . FWL weights are not necessarily positive, but both inverse probability weights and overlap weights are. IPW and OW weights are almost zero outside of the common support region of the covariate X .

Even with normalization and clipping⁶, the inverse probability weighting routine put a large amount of weight on a few observations (highest at 0.3). Because the weights add to 1 within the respective comparison set, most other weights — even those that are in the common support region — are down-weighted to zero. Overlap weights do not have such extreme weights, with its highest weight about one-tenth of the highest inverse probability weight. This makes overlap weighting more stable and the preferred weighting scheme. Though less prone to extreme weights, overlap weighting upweights observations in the common support region more aggressively than regression weights.

The top-left corner of each panel of Figure 3 also reports the mean bias and 95% confidence interval coverage of 500 simulations of the same data generating processes as in the top row of Figure 1. As was also discussed in the context of top-left panel of Figure 1, the negative weights associated with large negative quadratic left tail biases the treatment effect sufficiently enough to recover the wrong sign. Because of the bias, coverage of the confidence interval is also poor. Inverse probability weighting (center-left panel of Figure 3) does not improve matters. It recovers the correct sign but overshoots in that direction. The coverage is worse than the regression. Overlap

⁶A clipping parameter of 0.005 was chosen for the IPW and OW routines.

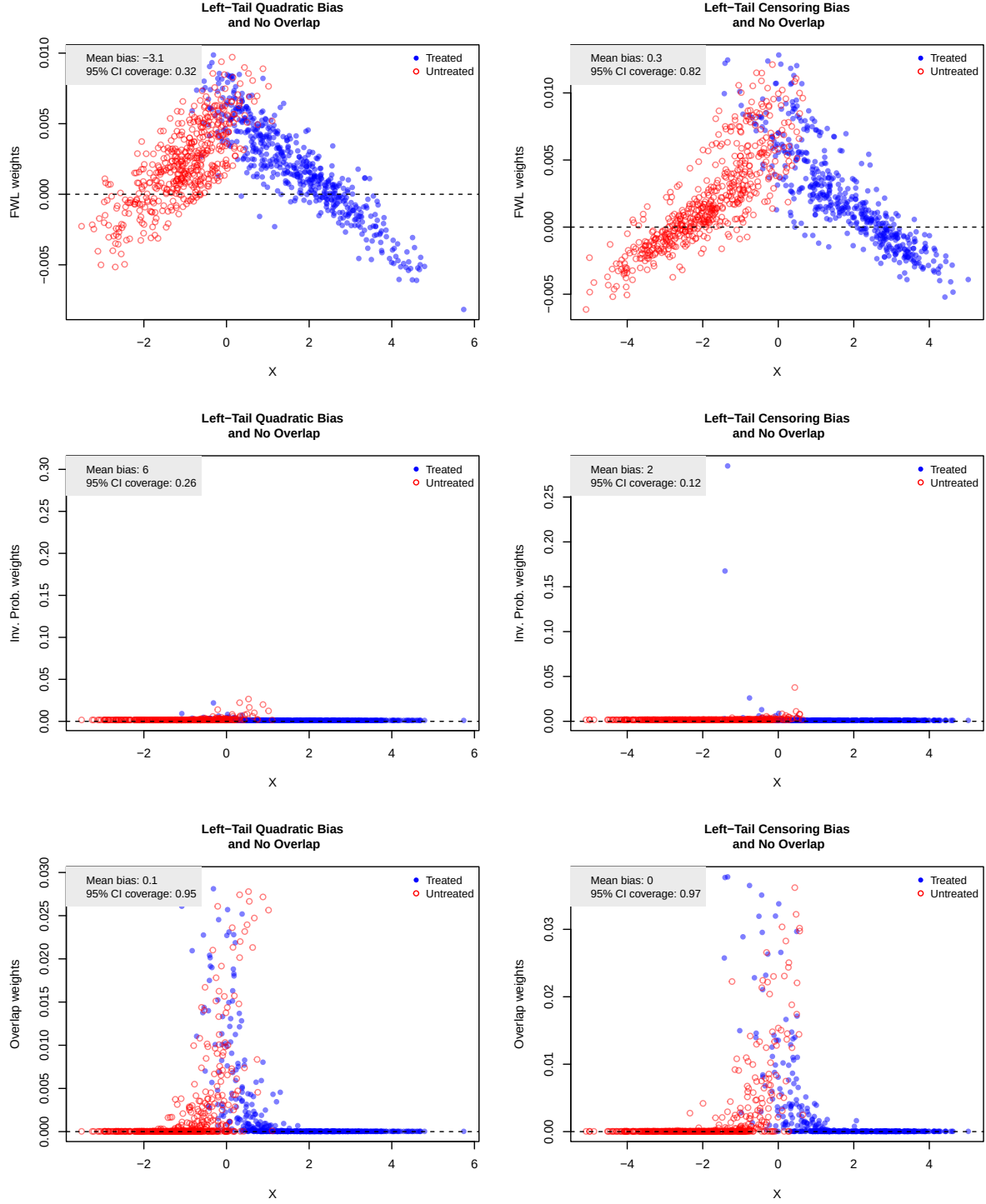


Figure 3: Comparison of weight distribution, bias, and 95% confidence interval coverage under three routines: regression, IPW, OW

weighting (bottom-left panel of Figure 3) produces the estimate closest to the true τ of 3 and achieves coverage of 0.95.

Under left-censoring, the regression’s bias is much less severe (top-right of Figure 3), but the coverage is poor. Inverse probability weighting (center-right of Figure 3) worsens bias and also coverage because it puts too much weight on few observations. Overlap weighting (bottom-right of Figure 3) outperforms both the regression and inverse probability weighting.

3.5 Mitigating and aggravating factors

Bias and coverage problems that arise under overlap violations could be mitigated by adding more covariates to increase the region of common support. But adding irrelevant covariates inflates standard errors. Removing a relevant covariate, while it may reduce the number of negative weights, can worsen other types of biases, such as selection or omitted variable bias, and would compromise the integrity of the causal identification exercise. If no common support can be found on the covariate space between treated and untreated sets, reporting the effect of negative weights is the best that can be done.

The negative weights themselves do not bias ATE s. For bias to arise they need to be associated with outcomes values that are large enough. Consider the difference in bias between the left-tail negative quadratic bias and the left-tail censoring in Figure 2. The distribution of negative weights under both data generating processes is very similar. However, the bias is worse with the left-tail quadratic bias because the negative weights are associated with larger negative outcomes. The smaller the outcomes associated with the negative weights, the smaller the bias. Some transformations of the outcome variable can therefore reduce bias. If some extremities of the covariate space has extreme observations, then winsorizing them can reduce bias if it has little effect on the sample sizes being considered. For non-zero outcomes, a log-transformed outcome can produce less bias under overlap violations. A log-transformation may also worsen bias if the transformation acts on outcomes between 0 and 1, and large negative tails are produced, as in top-left of Figure 1.

Another aggravating factor is the worsening of the lack of overlap itself.

Figure 4 shows the bias and coverage of confidence intervals under varying overlap conditions for regression $\hat{ATE}$ and overlap-weighted $\hat{ATE}^{IPW}$ under the same left-tailed quadratic simulation as in top-left of Figure 1. Although overlap weighting manages the bias better than regression weighting, bias is worsening as overlap is violated or the region of common support shrinks. Similarly, for the 95% confidence interval coverage.

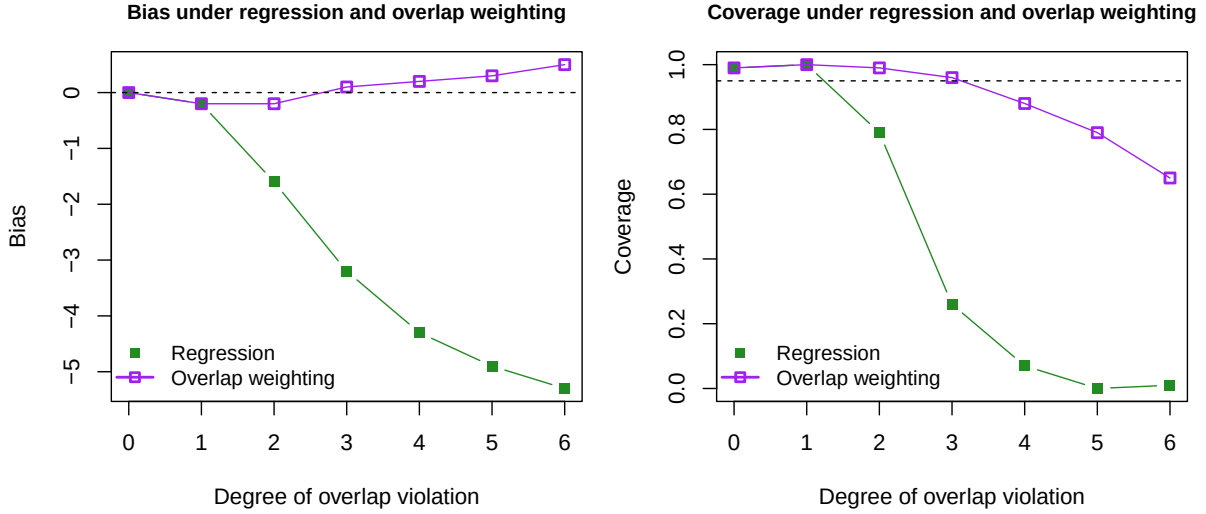


Figure 4: The bias and 95% confidence interval coverage under worsening overlap conditions of the left-tailed quadratic biased simulation

4 Applications

I apply the detect-and-correct approach described above to five studies that use panel data, has a binary or binarizable treatment, and is a non-staggered treatment setting, i.e. treatment status reverts. The studies are part of broader electoral accountability literature where the treatment is rare but damaging such as a natural disaster, mass shootings, or arson. Settings such as those are ripe for overlap violations because the type of voting districts such events occur are different to the type of districts where such events do not.

Studies try to control for these differences using covariates that might capture these differences, and uses fixed effects for unobservable differences. But as has been demonstrated with simulated data up to now, a lack of overlap can introduce bias as the regression extrapolates its internal propensity score beyond $[0, 1]$. I reanalyze the treatment effects recovered by the studies main specifications. In all but Markarian and Newman (2025) which uses a lagged dependent variable specification, the main specification is a two-way fixed effects specification with covariates.

For each relevant treatment effect, a three panel is used to summarize the detect-and-correct approach. In the left panel the FWL weights $w_i^{(1)}$ are plotted against the overlap weights $w_i^{OW,(1)}$. In addition, on its top-left corner the panel reports the number of treated units (size of T), the ‘Sum treated’ of outcomes Y^T , the ‘Sum treated (OW)’ of outcomes when weighted by $w_i^{OW,(1)}$,

the number of negative FWL weights among the treated units (size of T^-), and the outcome sum attributable to the negative weights Y^{T^-} . The center panel plots $w_i^{(0)}$ against the overlap weights $w_i^{OW,(0)}$ along with the corresponding quantities for the untreated units including Y^C and Y^{C^-} . The third panel reports the two estimates: regression and overlap weighting.

In reporting the estimates in the right panel, unless otherwise indicated, the point estimates labeled ‘Regression’ are those that were in the original paper. The only exceptions here are in Figure B.13 and the estimates in Markarian and Newman (2025) for which the reported estimates did not use population weights in the regression specification. The standard error, however, differs from the original papers as they are obtained by Eq. 12 than the more standard method for regression. This was so that the standard error calculation between regression and overlap and inverse probability weighting was comparable. The effect of this is that the standard errors are somewhat wider than those reported in the original papers and a few estimates the original papers report as significant are now insignificant in the right panel. However, this only affects estimates that are small.

The studies are discussed in order of publication date.

4.1 Gasper and Reeves (2011)

The first application is to the results in Table 1 and Table 2 of Gasper and Reeves (2011). Table 1 considers the effect of natural disasters on the incumbent governor’s vote share. Table 2 considers the effect of natural disasters on the incumbent president’s vote share.

There are two relevant treatment variables in Table 1: the number of disaster declarations by an incumbent governor for the county (‘Disaster Declarations’ Model 2 of Table 1), and the number of denials by the president for disaster declarations (‘Turn Downs’ Model 3 of Table 1). Their treatment effect coefficients were both positive and statistically significant at 95%. Both treatment variables are discrete⁷. To form clear treated and untreated sets, both were binarized with a cutoff at 0. The binarization did not change the sign or the significance of the estimates. Before binarization the coefficient on ‘Disaster Declarations’ in Model 1 was 3.53 with standard error 0.273. After binarization the coefficient was 2.684 with standard error 0.363. For ‘Turn

⁷‘Disaster Declarations’ has values 0 to 4 with 14,032 zeros, 1,219 ones, 286 twos, 38 threes, and 5 fours. And ‘Turn Downs’ has values 0 to 2 with 12,731 zeros, 2,559 ones, and 290 twos.

Downs’ in Model 2, the coefficient before binarization was 2.65 (s.e. 0.244), and after binarization was 3.32 (s.e. 0.281).

Table 2’s relevant variables are the similar. They are the number of disaster declarations for the county by the incumbent president (‘Disaster Declarations’ Model 2 of Table 2) and the number of requests turned down by the president (‘Turn Downs’ Model 3 of Table 2). ‘Disaster Declarations’ had a positive treatment estimate and ‘Turn Downs’ had a negative estimate. Both estimates were statistically significant at 95% level. These were also discrete, and were binarized with a cutoff at zero. The binarized coefficient on ‘Disaster Declarations’ was somewhat higher going from 0.52 in the non-binarized case to 0.82 when binarized. The coefficient on ‘Turn Downs’ became weaker going from -0.97 in the non-binarized case to -0.79 in the binarized case. The binarization preserved the significance.

The covariates used for Model 2 of Table 1 were logged dollars of weather damage per 10,000 citizens in the county, lagged presidential vote share, lagged governor vote share, and median income. For Model 3 of Table 1 those same covariates plus the original ‘Disaster Declarations’ variable was used. All covariates in all models were statistically significant at conventional levels. Covariates for Model 2 of Table 2 were similar but instead of lagged governor vote share, the president’s lagged vote share was used. In Model 3 of Table 2, the non-binarized disaster declarations of the president is a covariate.

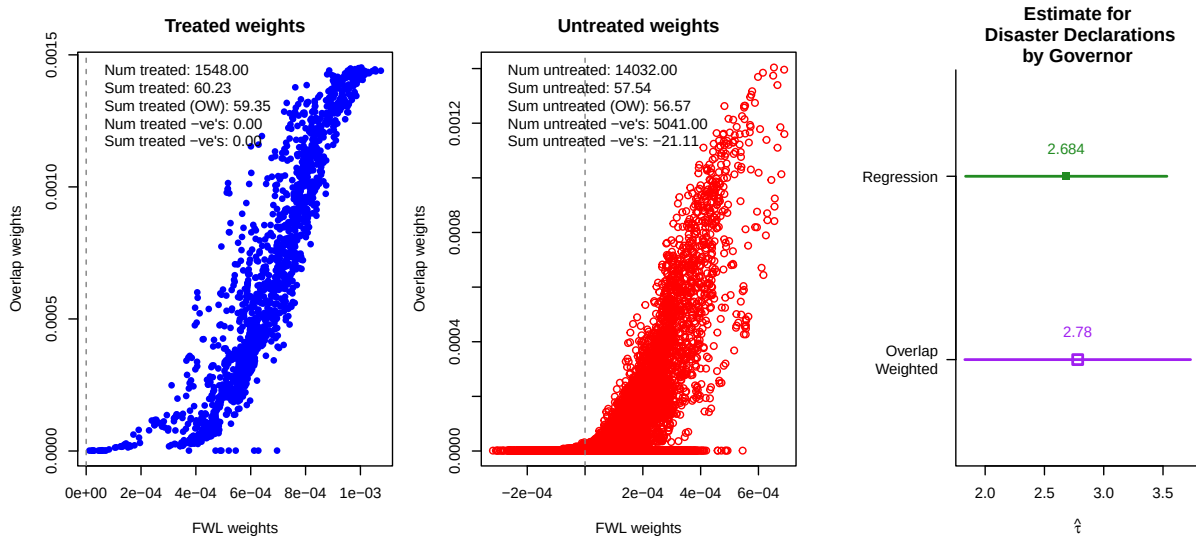


Figure 5: The FWL weights and overlap weights for the binarized treatment coefficient for governor’s ‘Disaster Declarations’ of Model 2 Table 1 in Gasper and Reeves (2011)

Figure 5 shows the FWL weights $w_i^{(0)}$ and $w_i^{(1)}$ from the regression with binarized ‘Disaster Declarations’ as treatment.

No treated unit has a negative FWL weight, but over a third of the untreated units (5041 out of 14032) are negative. These account for $Y^{C-} = -21$ of the untreated sum $Y^C = 57.54$. Overlap weighting sets these negative weights to be zero or positive and reweights those in the common support region. The overlap weighted untreated sum is a full outcome unit lower. Overlap weighting also more aggressively upweights the units in the common support region among the treated units causing the overlap weighted treated sum to also fall by about an outcome unit. So, the resulting estimates for governor’s Disaster Declarations do not change much (right panel of Figure 5).

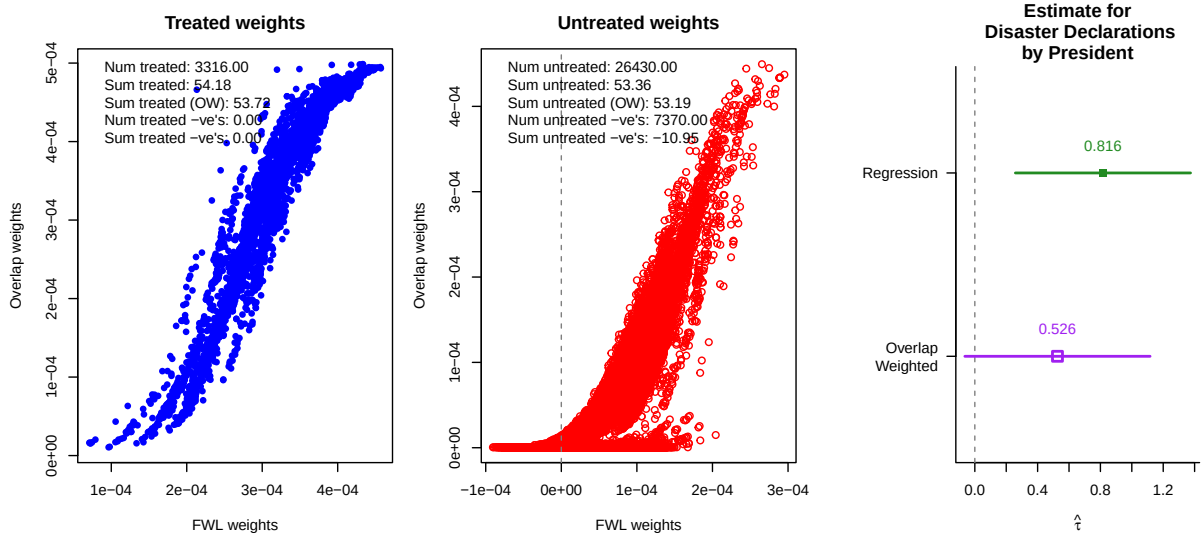


Figure 6: The FWL weights and overlap weights for the binarized treatment coefficient for president’s ‘Disaster Declarations’ of Model 2 Table 2 in Gasper and Reeves (2011)

In contrast, for the estimate associated with president’s Disaster Declarations, overlap weighting attenuates by about 0.3 outcome units making the estimate insignificant at 95% level, but still significant at 90% (right panel of Figure 6). The left and center panels of that same figure shows that treated units have no FWL negative weights, but about 30% of the untreated units has negative FWL weights and are associated with a Y^{C-} of -10.95.

The corresponding plot for ‘Turn Downs’ in Model 3 of Table 1 is reported in the Appendix as Figure B.1 and the ‘Turn Downs’ in Model 3 of Table 2 is reported in the Appendix as Figure B.2. The untreated sets of these models also had a substantial number of negative weights. But overlap weighting did not result in much change in the component sums or the resulting estimates, only

slightly weakening them.

In this application, despite overlap violations in the covariate space, regression and overlap weighted estimates agreed in sign. The next application illustrates that negative weights can revert the sign of the treatment estimate.

4.2 Hazlett and Mildenerger (2020)

The second application is from Hazlett and Mildenerger (2020). The outcome Y here is the precinct-block level vote share for pro-environmental propositions (ballot measures) in California. The treatment is the proximity to large fires. The study's main results includes a two-way fixed effects regression (precinct-block and election year) with a treatment indicator for whether a fire was experienced in the precinct with three covariates: the lagged Democratic vote share, average precipitation in the election cycle, and precipitation deviation from a historical average. The results of this regression is that exposure to fire is associated with a 6 percentage point increase in voting for pro-environmental propositions on the ballot. This is statistically significant at 95% level.

Another regression tests the effect of proximity with an expansive set of indicators for distance, imposing the precinct 35-40km from a fire to be the reference category. The results of this regression (in Figure 1 of the original paper) imply that closer to a fire more likely the precincts are to vote for pro-environmental propositions.

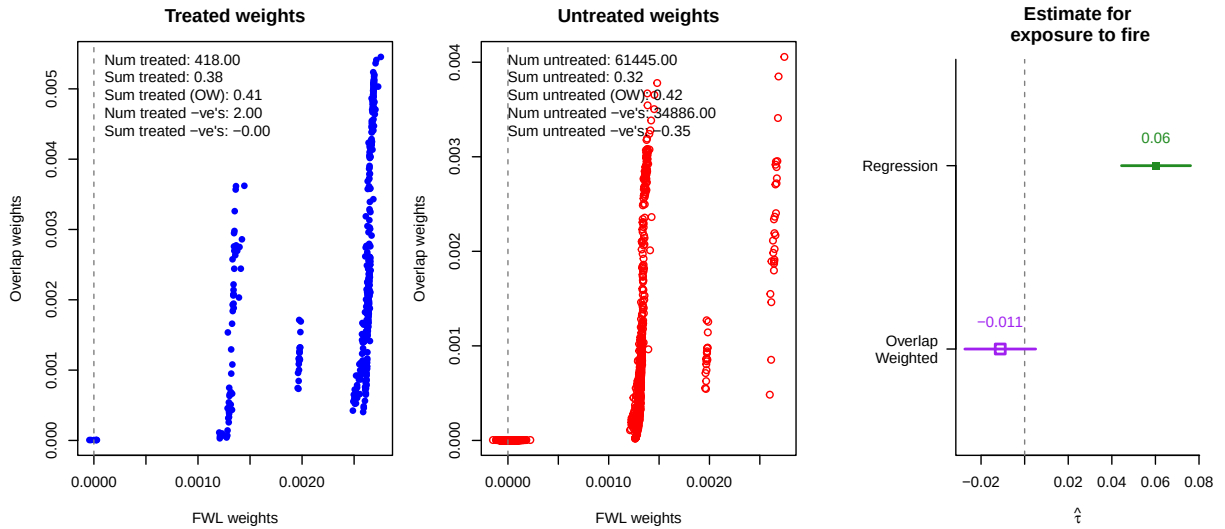


Figure 7: The FWL weights and overlap weights for fire exposure in the precinct from Hazlett and Mildenerger (2020)

Figure 7 shows that the regression treatment effect of 0.06 is an artefact of overlap violations. The center panel reports that the majority of the observations are associated with negative weights and they account for -0.35, overwhelming the Y^{T+} of 0.32. Under overlap weighting the sum from untreated units increases to 0.42. The sum from treated units increases to 0.41. In the case of the treated units, the change in sum is not due to suppression of negative weighted outcomes since there are too few (left panel), but from the overlap weighting scheme favoring the common support region more than the regression's FWL weights. The resulting overlap-weighted estimate is a different sign to the study's initial estimate of 0.06.

The overlap violation also affects the regression with distance indicators in Figure 1 of Hazlett and Mildenerger (2020).

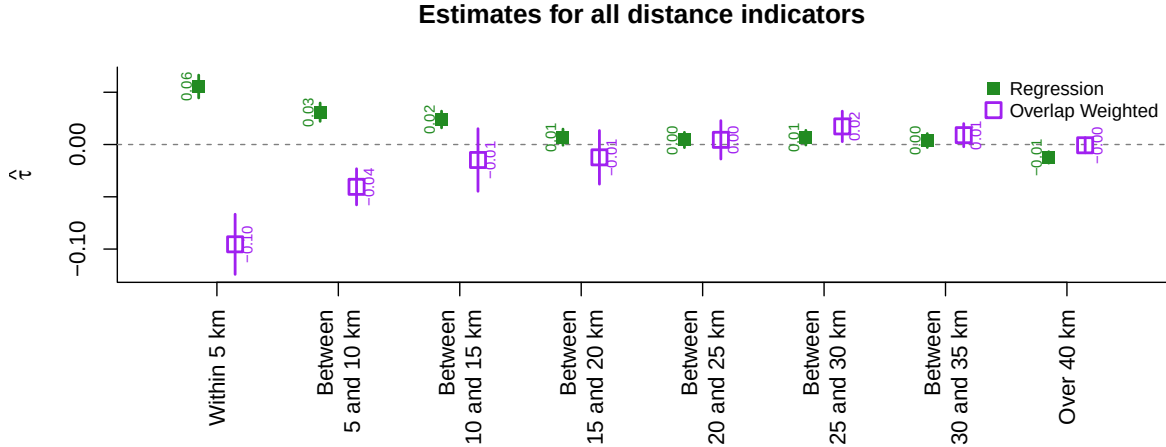


Figure 8: Regression vs Overlap estimates for Figure 1 of Hazlett and Mildenerger (2020)

The regression estimates for the first four distance categories (within 5 km, between 5 and 10 km, between 10 and 15 km, and between 15 and 20 km) are all a different sign to the overlap-weighted counterpart for those same categories. Figure 8 demonstrates all the categories with confidence intervals obtained using formula 12. Appendix Figures B.3 to B.10 demonstrates how a large number of negative weights can cause regression estimates to reverse in sign. For instance, in center panel of Appendix Figure B.3 more than half of the weights associated with controls are negative and they account for a sum of outcomes close to -0.85.

Even if the sum Y^{C-} is large compared to Y^{C+} , however, sign reversal does not always occur.

For the category of fire exposure within 20 to 25km (Appendix Figure B.7) the sum Y^{C^-} is -0.55 and Y^{C^+} is 0.49. There are no negative effects in the treatment set. And regression and overlap weighting estimates the same treatment estimate. In some cases, overlap weighting can strengthen the estimate, such as with the category of exposure between 25 and 30 km.

4.3 Garcia-Montoya, Arjona, Lacombe (2022)

The third study is García-Montoya et al. (2022). Here the objective of study is the effect of shootings in a county on county-level turnout and presidential vote share of the Democratic candidate. The study considers the effect of any shooting, then fatal and nonfatal shootings in turn, and then tests these with four specifications each. I focus only on the study's strongest result, which is the effect of shootings on Democratic presidential vote share. The relevant specification is the two-way fixed effects with controls plotted as the second row of each outcome in Figure 2 of that paper⁸. The covariates used are the total population, proportion of population that is non-white, and the change in unemployment rate. All treatments and covariates are statistically significant at 95% for Democratic vote share⁹.

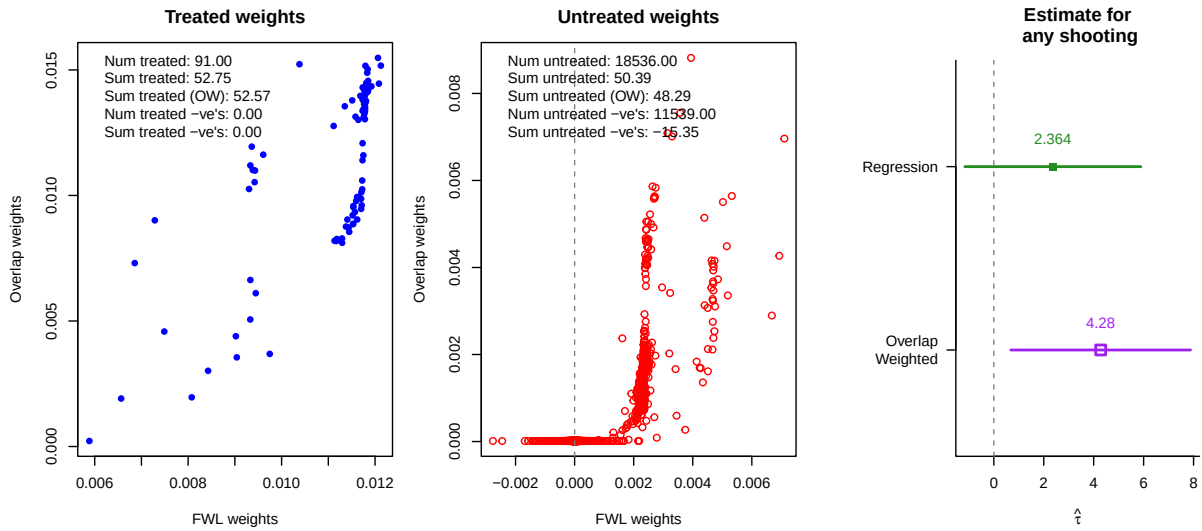


Figure 9: The FWL weights and overlap weights for any shooting in the county from row 2 of Figure 2 of Garcia-Montoya, Arjona, Lacombe (2022)

Figure 9 shows the FWL and overlap weights of the treated units in the left panel, and those

⁸Two of the other specifications are two-way fixed effects without covariates, which is not relevant to this paper's

⁹None of the specifications find that electoral turnout changes meaningfully

same weights for the untreated units in the center panel. The right panel shows the treatment effects estimated with regression and overlap-weighting. There are a sizeable number (62%) of negative weights in the untreated set, and they are also associated with a non-trivial Y^{C-} .

In this case, overlap weighting strengthens the treatment effect, by removing the effect of the outcomes associate with the negative weights. The counterpart to $Y^C = 50.39$ from the overlap weighting is 48.28 and the counterpart to $Y^T = 52.75$ from the overlap weighting is 52.57. Appendix Figures B.11 and B.12 demonstrate the weights and new estimates for the subset of fatal and non-fatal shootings, respectively. In both models there are a substantial number of negative weights and they contribute to a sizeable Y^{C-} component. Reweighting these negative weights to be overlap-weighted reduces the control sum in Figure B.11 reduces it from 49.795 to 46.135 with little effect on the treated sum, which increases the estimated effect from 1.78 to 5. Similarly, in Figure B.12, the overlap-weighted control sum is reduced to 47.5 by the reweighting and increases the estimate from 2.93 to 7.01.

4.4 Marsh (2023)

The fourth application study is Marsh (2023). This study tests the effect of church arsons, mass shootings, and natural disasters on county level turnout during presidential election cycles. The treatment effects are calculated using a two-way fixed effects at county and election year with covariates percentage of blacks, total population, and income.

In the original paper, the estimate for church arson is -0.013 and is statistically significant at 95% level ('Arson' column in Table 1). Figure 10 shows the relevant FWL and overlap weights for this estimate. There are a sizeable number of negative weights (72% of all control weights) in the regression's control sum and is overwhelmingly large. Removing the effect of the negative weights using overlap weighting, increases the control sum from 0.5 to 0.51. This reduces the estimate from -0.013 to -0.027 further strengthening it.

The treatment effect for mass shootings nor natural disasters in Table 1 were not significant in the original paper and overlap weighting corrections calculated effects quite similar to these. In the case of mass shootings, treatment effects were recalculated by dropping fatalities and injuries from the covariates, since these are almost perfectly correlated with the treatment indicator. The resulting treatment effect is very small, and even with many negative weights in its control sample,

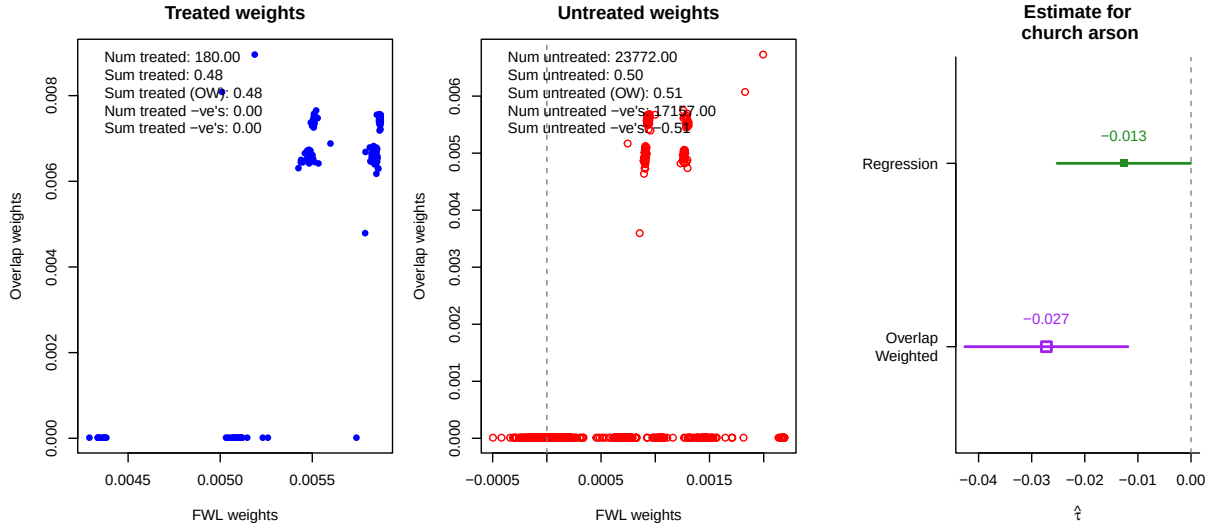


Figure 10: The FWL weights and overlap weights for effect of black church arsons in Table 1 of Marsh (2023)

reweighting with overlap weighting made little difference to the estimate. The plots for both these estimates are in the Appendix as Figure B.13 and Figure B.14.

4.5 Markarian and Newman (2025)

The final application is from Markarian and Newman (2025). Here the treatment effect of interest is the proximity to a single mass shooting (Cascade Mall) on the preferences for gun control as revealed by a ballot measure is studied. Proximity is indicated by 5 mutually exclusive dummy variables for the categories ‘5 miles or less’, ‘between 5 and 10 miles’, ‘between 10 and 15 miles’, ‘between 15 and 20 miles’, and ‘between 20 and 25 miles’. I focus on the results reported in Figure 1 Panel d of the original paper. Unlike the models in the previous application, no fixed effects are involved here, and the specification is instead a lagged dependent variable model. An exhaustive set of covariates are used including pro-gun control vote share in a previous proposition (lagged dependent variable), vote share for Obama, percentage of males, blacks, latinos, over-55 residents, and so on. The estimates associated with each of the categories in Figure 1 Panel d of the original paper are positive and significant with the largest effect for those closest to the Cascade Mall which dissipates as distance increases.

The Figure 11 shows the weights and the new overlap estimates for the indicator of being within 5 miles of the shooting. Negative weights make up 23% of the control sample’s weights contributing

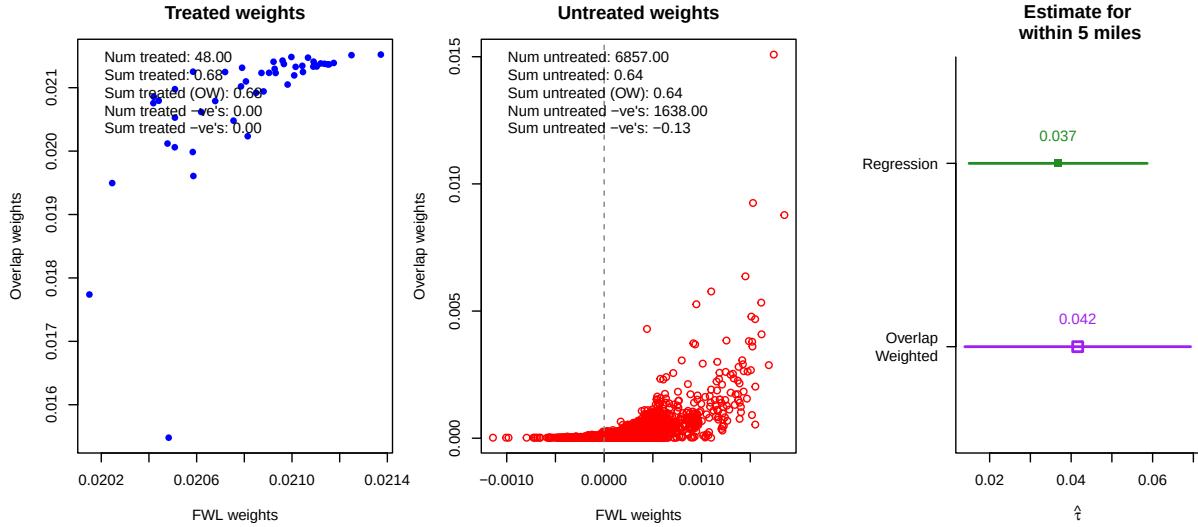


Figure 11: The FWL weights and overlap weights for effect of being within 5 miles of a mass shooting on voting for gun control from Figure 1d of Markarian and Newman (2025)

-0.13 to the sum $Y^C = 0.64$. Reweighting slightly improves the estimate from 0.037 to 0.042 since the overlap weighting reduces the control sum to 0.637 from 0.643. These differences are quite small.

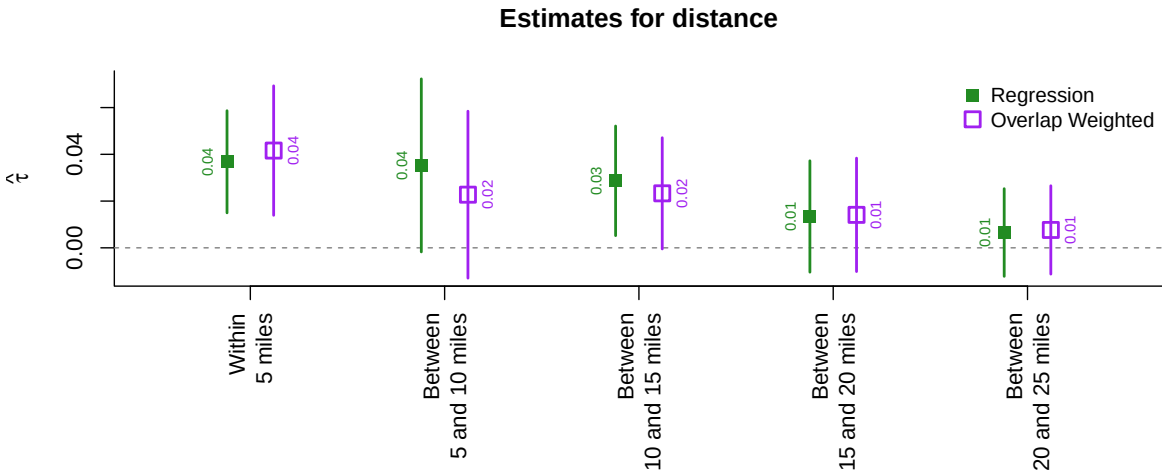


Figure 12: Estimates of all distance indicators on the effect of a mass shooting on voting for gun control from Figure 1d of Markarian and Newman (2025)

Figure 12 plots the estimates for the distance categories in Figure 1d. Overlap weighting makes little difference to the sign of the estimate. The overlap-weighted estimate for ‘between 5 and 10 miles’ is twice as weak as the regression estimate, which affects its significance.

5 Conclusion

Regression intending to recover treatment effects using covariate adjustment and sometimes fixed effects in observational empirical settings is an established method in the social sciences. Compared to many other causal methods, regressions are easily replicable, transparent, and implementable with little computational needs. However, competing methods such as synthetic controls and second-generation differences-in-differences estimators such as Callaway and Sant’Anna (2021) and De Chaisemartin and d’Haultfoeuille (2023) are now gaining popularity. But researchers that use regressions and these methods in tandem usually find that the estimates obtained from such a method is different to those obtained from regression. Some of these differences are driven by how the algorithm treats always-treated units, and previously treated units. But another difference is in how the algorithms adjusts covariates to gain covariate balance between treated and untreated sets. This paper is specifically about providing a diagnostic aid for when the covariate adjustment is the main driver of the discrepancy.

Synthetic controls do covariate adjustment in a convex optimization routine where all weights of the untreated candidates must sum to 1 and each must be positive. And if the treatment unit’s pre-treatment covariates exists outside of the convex hull of the pre-treatment covariates of the untreated candidates, users do not proceed with the estimation, or if they do so done with warnings. The Callaway and Sant’Anna (2021) estimator does covariate adjustment using a combination of inverse probability weighting and outcome regression, and if the propensity score is too close to the extremes, those observations are trimmed from the comparison.

Covariate adjustment in regressions and these two methods just discussed differ in two important ways. In those methods, weights used to reweight the treated and untreated sets are positive and if any of the units are outside of the common support region they are removed. In regressions, the weights do not have to be positive, and units with extreme propensity scores, as estimated internally using a linear probability model, are still retained in the comparison. These differences in covariate adjustment lead to vastly different results in situations where overlap is poor.

When these negative weights are combined with outcomes that are sufficiently large, the recovered treatment effects will be biased and confidence intervals will have poor coverage. In this paper, I demonstrate the conditions for the occurrence of such bias with simulated and real data. The bias

can manifest in several ways. It may change the sign completely. Or the bias can underestimate the effect, leaving what may have been statistically significant effect seem insignificant. In either case, the conclusions of the study are affected.

For practitioners, I recommend a detect-and-correct approach. The detection uses a decomposition of the regression estimate into treated and untreated components and then those components to positive and negative components within their respective sums. When the negative components are sizeable, bias is possible. To correct, an overlap weighting routine sets the negative weights to zero and reweights other units that have common support so that they add to 1. Other weighting methods that restrict weights to be positive are also valid. Comparisons in simulated data using both inverse probability and overlap weighting showed that overlap weighting was more robust as the common support region got sparser.

This paper considers the algorithmic differences in covariate weighting given the covariates chosen by the user. If a user excludes covariates, then negative weights due to overlap violation is not a concern. Of course, in most practical applications, users do not have access to all possible covariates. But omitting a covariate just to avoid the issue of negative weights is not recommended, as this can introduce omitted variable bias into the estimate. If the included covariates are highly correlated with treatment status leading to many negative weights, this might be important information to judge the integrity of the causal exercise.

The diagnostic tools I discuss here are also intended to help users decide whether regressions are the most appropriate specification. As simulations have demonstrated, overlap weighting corrections also become biased as covariate overlap worsens. The best approach in these situations is to calculate average treatment effects on the treated units that have sufficient overlap such as a synthetic control on multiple treated units. While panel data methods like synthetic control or Callaway and Sant’Anna (2021) cannot be applied directly when treatment status can reverse as in many of the applications here, it would improve the causal integrity of the exercise to adapt the setting to one of treatment non-reversal, rather than to persist within a regression framework.

References

- Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of economic literature*, 59(2), 391–425.
- Callaway, B., & Sant’Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of econometrics*, 225(2), 200–230.
- Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6), 656–664.
- Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1), 187–199.
- De Chaisemartin, C., & d’Haultfoeuille, X. (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The econometrics journal*, 26(3), C1–C30.
- Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, 387–401.
- García-Montoya, L., Arjona, A., & Lacombe, M. (2022). Violence and voting in the united states: How school shootings affect elections. *American Political Science Review*, 116(3), 807–826.
- Gasper, J. T., & Reeves, A. (2011). Make it rain? retrospection and the attentive electorate in the context of natural disasters. *American journal of political science*, 55(2), 340–355.
- Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1), 201–209.
- Hazlett, C., & Mildenberger, M. (2020). Wildfire exposure increases pro-environment voting within democratic but not republican areas. *American Political Science Review*, 114(4), 1359–1365.
- Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1), 1–24.
- Imai, K., & Kim, I. S. (2021). On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis*, 29(3), 405–415.
- Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521), 390–400.
- Li, F., Thomas, L. E., & Li, F. (2019). Addressing extreme propensity scores via the overlap weights. *American journal of epidemiology*, 188(1), 250–257.

- Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304), 993–1010.
- Lovell, M. C. (2008). A simple proof of the fwl theorem. *The Journal of Economic Education*, 39(1), 88–91.
- Markarian, G. A., & Newman, B. J. (2025). Exposure to mass shootings and voting directly on gun policy. *American Political Science Review*, 119(3), 1565–1572.
- Marsh, W. Z. (2023). Trauma and turnout: The political consequences of traumatic events. *American political science review*, 117(3), 1036–1052.
- Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12), 1393–1512.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

Appendix A

A.1 Statement of FWL theorem

For a linear regression of the form

$$y = X_1\beta_1 + X_2\beta_2 + u$$

where X_1 and X_2 are blocks of regressors the OLS estimate of β_2 is identical to the estimate obtained by the following three step procedure.

1. Regress y on X_1 and saving the residuals as $\tilde{y}$.
2. Regress each column of X_2 on X_1 and saving residuals as $\tilde{X}_2$.
3. Regress $\tilde{y}$ on $\tilde{X}_2$ to obtain

$$\hat{\beta}_2 = (\tilde{X}_2^\top \tilde{X}_2)^{-1}(\tilde{X}_2^\top \tilde{y})$$

For a proof see Lovell (2008).

A.2 Corollary on weights summing to 0

Suppose that X_2 is a single regressor which we rename x_2 and X_1 contains an intercept¹⁰, then the coefficient of $\hat{\beta}_2$ can be written as a weighted sum of the outcomes Y_i .

$$\hat{\beta}_2 = \sum_i w_i Y_i$$

where the weights sum to zero.

A proof sketch of this is as follows.

Let $M_{X_1} = I - X_1(X_1^\top X_1)^{-1}X_1^\top$ and

$$\tilde{Y} = M_{X_1} Y$$

$$\tilde{x}_2 = M_{X_1} x_2$$

From the FWL theorem,

$$\hat{\beta}_2 = \frac{\tilde{x}_2^\top \tilde{Y}}{\tilde{x}_2^\top \tilde{x}_2}$$

where the weights can be written as

$$w = \frac{\tilde{x}_2^\top}{\tilde{x}_2^\top \tilde{x}_2} M_{X_1}$$

or

$$w^\top = \frac{x_2 M_{X_1}}{x_2^\top M_{X_1} x_2}$$

Since the intercept is in X_1 , $M_{X_1} \mathbf{1} = 0$ where $\mathbf{1}$ is a vector of ones. So,

¹⁰The block of regressors X_1 can also contain fixed effects that are perfectly collinear with the intercept. Since such matrix of fixed effects can always be transformed in such a way to include an intercept, this corollary applies to those settings

$$\begin{aligned}
\sum_i w_i &= w^\top \mathbf{1} \\
&= \frac{x_2 M_{X_1} \mathbf{1}}{x_2^\top M_{X_1} x_2} \\
&= 0
\end{aligned}$$

A.3 Corollary on a binary indicator's weights on 1 summing to 1

Now suppose that x_2 is either 0 or 1. Then the weights associated with all $\{i : x_{2i} = 1\}$ sum to 1 and the weights associated with all $\{i : x_{2i} = 0\}$ sum to -1 .

A proof sketch is as follows:

Since

$$w = \frac{M_{X_1} x_2}{x_2^\top M_{X_1} x_2}$$

and from the previous corollary weights sum to 0 and that the weights multiplied by x_2 sum to one.

$$x_2^\top w = \frac{x_2^\top M_{X_1} x_2}{x_2^\top M_{X_1} x_2}$$

In summation

$$\begin{aligned}
\sum_i w_i &= 0 \\
\sum_i x_{2i} w_i &= 1
\end{aligned}$$

But because x_{2i} is binary

$$\sum_i x_{2i} w_i = \sum_{i: x_{2i}=1} w_i = 1$$

Using the fact that weights sum to 1, the sum over $\{i : x_{2i} = 0\}$ must sum to -1.

Appendix B

B.1 Additional results on Gasper and Reeves (2011)

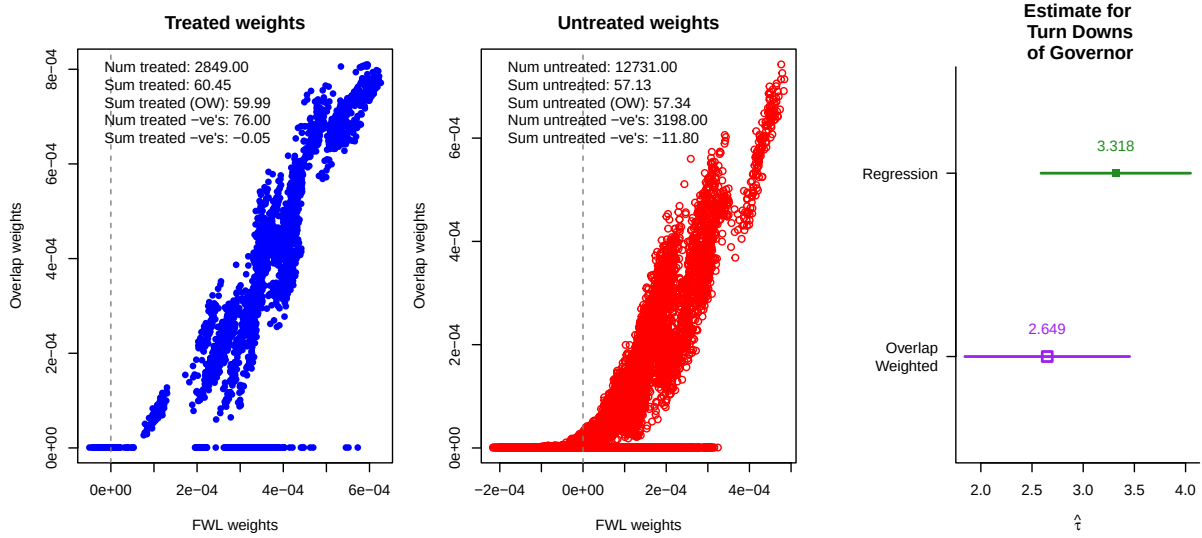


Figure B.1: The FWL weights and overlap weights for the binarized treatment coefficient of 'Turn Downs' of the Governor's request by President in Model 3 Table 1 in Gasper and Reeves (2011)

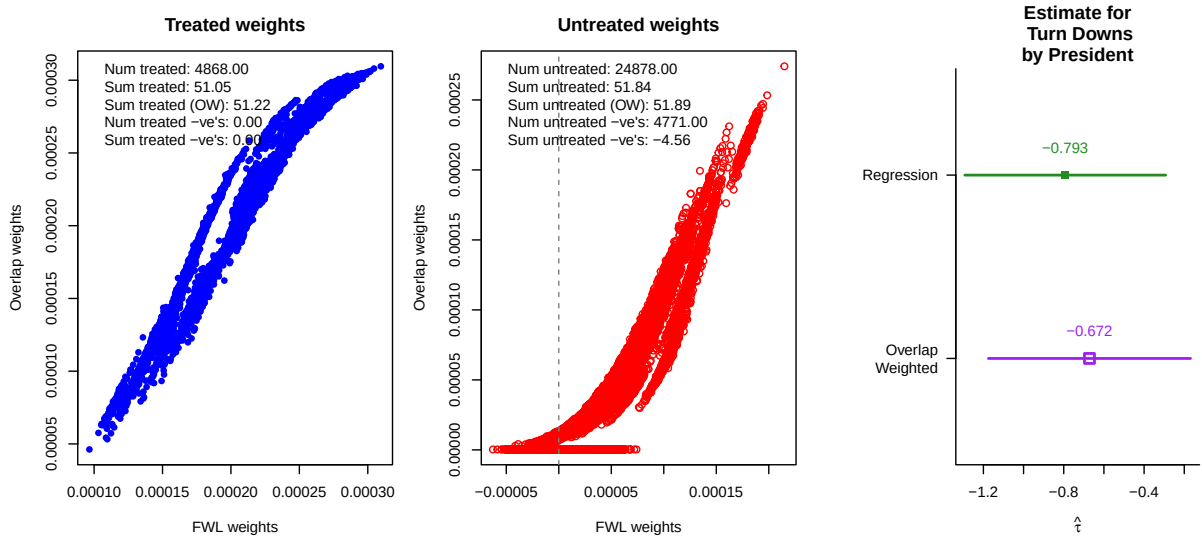


Figure B.2: The FWL weights and overlap weights for the binarized treatment coefficient of 'Turn Downs' of the Governor's request by President in Model 3 Table 1 in Gasper and Reeves (2011)

B.2 Additional results on Hazlett and Mildenerger (2020)

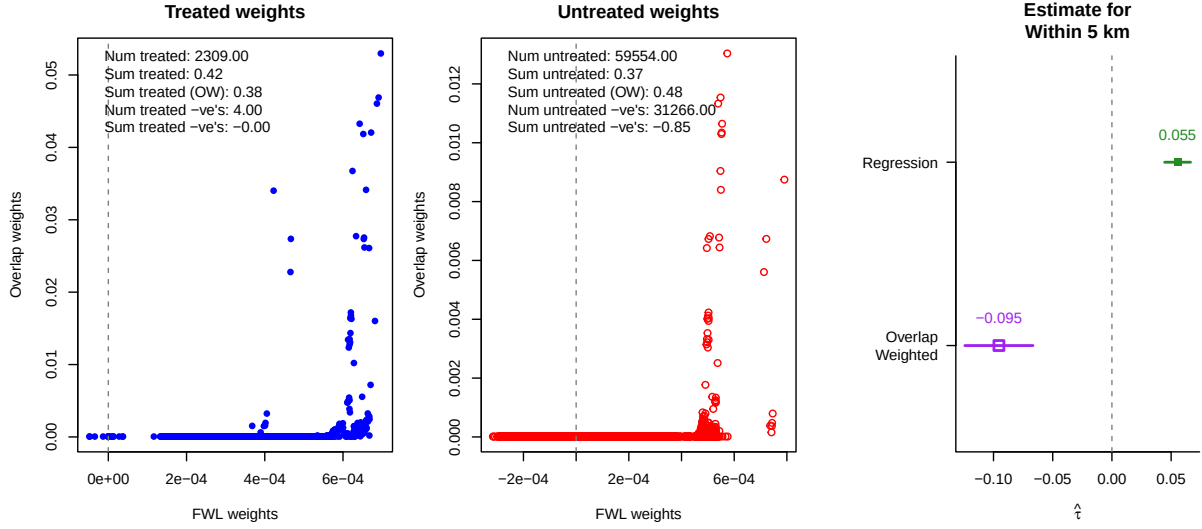


Figure B.3: The FWL weights and overlap weights for the distance category of within 5km in Figure 1 of Hazlett and Mildenerger (2020)

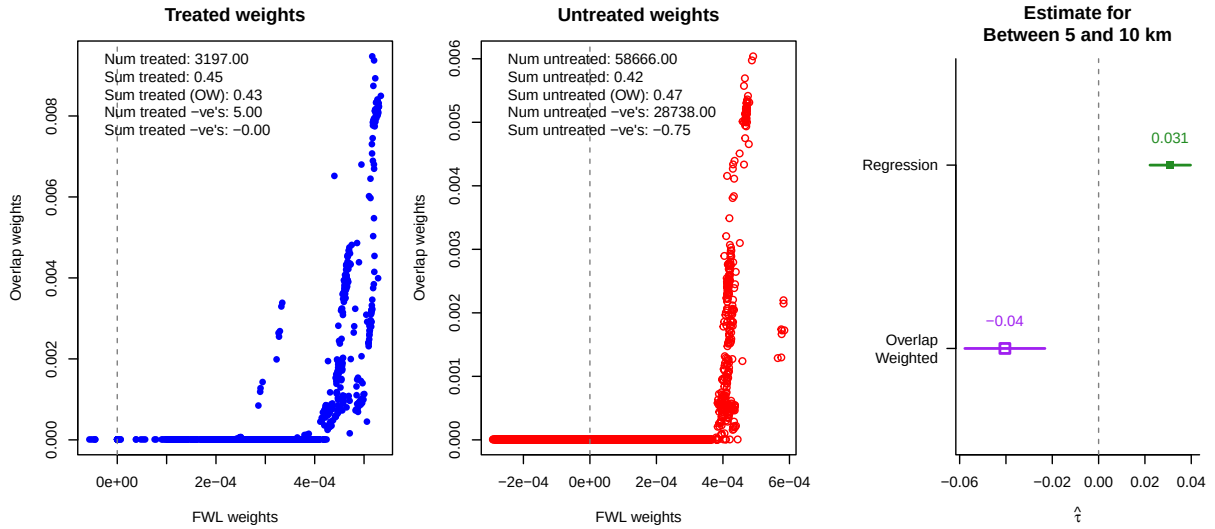


Figure B.4: The FWL weights and overlap weights for the distance category of 5 to 10km in Figure 1 of Hazlett and Mildenerger (2020)

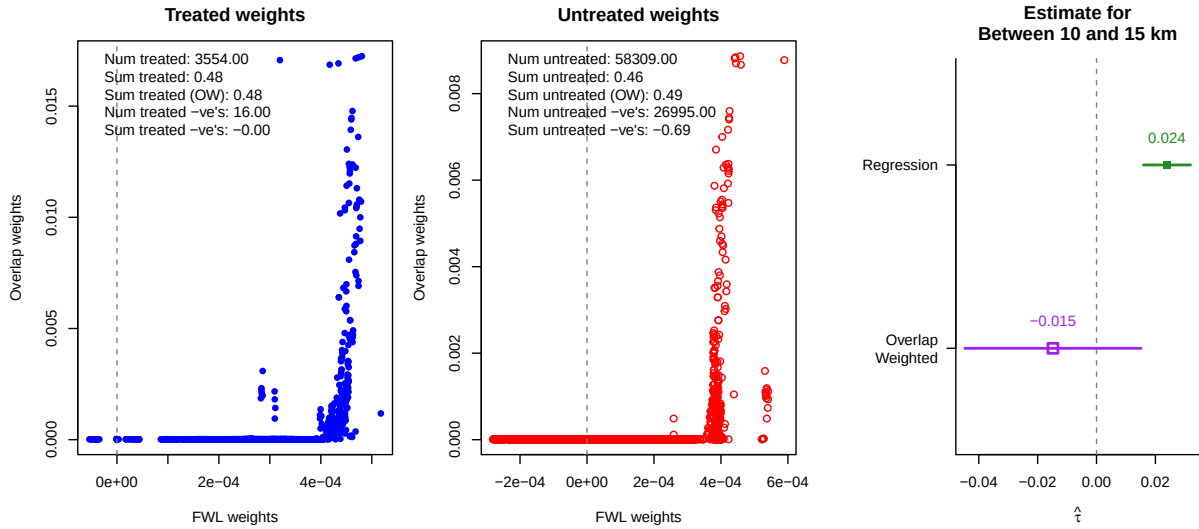


Figure B.5: The FWL weights and overlap weights for the distance category of 10 to 15km in Figure 1 of Hazlett and Mildenerger (2020)

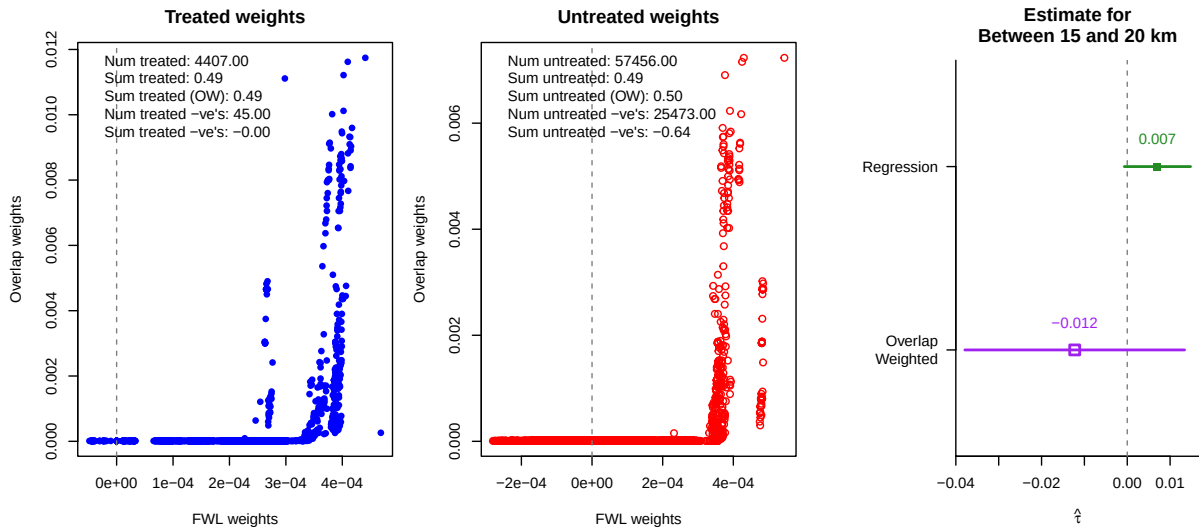


Figure B.6: The FWL weights and overlap weights for the distance category of 15 to 20km in Figure 1 of Hazlett and Mildenerger (2020)

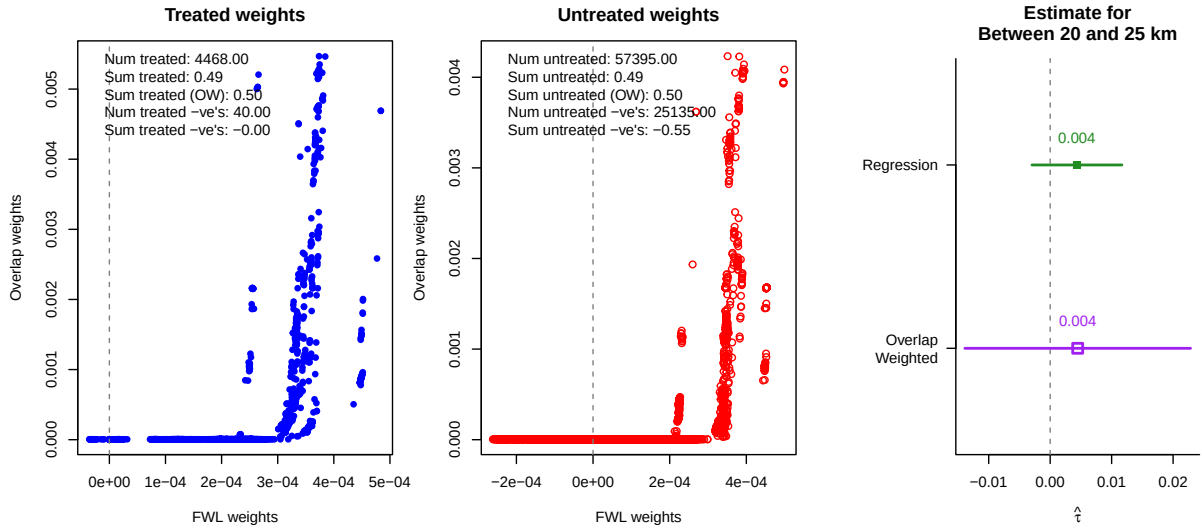


Figure B.7: The FWL weights and overlap weights for the distance category of 20 to 25km in Figure 1 of Hazlett and Mildenberger (2020)

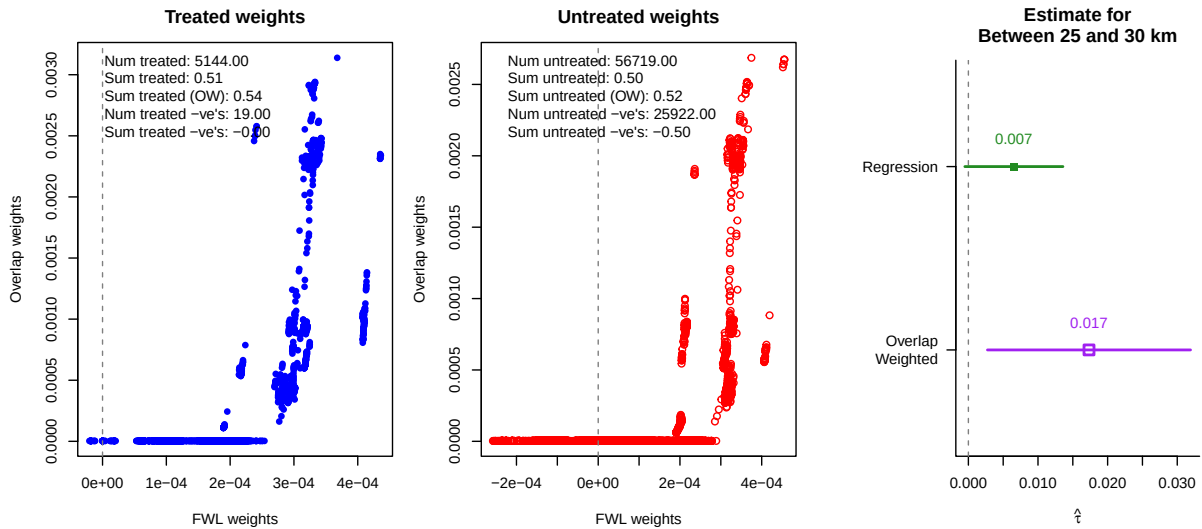


Figure B.8: The FWL weights and overlap weights for the distance category of 25 to 30km in Figure 1 of Hazlett and Mildenberger (2020)

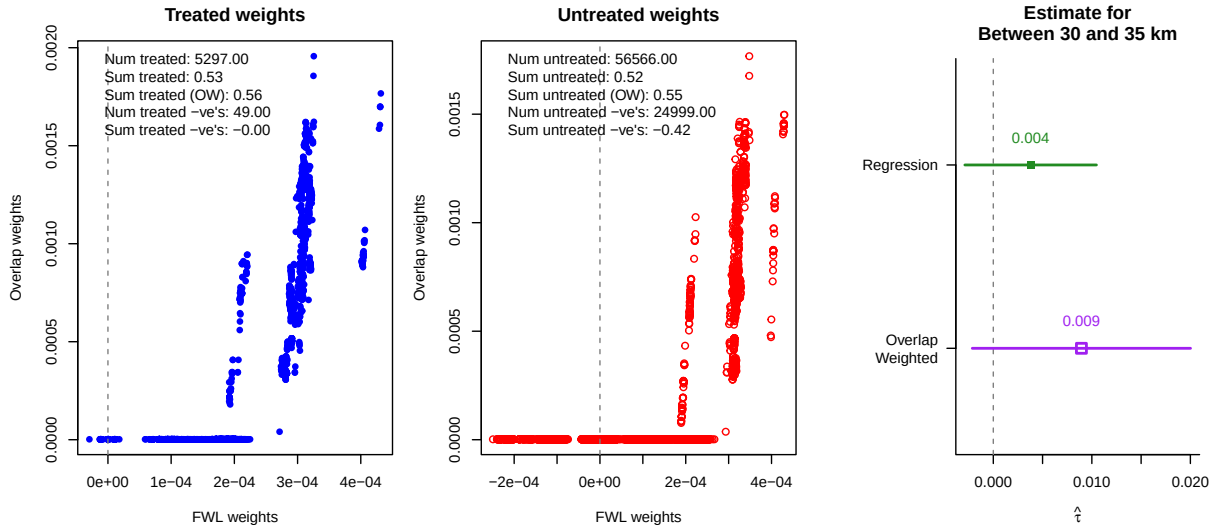


Figure B.9: The FWL weights and overlap weights for the distance category of 30 to 35km in Figure 1 of Hazlett and Mildenberger (2020)

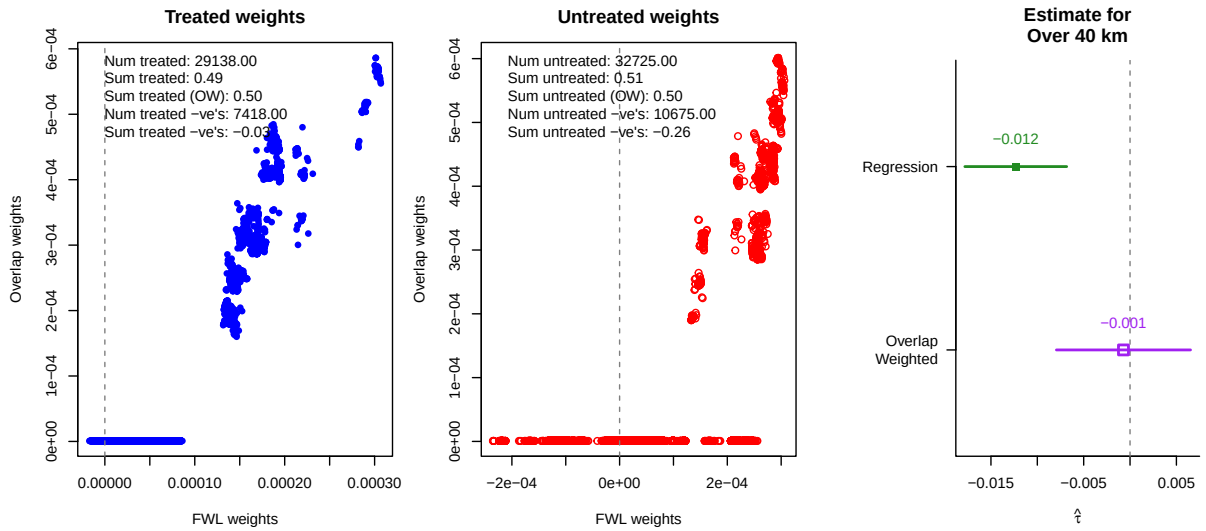


Figure B.10: The FWL weights and overlap weights for the distance category of over 40km in Figure 1 of Hazlett and Mildenberger (2020)

B.3 Additional results on Garcia-Montoya, Arjona, Lacombe (2022)

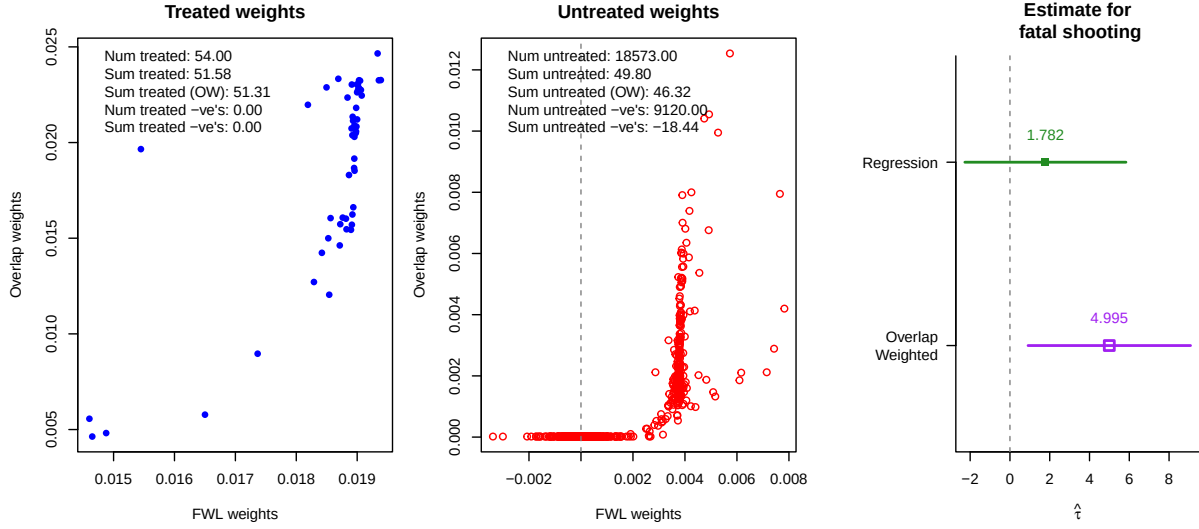


Figure B.11: The FWL weights and overlap weights for a fatal shooting in the county from row 2 of Figure 2 of Garcia-Montoya, Arjona, Lacombe (2022)

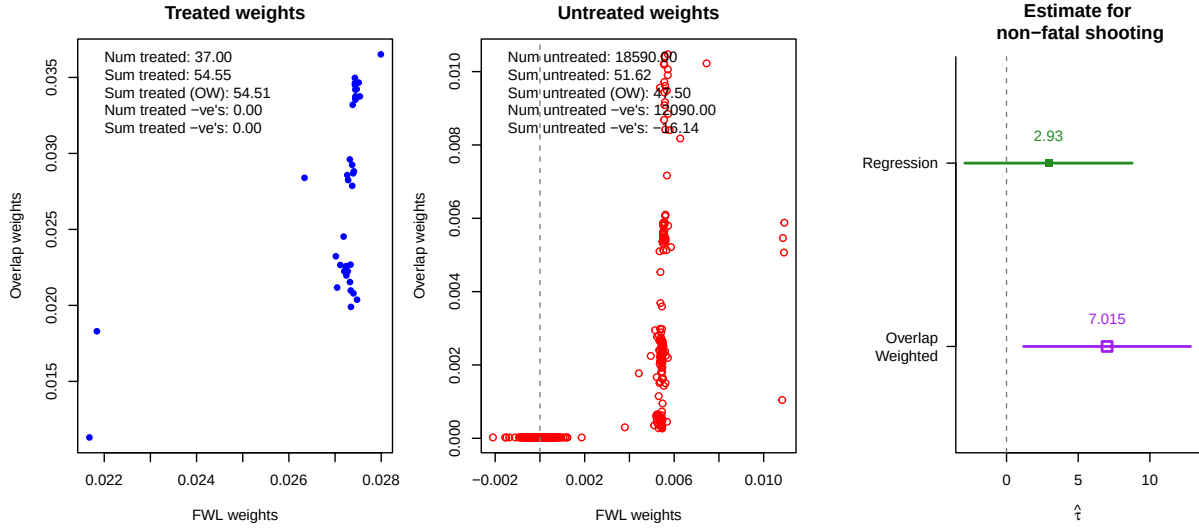


Figure B.12: The FWL weights and overlap weights for a nonfatal shooting in the county from row 2 of Figure 2 of Garcia-Montoya, Arjona, Lacombe (2022)

B.4 Additional results on Marsh (2023)

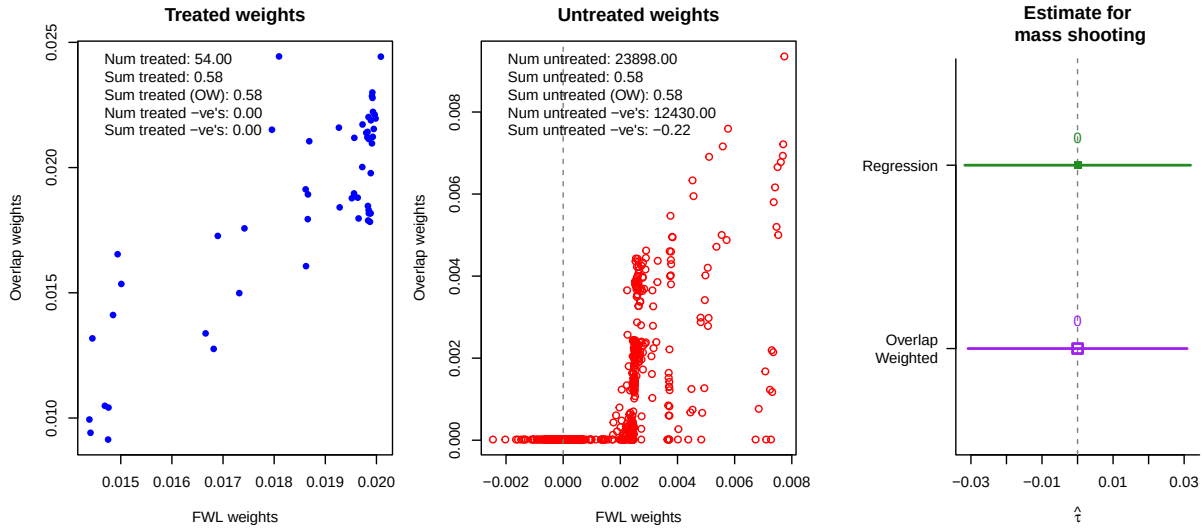


Figure B.13: The FWL weights and overlap weights for effect of shootings in Table 1 of Marsh (2023)

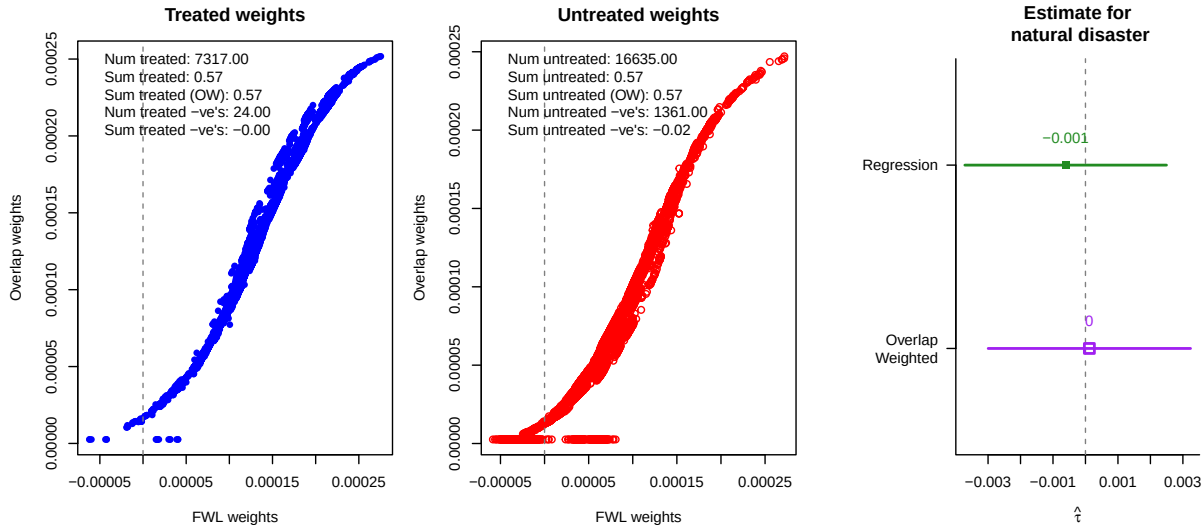


Figure B.14: The FWL weights and overlap weights for effect of natural disasters in Table 1 of Marsh (2023)