CONTAMINATION BIAS IN LINEAR REGRESSIONS

Paul Goldsmith-Pinkham Peter Hull Michal Kolesár

Oct 2022

- Interested in estimating “treatment effects” β in partly linear model

$$Y_i = \alpha + X_i'\beta + g(W_i) + U_i.$$

- Two key features:
  1. Multiple treatments: X_i is vector of (mutually exclusive) treatment indicators, $X_{ik} = \mathbb{1}\{D_i = k\}$
     - Underlying treatment $D_i \in \{0, \dots, K\}$
     - In paper: X_i treatment general vector, not necessarily discrete or mutually exclusive
  2. Necessary to include vector of controls W_i to prevent omitted variables bias (OVB)
- What is the interpretation of β if treatment effects are heterogeneous?

  Many examples…

1. Multi-armed RCT with variation across strata
    - D_i: set of treatments, W_i: strata FE
    - Examples: Project STAR (Krueger, 1999), RAND Health Insurance Experiment (Manning et al., 1987)
2. Value-added models in education, health, or development:
    - D_i: set of teachers, W_i: strata controls for as-if-random-assignment
    - Examples: evaluation of teachers (Kane & Staiger, 2008; Chetty et al., 2014), schools (Angrist et al., 2017; Angrist et al., 2021; Mountjoy & Hickman, 2020), or healthcare (Hull, 2018; Abaluck et al., 2021; Geruso et al., 2020), leader effectiveness (Easterly and Pennings, 2022)
3. Differences-in-differences and event studies
    - D_i: periods since treatment adoption, W_i: unit and time fixed effects (FE)
    - Our results related to Sun and Abraham (2021) and de Chaisemartin and D'Haultfœuille (2021).
4. First-stage + reduced-form in examiner designs (not focus of this talk):
    - D_i: set of judges, W_i: strata controls for as-if-random-assignment
    - Examples: Kling (2006), Maestas et al. (2013), Dobbie and Song (2015), and Arnold et al. (2020)

KNOWN SPECIAL CASE

- If D_i binary, so $X_i = D_i$, setting even more common and well-understood (e.g. Angrist & Pischke, 2009; Aronow & Samii, 2016)
- Influential result, known since at least Angrist (1998): β is (convex) weighted average treatment effect (ATE)
    - (Unweighted) ATE generally different
    - Weights proportional to variance of D_i conditional on W_i: strata with more variation in treatment receive more weight $\implies$ automatically deals with overlap issues
    - Used to justify estimating "treatment effect" of D_i using partly linear model

MAIN RESULTS

1. When D_i multi-valued, β no longer corresponds to convex combination of causal effects
    - β_k = weighted average of treatment effect of treatment k + contamination bias from other treatments, with weights summing to 0
    - Recent results in DiD literature arise as special case; negative weighting (Goodman-Bacon, 2021; de Chaisemartin & D'Haultfœuille, 2020) and contamination bias (Sun & Abraham, 2021; de Chaisemartin & D'Haultfœuille, 2021) conceptually distinct.
2. Provide solutions to contamination bias:

    (a) Estimate ATE directly: apply methods from ATE literature. Noisy with poor overlap.

    (b) Run one-treatment-at-a-time regression. New justification: implements weights that are "easiest to estimate" in that they minimize SEB for estimation of weighted average of treatment effects.

    - When D_i binary, SEB result gives new formal motivation for using partially linear model.

    (c) Use SEB bound to construct new estimator when considering all treatments equally

Simple example

General setting and result

Solutions

Empirical example and diagnostics

- To build intuition, first review Angrist's result when both W_i and D_i binary.
- Consider regression

$$Y_i = \alpha + \beta D_i + \gamma W_i + U_i,$$

with $D_i, W_i \in \{0, 1\}$. By definition, U_i mean-zero residual uncorrelated with (D_i, W_i)

- Stylized Project STAR example: D_i is small classroom dummy, Y_i is avg test score of student i
  - Randomization stratified: probability of assignment to small vs large classroom depends on school. W_i denotes school FE
  - Binary W_i: only 2 schools for simplicity

POTENTIAL OUTCOMES AND KEY ASSUMPTION

- To characterize β, use potential outcomes notation $Y_i(d)$
    - Individual treatment effect $\tau_i = Y_i(1) - Y_i(0)$, conditional treatment effect $\tau_1(w) = E[\tau_i \mid W_i = w]$
    - Observed outcome $Y_i = Y_i(0) + \tau_i D_i$
    - Propensity score: $p_1(W_i) = \Pr(D_i = 1 \mid W_i) = E[D_i \mid W_i]$
- Treatment (as good as) randomly assigned conditional on W_i: $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid W_i$
- Random assignment assumption delivers key result (Angrist, 1998):

$$\beta = \phi\tau_1(0) + (1-\phi)\tau_1(1), \quad \phi = \frac{\operatorname{var}(D_i \mid W_i = 0)\Pr(W_i = 0)}{\sum_{w=0}^{1} \operatorname{var}(D_i \mid W_i = w)\Pr(W_i = w)},$$

KEY FEATURES OF THIS ESTIMATOR

$$\beta = \phi\tau(0) + (1-\phi)\tau(1), \quad \phi = \frac{\operatorname{var}(D_i \mid W_i = 0)\Pr(W_i = 0)}{\sum_{w=0}^{1} \operatorname{var}(D_i \mid W_i = w)\Pr(W_i = w)},$$

- $\phi \in (0, 1)$
- No need to estimate propensity score
- Puts larger weight on strata with higher variation in D_i
    - $\neq$ ATE! (unless $\tau(w)$ constant or $p_1(w)$ constant across strata)
    - May lead to unusual or "unrepresentative" estimand (Aronow & Samii, 2016)
    - But this sort of weighting necessary to avoid loss of identification under overlap failure (e.g. $p_1(0) = 0$), or lack of precision under weak overlap ($p_1(0)$ close to 0)

MULTIPLE TREATMENTS

- Project STAR in fact had additional treatment arm in addition to small class ($D_i = 1$): full-time teaching aide ($D_i = 2$).

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma W_i + U_i,$$

- General notation:
  - $X_i = [X_{i1}, X_{i2}]', X_{ij} = \mathbb{1}\{D_i = j\}$
  - $Y_i = Y_i(0) + X_i' \tau_i$, where $\tau_{ik} = Y_k(k) - Y_i(0)$.
  - Let $\tau_k(W_i) = E[\tau_{ik} \mid W_i]$ and $p_k(w) = E[X_{ik} \mid W_i = w]$.
- Assignment still conditionally random, $(Y_i(0), Y_i(1), Y_i(2)) \perp X_i \mid W_i$

CAUSAL INTERPRETATION OF β_1

Due to FWL,

$$\begin{aligned}\beta_1 &= \frac{E[\tilde{\tilde{X}}_{i1} Y_i]}{E[\tilde{\tilde{X}}_{i1}^2]} = \frac{E[\tilde{\tilde{X}}_{i1} Y_i(0)]}{E[\tilde{\tilde{X}}_{i1}^2]} + \frac{E[\tilde{\tilde{X}}_{i1} X_{i1}\tau_{i1}]}{E[\tilde{\tilde{X}}_{i1}^2]} + \frac{E[\tilde{\tilde{X}}_{i1} X_{i2}\tau_{i2}]}{E[\tilde{\tilde{X}}_{i1}^2]} \\ &= E[\lambda_{11}(W_i)\tau_1(W_i)] + E[\lambda_{12}(W_i)\tau_2(W_i)],\end{aligned}$$

where $\lambda_{11}(W_i) = \frac{E[\tilde{\tilde{X}}_{i1} X_{i1} | W_i]}{E[\tilde{\tilde{X}}_{i1}^2]} \geq 0$, and $\lambda_{12}(W_i) = \frac{E[\tilde{\tilde{X}}_{i1} X_{i2} | W_i]}{E[\tilde{\tilde{X}}_{i1}^2]} \neq 0$ in general.

Key point $\tilde{\tilde{X}}_{i1}$ is residual from regressing X_{i1} on W_i, constant, and X_{i2}

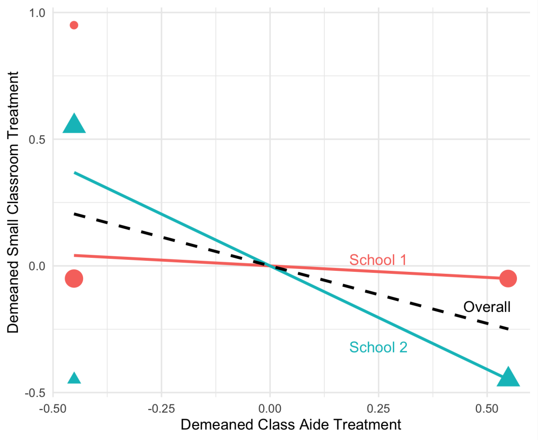
- $\tilde{\tilde{X}}_{i1} \neq X_{i1} - E[X_{i1} \mid W_i, X_{i2}]$, since X_{i2} depends non-linearly on X_{i1}
- As a result, β_1 contaminated by τ_{i2}.

$$\beta_1 = E[\lambda_{11}(W_i)\tau_1(W_i)] + E[\lambda_{12}(W_i)\tau_2(W_i)], \qquad \lambda_{12}(W_i) = \frac{E[\tilde{\tilde{X}}_{i1}X_{i2} \mid W_i]}{E[\tilde{\tilde{X}}_{i1}^2]} \neq 0.$$

- Why does this second contamination bias term arise?
- Consider single residualization:
  - $\tilde{X}_{ik} = X_{ik} - E[X_{ik} \mid W_i] = X_{ik} - p_k(W_i)$
  - $\tilde{\tilde{X}}_{i1}$ is residual from regressing $\tilde{X}_{i1}$ on $\tilde{X}_{i2}$, and relationship between them varies by school
  - If $p_k(0) \neq p_k(1)$, then line of best fit averages across this relationship and does not isolate conditional (i.e. within-school) variation in $X_{i1} \implies$ variation in $\tilde{\tilde{X}}_{i1}$ predicts X_{i2} within-schools

STYLIZED EXAMPLE 1

- Two (equal-sized) schools vary significantly in treatment assignment
  - School 0: $p_1(0) = 0.05$, $p_2(0) = 0.45$
  - School 1: $p_1(1) = 0.45$, $p_2(1) = 0.45$
- Under our formula, $\lambda_{12}(0) = 99/106$, $\lambda_{12}(1) = -99/106$
- To illustrate potential magnitude of contamination bias:
  - $\tau_1(W_i) = 0$, $\tau_2(0) = 0$, $\tau_2(1) = 1$
  - Then, $\beta_1 \approx -0.47$



STYLIZED EXAMPLE 2: NO OVERLAP

- Consider multiple strata, included as indicators, but only units in stratum $W_i = 0$ receive treatment 2. Let $n_k(w) = \sum_i \mathbb{1}\{W_i = w, X_i = k\}$.
- Then

$$\hat{\beta} = \begin{pmatrix} \sum_w \hat{\lambda}_{11}(w)\hat{\tau}_1(w) \\ \frac{n_1(0)}{n_0(0)+n_1(0)} \sum_{w\neq 0} \lambda(w)\left[\hat{\tau}_1(w) - \hat{\tau}_1(0)\right] + \hat{\tau}_2(0) \end{pmatrix},$$

where $\hat{\lambda}_{11}(w) = (n_0(w) + n_1(w))\hat{V}_1(w)/\sum_w (n_0(w) + n_1(w))\hat{V}_1(w)$. and $\hat{V}_1(w) = n_1(w)n_0(w)/(n_0(w) + n_1(w))^2$ is an estimate of $\operatorname{var}(X_{i1} \mid W_i = w, D_i \neq 2)$.

Simple example

General setting and result

Solutions

Empirical example and diagnostics

GENERAL SETTING AND RESULT

- Interested in effect of X_i on Y_i estimated by a partially linear model,

$$Y_i = X_i'\beta + g(W_i) + U_i,$$

where β and $g(\cdot)$ defined as minimizers of expected squared residuals $E[U_i^2]$:

$$(\beta, g) = \underset{\tilde{\beta}\in\mathbb{R}^K, \tilde{g}\in\mathcal{G}}{\operatorname{argmin}} \; E[(Y_i - X_i'\tilde{\beta} - \tilde{g}(W_i))^2]$$

for some linear space of functions $\mathcal{G}$.

  - Linear covariate adjustment: $\mathcal{G} = \{\alpha + w'\gamma \colon (\alpha, \gamma')' \in \mathbb{R}^{1+\dim(W_i)}\}$,
  - Allow for flexible adjustments if $\mathcal{G}$ large class of "nonparametric" functions (Robinson, 1988).

EXAMPLES

$$Y_i = X_i'\beta + g(W_i) + U_i,$$

Multi-armed RCT

W_i are strata indicators, X_i treatment indicators.

Event study / Two-way fixed effects

Panel data, $i = (j, t)$ where j is unit and t is time. $g(\cdot)$ is linear, and W_i contains unit and time indicators. X_i contains leads and lags relative to (deterministic) treatment adoption date $A(j)$. Or indicators for multiple treatments (“mover regressions”).

- Let $\tilde{X}_i$ denote residual from projecting X_i onto $\mathcal{G}$. Then by projection theorem,

$$\beta = E[\tilde{X}_i \tilde{X}_i']^{-1} E[\tilde{X}_i Y_i].$$

- Hence, by FWL

$$\beta_k = \frac{E[\tilde{\tilde{X}}_{ik} Y_i]}{E[\tilde{\tilde{X}}_{ik}^2]}.$$

  Where $\tilde{\tilde{X}}_{ik}$ is residual from regressing $\tilde{X}_{i,k}$ on $\tilde{X}_{i,-k}$. Equivalently, $\tilde{\tilde{X}}_{ik} = X_{ik} - E^*[X_{ik} \mid X_{i,-k}, W_i]$ where E^* is projection on $X_{i,-k}'\delta + g(W_i)$.

- Does β_k have a causal interpretation?

ASSUMPTIONS

Assumption 1: Random assignment

$E[Y_i(k) \mid D_i, W_i] = E[Y_i(k) \mid W_i]$ for all k.

- Let $\mu_0(w) = E[Y_i(0) \mid W_i = w]$.

Assumption 2: Correct outcome or assignment model

Either $\mu_0 \in \mathcal{G}$ or $p_k \in \mathcal{G}$ for all k.

- $\mu_0 \in \mathcal{G}$ is "parallel trends" assumption in event studies. p_k degenerate and not in $\mathcal{G}$.
- $p_k \in \mathcal{G}$ automatic in stratified RCT example

Main result

Suppose Assumptions 1 and 2 hold. Then

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

where

$$\lambda_{kk}(W_i) = \frac{E[\tilde{\tilde{X}}_{ik} X_{ik} \mid W_i]}{E[\tilde{\tilde{X}}_{ik}^2]}, \qquad \lambda_{k\ell}(W_i) = \frac{E[\tilde{\tilde{X}}_{ik} X_{i\ell} \mid W_i]}{E[\tilde{\tilde{X}}_{ik}^2]}.$$

The weights satisfy $E[\lambda_{kk}(W_i)] = 1$ and $E[\lambda_{k\ell}(W_i)] = 0$. Furthermore, $\lambda_{kk}(W_i) \geq 0 \iff E^*[X_{ik} \mid X_{i,-k} = 0, W_i] \leq 1$. A sufficient condition is $p_k \in \mathcal{G}$.

INTERPRETATION

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

Two conceptually distinct issues:

1. If $p_1 \notin \mathcal{G}$, then weights not necessarily positive, even with $K = 1$.
   - If $p_1 \in \mathcal{G}$, $\lambda_{11}(W_i) = \text{var}(X_i \mid W_i)/E[\text{var}(X_i \mid W_i)] \geq 0$
2. If $K > 1$, then additional contamination bias present unless
   2.1 $E[X_{ik} \mid X_{i,-k}, W_i] = E^*[X_{ik} \mid X_{i\ell}, W_i]$: no non-linear dependence across treatments, or treatment completely randomized $\implies \lambda_{kl}(W_i) = 0$
   2.2 Uncorrelated heterogeneity: $E[\lambda_{k\ell}(W_i)\tau_\ell(W_i)] = 0$

RELATIONSHIP WITH DID LITERATURE

- In DiD or event studies, focus on two-way FE specification: $\mathcal{G}$ is a set of unit + time effects. D_i may index time since treatment adoption, or else static multivalued treatment
- "Model-based" parallel trends assumption $E[Y_i(0) \mid W_i = w] \in \mathcal{G}$, but $p \notin \mathcal{G}$ virtually by design. Then
  - If $K = 1$, weights $\lambda_{kk}(W_i)$ may not be positive (de Chaisemartin & D'Haultfœuille, 2020; Goodman-Bacon, 2021)
  - If $K > 1$, also contamination bias (de Chaisemartin & D'Haultfœuille, 2021; Sun & Abraham, 2021)
- Our result shows issue not specific to two-way FE specification of $\mathcal{G}$. Instead, issue is:
  1. Contamination bias: non-linear dependence among treatments
  2. Own weights could be negative: propensity score not in $\mathcal{G}$ (restrictive here, albeit implied by setup in Athey and Imbens, 2022)

Simple example

General setting and result

Solutions

Empirical example and diagnostics

SOLUTION 1: ESTIMATE ATE

- Focus on $p_k \in \mathcal{G}$ case
- Most principled solution: estimate ATEs $E[\tau(W_i)]$ directly, using your favorite method (propensity score weighting, matching, regression etc). E.g. regression implementation:

$$Y_i = X_i'\beta + q_0(W_i) + \sum_{k=1}^{K} X_{ik}\left(q_k(W_i) - E[q_k(W_i)]\right) + U_i, \qquad \beta = E[\tau(W_i)]$$

  Note relevant for IV setting: adds many more additional controls
- Key issue in practice: poor overlap $\implies$ large standard errors. Salient once K moderate.
  - Suppose new Medicaid enrollees randomized among K plans. Monthly variation in capacity means month FE necessary.
  - Little/no enrollment in plan k in some months $\implies$ ATE estimate very noisy/unidentified.

SOLUTIONS 2 AND 3: FOCUS ON WEIGHTED ATE THAT IS EASY TO ESTIMATE

- Regression approach adapts to overlap by putting less weight on strata with less overlap
    - This is why it is popular in practice with binary treatment!
    - Can we generalize this advantage without introducing contamination bias?
- Our strategy:
    1. Derive semiparametric efficiency bound for such a weighted ATE under idealized conditions
    2. Identify which weights lead to smallest efficiency bound
    3. Construct efficient feasible estimator of this weighted ATE
- Consider conditional potential outcome contrasts: $\sum_{k=0}^{K} c_k \mu_k(W_i)$, where $\mu_k(W_i) = E[Y_i(k) \mid W_i]$ and c is a $(K+1)$-dimensional vector.
    - If we set $c_k = 1$, $c_0 = -1$ and all other entries of c to zero: $\sum_{k=0}^{K} c_k \mu_k(W_i) = \tau_k(W_i)$.
    - Conditions on realized $\{W_i\}_{i=1}^{n}$

STEP 1: SEMIPARAMETRIC EFFICIENCY BOUND

Proposition

Consider i.i.d. sample of size N that satisfies Assumption 1. Suppose $p_k(W_i)$ known, and let $\sigma_k^2(W_i) = \operatorname{var}(Y_i(k) \mid W_i)$. Consider estimating weighted avg of contrasts

$$\theta_{\lambda,c} = \frac{1}{\sum_{i=1}^{N} \lambda(W_i)} \sum_{i=1}^{N} \lambda(W_i) \sum_{k=0}^{K} c_k \mu_k(W_i),$$

where λ and c are known. Then, conditional on $\{W_i\}_{i=1}^{N}$,

$$\mathcal{V}_{\lambda,c} = E\Big[\textstyle\sum_{k=0}^{K} \frac{\lambda(W_i)^2 c_k^2 \sigma_k^2(W_i)}{p_k(W_i)} \Big] / E[\lambda(W_i)]^2.$$

is a.s. the semiparametric efficiency bound.

SOLUTION 2: ONE-TREATMENT-AT-A-TIME REGRESSION

- Suppose we are interested in single contrast, so contrast vector c^k sets $c_j^k = 1$ if $j = k$, $c_j^k = -1$ if $j = 0$, and $c_j^k = 0$ otherwise. Suppose further conditional variance homoskedastic: $\sigma_k^2(W_i) = \sigma^2$.
- Minimizing $\mathcal{V}_{\lambda,c^k}$ over weights λ yields efficient weights and minimal asymptotic variance

$$\lambda^k(W_i) = \frac{p_0(W_i)p_k(W_i)}{p_0(W_i) + p_k(W_i)}, \; \mathcal{V}_{\lambda^k,c^k} = \sigma^2 E\left[\frac{p_0(W_i)p_k(W_i)}{p_0(W_i) + p_k(W_i)}\right]^{-1},$$

- Precisely the weighting of one-treatment-at-a-time regression when we fit partially linear model only using observations with $D_i \in \{0, k\}$
- Result gives formal justification for using regression to estimate effects of single treatment.

SOLUTION 3: EFFICIENTLY WEIGHTED CONTRASTS

- One-treatment-at-a-time regression estimates not comparable across treatments (same is true for usual estimator based on full data, even in absence of contamination bias).
- If consider all $K(K+1)$ potential contrasts $\mu_j(W_i) - \mu_k(W_i)$ equally important, then natural to minimize average variance $\int V_{\lambda,c} dF(c)$, where F uniform over contrasts
- Equivalent to setting $c_k^2 = 2/(K+1)$, which under homoskedasticity leads to

$$\lambda^{\mathrm{C}}(W_i) = \frac{1}{\sum_{k=0}^{K} p_k(W_i)^{-1}}.$$

$$\lambda^{\mathrm{C}}(W_i) = \frac{1}{\sum_{k=0}^{K} p_k(W_i)^{-1}}$$

- Weights generalize intuition behind single binary treatment: place more weight on strata with evenly distributed treatments, less weight on strata with overlap problem.
- Weights are the same for every treatment contrast c, so $\beta_{\lambda^{\mathrm{C}},k} - \beta_{\lambda^{\mathrm{C}},\ell}$ is a convex weighted average of relative causal effects for all k, ℓ.
- Efficient estimator: weighted regression of Y_i onto X_i, weighting each observation by $\hat{\lambda}^{\mathrm{C}}(W_i)/\hat{p}_{D_i}(W_i)$.

WHY LOOK FOR BRIGHTEST LIGHT?

Three motivations for Solutions 2 and 3:

1. Robustness concern: want to estimate given contrast as efficiently as possible, at least under benchmark of constant treatment effects, while being robust to the possibility that effects heterogeneous.
2. Gives bound on information available in the data: if these weights λ^{C} or λ^{k} yield overly large standard errors, inference on other treatment effects (such as the unweighted ATE) will be at least as uninformative.
3. General efficiency vs robustness trade-off: estimators of unweighted ATE most robust, but least efficient. OLS most efficient, but not at all robust to heterogeneity in treatment effects. Weights λ^{k}/λ^{C} land closer to the middle.

Simple example

General setting and result

Solutions

Empirical example and diagnostics

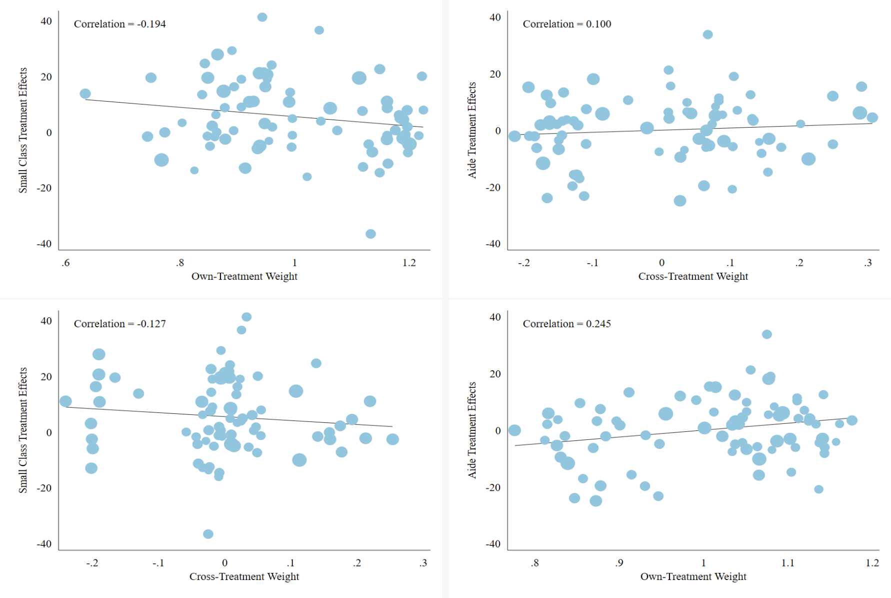
PROJECT STAR

A. Contamination Bias Estimates

	Regression	Own	Bias	Worst-Case Bias	
	Coefficient	Effect		Negative	Positive
Small Class Size	5.357	5.202	0.155	-1.654	1.670
	(0.778)	(0.778)	(0.160)	(0.185)	(0.187)
Teaching Aide	0.177	0.360	-0.183	-1.529	1.530
	(0.720)	(0.714)	(0.149)	(0.176)	(0.177)

B. Treatment Effect Estimates

	Unweighted	Efficiently-Weighted	
	(ATE)	One-at-a-time	Common
Small Class Size	5.561	5.295	5.563
	(0.763)	(0.775)	(0.764)
Teaching Aide	0.070	0.263	−0.003
	(0.708)	(0.715)	(0.712)


Correlation = -0.194
Small Class Treatment Effects
40
20
0
-20
-40
.6
.8
1
1.2
Own-Treatment Weight
Correlation = 0.100
Aide Treatment Effects
40
20
0
-20
-40
-.2
-.1
0
.1
.2
.3
Cross-Treatment Weight
Correlation = -0.127
Small Class Treatment Effects
40
20
0
-20
-40
-.2
-.1
0
.1
.2
.3
Cross-Treatment Weight
Correlation = 0.245
Aide Treatment Effects
40
20
0
-20
-40
.8
.9
1
1.1
1.2
Own-Treatment Weight

EMPIRICAL EXAMPLE: TAKEAWAY

- Considerable variation in weights as well as in treatment effects
- But heterogeneity in treatments appears to be uncorrelated with weights, so contamination bias small
    - Generality of this result open question
    - Issues we flag in this paper may be salient under “optimal” stratification.
- Diagnostics implementation details in paper and stata package `multe`