July 25, 2023

Note: Contamination Bias in Linear Regressions

as in Goldsmith-Pinkham et al. (2022) *by Sai Zhang*

Key points: The contamination bias arises in multiple-treatment regression even when the treatment assignment is as good as random, due to the **inherent nonlinear dependence** of mutually exclusive treatment indicators.

Disclaimer: *This note is built on Goldsmith-Pinkham et al. (2022).*

1 Motivation

Consider the regression

$$Y_i = \alpha + \beta D_i + \gamma W_i + U_i \tag{1}$$

where

- $D_i \in \{0, 1\}$ is a single treatment indicator
- $W_i \in \{0, 1\}$ is a single binary control
- U_i is a mean-zero residual uncorrelated with D_i and W_i

Assume the (within-strata) treatment assignment is random, i.e., conditionally independent of potential outcomes given the control:

$$(Y_i(0), Y_i(1)) \perp D_i \mid W_i \tag{2}$$

where $Y_i(d)$ is the outcome of individual i when $D_i = d$, i's treatment effect is given by $\tau_i = Y_i(1) - Y_i(0)$, and the realized outcome is $Y_i = Y_i(0) + \tau_i D_i$.

By Angrist (1998), β in Eq (1) identifies a weighted average of within-strata ATEs with **convex** weights:

$$\beta = \phi\tau(0) + (1-\phi)\tau(1) \qquad \text{where } \phi = \frac{\text{var}(D_i \mid W_i = 0)\Pr(W_i = 0)}{\sum_{w=0}^{1}\text{var}(D_i \mid W_i = w)\Pr(W_i = w)} \in [0, 1] \tag{3}$$

and

$$\tau(w) = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid W_i = w\right]$$

is the ATE in the strata indexed by control $W_i = w$.

By appying the Frisch-Waugh-Lovell (FWL) Theorem, β can be written as the univariate regression coefficient of regression Y_i on $\tilde{D}_i$[1]:

$$\begin{aligned}\beta = \frac{\mathbb{E}\left[\tilde{D}_i Y_i\right]}{\mathbb{E}\left[\tilde{D}_i^2\right]} &= \frac{\mathbb{E}\left[\tilde{D}_i Y_i(0)\right]}{\mathbb{E}\left[\tilde{D}_i^2\right]} + \frac{\mathbb{E}\left[\tilde{D}_i D_i \tau_i\right]}{\mathbb{E}\left[\tilde{D}_i^2\right]} \\ &= \frac{\mathbb{E}\left[\mathbb{E}\left[\tilde{D}_i Y_i(0) \mid W_i\right]\right]}{\mathbb{E}\left[\tilde{D}_i^2\right]} + \frac{\mathbb{E}\left[\mathbb{E}\left[\tilde{D}_i D_i \tau_i \mid W_i\right]\right]}{\mathbb{E}\left[\tilde{D}_i^2\right]} \\ &= \frac{\mathbb{E}\left[\mathbb{E}\left[\tilde{D}_i \mid W_i\right]\mathbb{E}\left[Y_i(0) \mid W_i\right]\right]}{\mathbb{E}\left[\tilde{D}_i^2\right]} + \frac{\mathbb{E}\left[\mathbb{E}\left[\tilde{D}_i D_i \mid W_i\right]\mathbb{E}\left[\tau_i \mid W_i\right]\right]}{\mathbb{E}\left[\tilde{D}_i^2\right]} \\ &= 0 + \frac{\mathbb{E}\left[\text{var}(D_i \mid W_i)\tau(W_i)\right]}{\mathbb{E}\left[\text{var}(D_i \mid W_i)\right]}\end{aligned} \tag{4}$$

for the derivation in Eq. (4) to work, the key underlying point is that $\mathbb{E}\left[\tilde{D}_i \mid W_i\right] = 0$, i.e., $\tilde{D}_i$ is **mean-independent** of W_i and the propensity score $\mathbb{E}\left[D_i \mid W_i\right]$ is **linear** since W_i is **binary**.

Where contamination bias arises Now add an **additional treatment arm**: consider 2 **mutually exclusive** interventions: $D_i \in \{0, 1, 2\}$, represented by a vector of 2 treatment indicators $\mathbf{X}_i = (X_{i1}, X_{i2})'$ where

$$X_{i1} = \mathbf{1}\{D_i = 1\} \qquad X_{i2} = \mathbf{1}\{D_i = 2\}$$

this yields the regression

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma W_i + U_i \tag{5}$$

now the observe outcome is given by $Y_i = Y_i(0) + \tau_{i1} X_{i1} + \tau_{i2} X_{i2}$

$$\tau_{i1} = Y_i(1) - Y_i(0) \qquad \tau_{i2} = Y_i(2) - Y_i(0)$$

hence, some heterogeneity in treatment effect emerges. Still, assume $\mathbf{X}_i$ is conditionally independent of $Y_i(d)$ given control W_i

$$(Y_i(0), Y_i(1), Y_i(2)) \perp \mathbf{X}_i \mid W_i$$

If we still use FWL theorem to derive β_1

$$\beta_1 = \frac{\mathbb{E}\left[\tilde{X}_{i1} Y_i\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]} = \frac{\mathbb{E}\left[\tilde{X}_{i1} Y_i(0)\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]} + \frac{\mathbb{E}\left[\tilde{X}_{i1} X_{i1} \tau_{i1}\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]} + \frac{\mathbb{E}\left[\tilde{X}_{i1} X_{i2} \tau_{i2}\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]} \tag{6}$$

where $\tilde{X}_{i1}$ is obtained by running $X_{i1} = a + bX_{i2} + cW_i + \tilde{X}_{i1}$. Now, the issues is that X_{i1} and X_{i2} are **mutually exclusive**:

- If $X_{i2} = 1$: $X_{i1} = 0$, **not** depends on W_i
- If $X_{i2} = 0$: the mean of X_{i1} depends on W_i

hence in general

$$\tilde{X}_{i1} \neq X_{i1} - \mathbb{E}\left[X_{i1} \mid W_i, X_{i2}\right]$$

[1] $\tilde{D}_i$ is the residual of regressing D_i on W_i and a constant:

$$D_i = a + bW_i + \tilde{D}_i$$

which means that we can only derive β_1 from Eq. (6) as

$$\begin{aligned}\beta_1 &= \frac{\mathbb{E}\left[\tilde{X}_{i1}Y_i(0)\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]} + \frac{\mathbb{E}\left[\tilde{X}_{i1}X_{i1}\tau_{i1}\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]} + \frac{\mathbb{E}\left[\tilde{X}_{i1}X_{i2}\tau_{i2}\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]} \\ &= 0 + \mathbb{E}\left[\lambda_{11}(W_i)\tau_1(W_i)\right] + \mathbb{E}\left[\lambda_{12}(W_i)\tau_2(W_i)\right] \end{aligned} \tag{7}$$

breakdown each term:

- $\mathbb{E}\left[\tilde{X}_{i1}Y_i(0)\right]/\mathbb{E}\left[\tilde{X}_{i1}^2\right]$: FWL regression residuals are **uncorrelated** with $Y_i(0)$

$$\begin{aligned} &X_{i1} = a + bX_{i2} + cW_i + \tilde{X}_{i1} \\ \xRightarrow[\text{on both sides}]{\text{purge } W_i} &\tilde{\tilde{X}}_{i1} = \mu_1\tilde{\tilde{X}}_{i2} + \tilde{X}_{i1} \Rightarrow \tilde{X}_{i1} = \tilde{\tilde{X}}_{i1} - \mu_1\tilde{\tilde{X}}_{i2} \xRightarrow{(Y_i(0),Y_i(1),Y_i(2))\perp\mathbf{X}_i|W_i} \mathbb{E}\left[\tilde{X}_{i1}Y_i(0)\right] = 0 \end{aligned}$$

- $\mathbb{E}\left[\tilde{X}_{i1}X_{i1}\tau_{i1}\right]/\mathbb{E}\left[\tilde{X}_{i1}^2\right]$: similarly to Eq. (4),

$$\frac{\mathbb{E}\left[\tilde{X}_{i1}X_{i1}\tau_{i1}\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]} = \frac{\mathbb{E}\left[\mathbb{E}\left[\tilde{X}_{i1}X_{i1}\tau_{i1} \mid W_i\right]\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]} \xRightarrow{(Y_i(0),Y_i(1),Y_i(2))\perp\mathbf{X}_i|W_i} = \mathbb{E}\left[\underbrace{\frac{\mathbb{E}\left[\tilde{X}_{i1}X_{i1} \mid W_i\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]}}_{\equiv\lambda_{11}(W_i)}\tau_1(W_i)\right]$$

  here, $\lambda_{11}(W_i)$ is still non-negative and average to one, hence similar to Eq. (4), this term is still a convex average of the conditional ATEs $\tau_1(W_i)$.

- $\mathbb{E}\left[\tilde{X}_{i1}X_{i2}\tau_{i2}\right]/\mathbb{E}\left[\tilde{X}_{i1}^2\right]$: on the contrary,

$$\frac{\mathbb{E}\left[\tilde{X}_{i1}X_{i2}\tau_{i2}\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]} = \frac{\mathbb{E}\left[\mathbb{E}\left[\tilde{X}_{i1}X_{i2}\tau_{i2} \mid W_i\right]\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]} \xRightarrow{(Y_i(0),Y_i(1),Y_i(2))\perp\mathbf{X}_i|W_i} = \mathbb{E}\left[\underbrace{\frac{\mathbb{E}\left[\tilde{X}_{i1}X_{i2} \mid W_i\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]}}_{\equiv\lambda_{12}(W_i)}\tau_2(W_i)\right]$$

  here $X_{i2} \neq X_{i1} - \mathbb{E}\left[X_{i1} \mid W_i, X_{i2}\right]$, hence $\lambda_{12}(W_i)$ is generally **non-zero**. This term is essentially the **contamination bias**.

How to simply understand contamination bias? As shown above,

$$\mathbb{E}\left[\frac{\mathbb{E}\left[\tilde{X}_{i1}X_{i2} \mid W_i\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]}\tau_2(W_i)\right] \equiv \mathbb{E}\left[\lambda_{12}(W_i)\tau_2(W_i)\right] \neq 0$$

arises because $\tilde{X}_{i1}$ is **uncorrelated** with X_{i2} by construction, but **NOT** conditionally independent of X_{i2}. To understand this, consider a two-step residualization:

- **Step 1**: first, demean X_{i1} and X_{i2}, conditional on W_i

$$\hat{X}_{i1} = X_{i1} - \mathbb{E}\left[X_{i1} \mid W_i\right] = X_{i1} - p_1(W_i) \qquad \hat{X}_{i2} = X_{i2} - \mathbb{E}\left[X_{i2} \mid W_i\right] = X_{i2} - p_2(W_i)$$

  where $p_j(W_i) = \mathbb{E}\left[X_{ij} \mid W_i\right]$ gives the propensity score for treatment j

- **Step 2**: run a bivarite regression

$$\hat{X}_{i1} = \alpha \hat{X}_{i2} + \tilde{X}_{i1}$$

Therefore, when the propensity scores vary across different strata ($W_i = w_a$ v.s. $W_i = w_b$), that is

$$p_j(w_a) \neq p_j(w_b)$$

the regression in Step 2 would also preserve this strata heterogeneity, leading to the *contamination weight* $\lambda_{12}(W_i)$ non-zero.

A numerical example Consider $W_i \in \{0, 1\}$ and a two-arm treatment assignment $D_i \in \{0, 1, 2\}$

	$D_i = 0$	$D_i = 1$	$D_i = 2$
$W_i = 0$	50%	5%	45%
$W_i = 1$	10%	45%	45%

and the 2 strata have equal number of observations, then we have the propensity score

$$p_1(0) = 0.05 \qquad p_2(0) = 0.45 \qquad p_1(1) = p_2(1) = 0.45$$

Then the *contamination weights* are

$$\lambda_{12}(W_i = 0) = \frac{\mathbb{E}\left[\tilde{X}_{i1} X_{i2} \mid W_i = 0\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]} = \frac{99}{106} \qquad \lambda_{12}(W_i = 1) = \frac{\mathbb{E}\left[\tilde{X}_{i1} X_{i2} \mid W_i = 1\right]}{\mathbb{E}\left[\tilde{X}_{i1}^2\right]} = -\frac{99}{106}$$

If we calculate the conditional correlation of the 2 within-strata residualzied treatments

$$\text{corr}\left(\tilde{X}_{i1}, \tilde{X}_{i2} \mid W_i\right) = -\sqrt{\frac{p_1(W_i)}{1 - p_1(W_i)}} \cdot \sqrt{\frac{p_2(W_i)}{1 - p_2(W_i)}}$$

then we have

$$\text{corr}\left(\tilde{X}_{i1}, \tilde{X}_{i2} \mid W_i = 0\right) = -0.2075 \qquad \text{corr}\left(\tilde{X}_{i1}, \tilde{X}_{i2} \mid W_i = 1\right) = -0.8182$$

the overall regression of $\tilde{X}_{i1}$ on $\tilde{X}_{i2}$ would be correlated with X_{i2} within each strata, hence misspecified (see Figure 1 for an illustration).

When is the contamination bias 0? We have derived

$$\beta_1 = \mathbb{E}\left[\lambda_{11}(W_i)\tau_1(W_i)\right] + \underbrace{\mathbb{E}\left[\lambda_{12}(W_i)\tau_2(W_i)\right]}_{\text{contamination bias}}$$

now consider 2 scenarios where the contamination bias vanishes

- **Case 1**: constant treatment effects of the 2nd treatment arm $\tau_2(W_i) \equiv \tau_2$, then

$$\beta_1 = \mathbb{E}\left[\lambda_{11}(W_i)\tau_1(W_i)\right] + \mathbb{E}\left[\lambda_{12}(W_i)\tau_2(W_i)\right] = \mathbb{E}\left[\lambda_{11}(W_i)\tau_1(W_i)\right] + \underbrace{\mathbb{E}\left[\lambda_{12}(W_i)\right]}_{=0}\tau_2 = \mathbb{E}\left[\lambda_{11}(W_i)\tau_1(W_i)\right]$$

  more generally, the **less heterogeneous** the treatment effect of the 2nd treatment arm $\tau_2(W_i)$ is (or the **less correlated** it is with $\lambda_{11}(W_i)$), the smaller the contamination bias is.

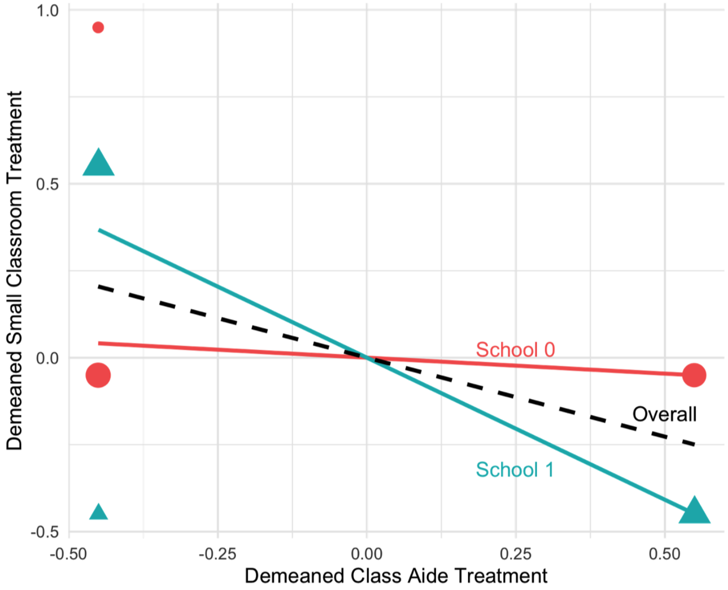

Figure 1: An Example of Contamination Bias (Goldsmith-Pinkham et al., 2022, Figure 1)

- **Case 2**: X_{i1} and X_{i2} are **independent** conditional on W_i[2], that is

  $$\mathbb{E}\left[X_{i1} \mid W_i, X_{i2}\right] = \mathbb{E}\left[X_{i1} \mid W_i\right]$$

  which solves the issue naturally, guaranteeing

  $$\tilde{X}_{i1} = X_{i1} - \mathbb{E}\left[X_{i1} \mid W_i, X_{i2}\right] = X_{i1} - \mathbb{E}\left[X_{i1} \mid W_i\right]$$

  naturally, conditional independence of X_{i1} and X_{i2} brings contamination bias to 0, reducing it to a single treatment problem.

How to solve the issue? An intuitive and simple solution to the problem is just including an interaction term between W_i and X_{i2} in Eq. (5)

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma W_i + \xi X_{i2} \times W_i + V_i$$

then the FWL regression

$$X_{i1} = a + bX_{i2} + cW_i + dX_{i2} \times X_i + \tilde{X}_{i1}$$

is saturated and capture the nonlinearity in $\mathbb{E}\left[X_{i1} \mid W_i, X_{i2}\right]$.

2 General Characterization

Now, consider a general characterization of the intuition above. For a partially linear model

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + g(\mathbf{W}_i) + U_i \tag{8}$$

where

[2] That is, they are no longer mutually exclusive: individuals can be assigned to both treatment arms with a non-zero probability.

- treatment indictors $\mathbf{X}_i = (X_{i1}, \cdots, X_{iK})'$: $X_{ik} = \mathbf{1}\{D_i = k\}$ for a $K-$arm mutually exclusive treatment assignment $D_k \in \{0, \cdots, K\}$
- vector of control variables $\mathbf{W}_i$

and $\boldsymbol{\beta}$ and g are defined as

$$(\boldsymbol{\beta}, g) = \arg \min_{\tilde{\beta} \in \mathbb{R}^K, \tilde{g} \in \mathcal{G}} \mathbb{E}\left[\left(Y_i - \mathbf{X}_i' \tilde{\boldsymbol{\beta}} - \tilde{g}(\mathbf{W}_i)\right)^2\right] \tag{9}$$

this characterization includes 2 of the most common applications

- **multi-arimed RCT**: $\mathbf{W}_i$ is the vector of indicators for experimental strata, within which $\mathbf{X}_i$ is randomly assigned to individual i, g is linear
- **two-way FEs**: For a fixed unit $j \in \{1, \cdots, J\}$ and period $t \in \{1, \cdots, T\}$, $W_i = (J_i, T_i)$ indicates the underlying unit and period for each observation i in a panel data where $J_i = j, T_i = t$, and $g(W_i) = \alpha + (\mathbf{1}\{J_i = 2\}, \cdots, \mathbf{1}\{J_i = n\}, \mathbf{1}\{T_i = 2\}, \mathbf{1}\{T_i = T\})' \boldsymbol{\gamma}$ includes unit and period indicators. $\mathbf{X}_i$ contains **lead** and **lag** indicators relative to a treatment adoption date $A(j) \in \{1, \cdots, T\}$

2.1 Derive Treatment Coefficients

To derive β, solve the minimization problem in Eq. (9), get

$$\boldsymbol{\beta} = \mathbb{E}\left[\tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_i'\right]^{-1} \mathbf{E}[\tilde{\mathbf{X}}_i Y_i] \tag{10}$$

where $\tilde{\mathbf{X}}_i$ is the residual vector from projecting $\mathbf{X}_i$ onto the controls:

$$\tilde{\mathbf{X}}_{ik} = \mathbf{X}_{ik} - \arg \min_{\tilde{g} \in \mathcal{G}} \mathbb{E}\left[\left(\tilde{\mathbf{X}}_{ik} - \tilde{g}\left(\mathbf{W}_i\right)\right)^2\right]$$

Following FWL theorem, each treatment coefficient of $\boldsymbol{\beta}$ can be written as

$$\boldsymbol{\beta}_k = \frac{\mathbb{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik} Y_i\right]}{\mathbb{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik}^2\right]}$$

where $\tilde{\tilde{\mathbf{X}}}_{ik}$ is the residual from regression $\tilde{\mathbf{X}}_{ik}$ on $\tilde{\mathbf{X}}_{i,-k} = \left(\tilde{\mathbf{X}}_{i1}, \cdots, \tilde{\mathbf{X}}_{i,k-1}, \tilde{\mathbf{X}}_{i,k+1}, \cdots, \tilde{\mathbf{X}}_{iK}\right)'$.

2.2 Causal interpretation

Let $Y_i(k)$ denote the potential outcome of unit i when $D_i = k$, then observed outcomes are given by

$$Y_i = Y_i(D_i) = Y_i(0) + \mathbf{X}_i' \tau_i$$

where τ_i is the vector of treatment effects, $\tau_{ik} = Y_i(k) - Y_i(0)$.

- **conditional ATE**: The conditional expectation of the vector of treatment effects given by the controls is

  $$\boldsymbol{\tau}(\mathbf{W}_i) = \mathbb{E}\left[\tau_i \mid \mathbf{W_i}\right]$$

  and $\tau_k(\mathbf{W}_i)$ is the **conditional ATE** of the kth treatment.
- **propensity scores**: let

  $$\mathbf{p}(\mathbf{W}_i) = \mathbb{E}\left[\mathbf{X}_i \mid \mathbf{W_i}\right]$$

  denote the vector of propensity scores, with components being $\mathbf{p}_k(\mathbf{W}_i) = \Pr\left(D_i = k \mid \mathbf{W}_i\right)$

2 assumptions

A1 **mean-independence** of the potential outcomes and treatment, conditional on the controls:

$$\mathbb{E}\left[Y_i(k) \mid D_i, \mathbf{W}_i\right] = \mathbb{E}\left[Y_i(k) \mid \mathbf{W}_i\right], \forall k$$

a *sufficient* condition for this assumption is the conditional independence of treatment and potential outcomes (e.g., random assignment)

$$(Y_i(0), \cdots, Y_i(K)) \perp D_i \mid \mathbf{W}_i$$

A2 **propensity scores** must be captured by the covariate adjustment function family $\mathcal{G}$:

$$\mathbf{p}_k\left(w\right) = \mathbb{E}\left[\mathbf{X}_{ik} \mid \mathbf{W}_i = w\right] \in \mathcal{G}, \forall k \tag{11}$$

$$\mu_0\left(w\right) = \mathbb{E}\left[Y_i(0) \mid \mathbf{W}_i = w\right] \in \mathcal{G} \tag{12}$$

one of Eq. (11) and Eq. (12) must be satisfied, then OVB can be avoided.

Two simple cases Under 2 scenarios, there would be no bias

- **homogeneous** (constant within treatment arm k) **treatment effect**: Consider $\tau_{ik} = \tau_k$ for all k, then

$$Y_i = Y_i(0) + \mathbf{X}_i'\boldsymbol{\tau}$$

where $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_k)'$, the coefficient vector identifies $\boldsymbol{\tau}$ is

$$\begin{aligned}
\boldsymbol{\beta} = \mathbb{E}\left[\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i'\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{X}}_iY_i\right] &= \mathbb{E}\left[\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i'\right]^{-1}\left(\mathbb{E}\left[\tilde{\mathbf{X}}_iY_i(0)\right] + \mathbb{E}\left[\tilde{\mathbf{X}}_i\mathbf{X}_i'\right]\boldsymbol{\tau}\right)\\
&= \mathbb{E}\left[\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i'\right]^{-1}\left(\mathbb{E}\left[\tilde{\mathbf{X}}_iY_i(0)\right] + \mathbb{E}\left[\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i'\right]\boldsymbol{\tau}\right)\\
&= \mathbb{E}\left[\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i'\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{X}}_iY_i(0)\right] + \boldsymbol{\tau} = \mathbb{E}\left[\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i'\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{X}}_iY_i(0)\right] + \boldsymbol{\tau}
\end{aligned}$$

by the Law of Iterated Expectations, $\mathbb{E}\left[\tilde{\mathbf{X}}Y_i(0)\right] = \mathbb{E}\left[\tilde{\mathbf{X}}_i\mathbb{E}\left[Y_i(0) \mid \mathbf{W}_i\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\tilde{\mathbf{X}}_i \mid \mathbf{W}_i\right]\mathbb{E}\left[Y_i(0) \mid \mathbf{W}_i\right]\right]$, then

  - under Eq. (11), $\mathbb{E}\left[\tilde{\mathbf{X}}_i \mid \mathbf{W}_i\right] = 0$
  - under Eq. (12), $\mathbb{E}\left[\tilde{\mathbf{X}}_i\mathbb{E}\left[Y_i(0) \mid \mathbf{W}_i\right]\right] = \mathbb{E}\left[\tilde{\mathbf{X}}_i\mu_0(\mathbf{W}_i)\right] = 0$

  no omitted variable bias.

- **heterogeneous** treatment effects, **single** treatment arm: the coefficient is

$$\boldsymbol{\beta} = \mathbb{E}\left[\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i'\right]^{-1}\mathbf{E}[\tilde{\mathbf{X}}_iY_i] = \frac{\mathbb{E}\left[\tilde{\mathbf{X}}_i\mathbf{X}_i\tau_i\right]}{\mathbb{E}\left[\tilde{\mathbf{X}}_i^2\right]} = \mathbb{E}\left[\frac{\mathbb{E}\left[\tilde{\mathbf{X}}_i\mathbf{X}_i \mid \mathbf{W}_i\right]}{\mathbb{E}\left[\tilde{\mathbf{X}}_i\mathbf{X}_i\right]} \cdot \tau(\mathbf{W}_i)\right]$$

  $\boldsymbol{\beta}$ is just a weighted average of heterogeneous treatment effects with weight $\lambda_{11}(\mathbf{W}_i) = \frac{\mathbb{E}[\tilde{\mathbf{X}}_i\mathbf{X}_i|\mathbf{W}_i]}{\mathbb{E}[\tilde{\mathbf{X}}_i\mathbf{X}_i]}$. And,

  - under Eq. (11), $\mathbb{E}\left[\tilde{\mathbf{X}}_i\mathbf{X}_i \mid \mathbf{W}_i\right] = \mathbb{E}\left[\tilde{\mathbf{X}}_i^2 \mid \mathbf{W}_i\right] = \mathrm{Var}\left(\mathbf{X}_i \mid \mathbf{W}_i\right)$, the weight is **non-negative**
  - under Eq. (12) but *not* Eq. (11), negative weights could arise.

Contamination bias In the general model (8), contamination bias arises

Proposition 2.1: Contamination Bias in the Partial Linear Model (8)

Under Assumption 1 and 2, the treatment coefficients in the partial linear model (8) identify

$$\boldsymbol{\beta}_k = \mathbb{E}\left[\lambda_{kk}(\mathbf{W}_i)\tau_k(\mathbf{W}_i)\right] + \sum_{l \neq k} \mathbb{E}\left[\lambda_{kl}(\mathbf{W}_i)\tau_l(\mathbf{W}_i)\right] \tag{13}$$

where the weights

$$\lambda_{kk}(\mathbf{W}_i) = \frac{\mathbb{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik}\mathbf{X}_{ik} \mid \mathbf{W}_k\right]}{\mathbf{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik}^2\right]} \qquad \lambda_{kl}(\mathbf{W}_i) = \frac{\mathbb{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik}\mathbf{X}_{il} \mid \mathbf{W}_k\right]}{\mathbf{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik}^2\right]}$$

satisfy that

$$\mathbb{E}\left[\lambda_{kk}(\mathbf{W}_i)\right] = 1 \qquad \mathbb{E}\left[\lambda_{kl}(\mathbf{W}_i)\right] = 0$$

and under Eq (11), $\lambda_{kk}(\mathbf{W}_i) \geq 0, \forall k$.

here, the coefficient is decomposed into 2 terms:

- a weighted average of conditional ATEs: $\mathbb{E}\left[\lambda_{kk}(\mathbf{W}_i)\tau_k(\mathbf{W}_i)\right]$
  - It generalizes the coefficient in the single treatment case
  - The weights $\lambda_{kk}\left(\mathbf{W}_i\right)$ average to 1, and convex under Eq. (11)
- **Contamination bias**: a weighted average of treatment effects of *other* treatments $\tau_l(\mathbf{W}_i)$

When there is no contamination bias The contamination bias persists generally, with several exceptions:

- $\lambda_{kl}(\mathbf{W}_i) = \mathbb{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik}\mathbf{X}_{il} \mid \mathbf{W}_k\right] / \mathbf{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik}^2\right] = 0$ almost surely, $\forall l \neq k$. Consider

  $$\mathbb{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik} \mid \mathbf{X}_{i,-k}, \mathbf{W}_i\right] = 0$$

  or equivalently,

  $$\mathbb{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik} \mid \mathbf{X}_{i,-k}, \mathbf{W}_i\right] = \mathbf{X}'_{i,-k}\boldsymbol{\alpha} + g_k(\mathbf{W}_i)$$

  However, when the treatment arms are mutually exclusive, that is, $\mathbf{X}_{ik} = 0$ if the unit is assigned to one of the other treatments regardless of $\mathbf{W}_i$, hence it's always true that

  $$\alpha_l = -g_k(\mathbf{W}_i)$$

  for all elements α_l of $\boldsymbol{\alpha}$, implying that the assignment of treatment does not depend on $\mathbf{W}_i$, which is generally not ture unless the propensity score $p_k(\mathbf{W}_i)$ is **constant**.
- the conditional effects of the other treatments are **homogeneous** s.t. $\tau_l(\mathbf{W}_i) = \tau_l$
- the weights $\lambda_{kl}(\mathbf{W}_i)$ and conditional ATEs $\tau_l(\mathbf{W}_i)$ are **uncorrelated** with each other
  - more generally, contamination bias is less concerning when $\lambda_{kl}(\mathbf{W}_i)$ and $\tau_l(\mathbf{W}_i)$ are weakly correlated: that is, the factors influencing treatment effect heterogeneity are largely **unrelated** to the factors influencing the treatment assignment process.

Some remarks on the general characterization of contamination bias

R1 in the multiple treatment case, contamination bias generally arises regardless of Assumption 2, when one use an **additive** covariate adjustment, even the covariate specifications are flexible.

R2 the contamination bias can be **bounded** by the identified contamination weights $\lambda_{kl}(\mathbf{W}_i)$ and the heterogeneity in conditional ATEs $\tau_l(\mathbf{W}_i)$

R3 **negative weighting**: when the treatments are ***not*** mutually exclusive, $\lambda_{kk}(\mathbf{W}_i)$ may still be negative under Eq. (11).

Note: for the two-way FE regressions, none of the recent alternative specifications for g[3] are flexible enough to capture the degenerate propensity scores, hence Eq. (11) in general won't hold.

R4 **IV**: following Prop.2.1,

- the first-stage coefficients on the instruments $\boldsymbol{\beta}_k$ will generally **not** be convex weighted average of the true first-stage effects τ_{ik}, hence, monotonicity condition might not hold
- this new monotonicity concern is especially important in ***judge IV*** designs (conditional random assignment of decision-makers and leave-one-out leniency measure)

R5 descriptive (instead of casual) regressions also suffer from contamination bias.

3 Bias-Aware Estimation of ATE

A simple implementation is **expanding** the partial linear model to include treatment interaction terms. Consider

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + q_0(\mathbf{W}_i) + \sum_{k=1}^{K} X_{ik}\left(q_k(\mathbf{W}_i) - \mathbb{E}\left[q_k(\mathbf{W}_i)\right]\right) + \dot{U}_i \tag{14}$$

where $q_k \in \mathcal{G}$, $\mathcal{G}$ consists of linear functions, $k = 0, \cdots, K$, and again as in Eq. (9),

$$(\boldsymbol{\beta}, q_k) = \arg\min_{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^K, \tilde{q}_k \in \mathcal{G}} \mathbb{E}\left[\dot{U}_i^2\right]$$

Define $\mu_k(w) = \mathbb{E}\left[Y_i(k) \mid \mathbf{W}_i = w\right]$ for $k = 0, \cdots, K$, s.t. $\tau_k(w) = \mu_k(w) - \mu_0(w)$. Under Assumption 1 with rich enough $\mathcal{G}$, then

$$\begin{aligned} \boldsymbol{\beta} &= \boldsymbol{\tau} && \text{unconditional ATEs} \\ q_k(w) &= \tau_k(w) && \text{conditional ATEs} \end{aligned}$$

Issues of weak overlapping the proposed estimator

- achieves *semiparametric efficiency* bound under **strong overlap**, that is, when the propensity score is **bounded away** from 0 and 1
- may be *imprecise* and with performing poorly in *finite sample*.

And, weak overlapping tend to be **more severe** with multiple treatments: the more treatment arms are added, the closer to 0 some propensity scores become. Hence, consider the following estimation of **weighted** averages of conditional ATEs downweighting these counterfactuals with extreme propensity scores.

[3] For example, linear trends, interacted FEs, or other extensions of the basic parallel trends model.

3.1 Efficient Weight Averages of Treatment Effects

Consider a weighted average of conditional potential outcome contrasts

$$\frac{\sum_{i=1}^{N} \lambda(\mathbf{W}_i) \sum_{k=0}^{K} c_k \mu_k(\mathbf{W}_i)}{\sum_{i=1}^{N} \lambda(\mathbf{W}_i)}$$

where $\mu_k(\mathbf{W}_i) = \mathbb{E}[Y_i(k) \mid \mathbf{W}_i]$, $\mathbf{c}$ is a $K+1$ dimension contrast vector with elements c_k, and $\lambda(\mathbf{W}_i)$ is some weighting scheme.

Choosing contrast vector c Stemming from this, two alternative estimations to the one based on Eq. (14) can be established:

i **One-at-a-time** Comparisons: **separately** estimating the effect of each treatment k: set $c_k = 1, c_0 = -1$ and other elements of $\mathbf{c}$ as 0, which leads to

$$\frac{\sum_{i=1}^{N} \lambda(\mathbf{W}_i) \sum_{k=0}^{K} c_k \mu_k(\mathbf{W}_i)}{\sum_{i=1}^{N} \lambda(\mathbf{W}_i)} = \frac{\sum_{i=1}^{N} \lambda(\mathbf{W}_i) \sum_{k=0}^{K} \tau_k(\mathbf{W}_i)}{\sum_{i=1}^{N} \lambda(\mathbf{W}_i)}$$

ii **Simultaneous** Comparisons across **all** treatment arms for all $K(K+1)$ contrasts, that is, weighted averages $\mu_j(\mathbf{W}_i) - \mu_k(\mathbf{W}_i)$ for all $j \neq k$, $j, k = 0, \cdots, K$: set $c_j = 1$ with probability $1/(K+1)$ and $= -1$ with probability $1/(K+1)$

Choosing weight scheme $\lambda(\mathbf{W}_i)$ Given the contrast vector, the weighting scheme $\lambda(\mathbf{W}_i)$ should lead to the **smallest** possible **standard errors** (easiest-to-estimate):

- for **robustness**
- for an **upper bound of the information** available in the data: if the weighting scheme yields small SEs when the SEs for the unweighted ATE are large, it can be concluded that the data is informative about some treatment effects even if it is not about the unweighted average
- for an **intermediate** point along a robustness-efficiency *possibility frontier*
  - Eq (14) gives the **most robust** to treatment effect heterogeneity
  - Eq (8) gives the **most efficient** estimation, while suffering from contamination bias

3.2 Easiest-to-estimate Weighting Scheme

The easiest-to-estimate weighting scheme for multiple treatments is derived in 2 steps:

Step 1: Efficiency Benchmark for A Given Weighted Average Under Assumption 1, an i.i.d. sample of size N, with **known**, **degenerate** propensity scores $p_k(\mathbf{W}_i)$, let $\sigma_k^2(\mathbf{X}_i) = \mathrm{Var}(Y_i(k) \mid \mathbf{W}_i)$, consider estimating the weighted average of contrasts

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

with **known** weighting function λ and contrasts $\mathbf{c}$. Assume $\mathbb{E}[\lambda(\mathbf{W}_i)] \neq 0$, second moments of $\lambda(\mathbf{W}_i)$, $\mu(\mathbf{W}_i)$ are bounded, then we have

Proposition 3.1: Easiest-to-Estimate Weighting Scheme: Step 1

conditional on the controls $\mathbf{W}_i$, the semiparametric efficiency bound is almost surely given by

$$\mathcal{V}_{\lambda,\mathbf{c}} = \frac{1}{\mathbb{E}\left[\lambda(\mathbf{W}_i)\right]^2}\mathbb{E}\left[\sum_{k=0}^{K} \frac{\lambda(\mathbf{W}_i)^2 c_k^2 \sigma_k^2(\mathbf{W}_i)}{p_k(\mathbf{W}_i)}\right] \tag{15}$$

this proposition establishes a lower bound on the asymptotic variance of any regular estimator of $\theta_{\lambda,\mathbf{c}}$ under known propensity scores.

Step 2: Minimizing the Efficiency Bound over Weighting Schemes In Step 2, choose λ to minimize Eq. (15), which gives

$$\lambda^*_{\mathbf{c}}(\mathbf{W}_i) = \left(\sum_{k=0}^{K} \frac{c_k^2 \sigma_k^2(\mathbf{W}_i)}{p_k(\mathbf{W}_i)}\right)^{-1} \geq 0 \tag{16}$$

It is **non-negative**, and the asymptotic variance of the easiest-to-estiate weighting is

$$\mathcal{V}_{\lambda^*_c,\mathbf{c}} = \mathbb{E}\left[\left(\sum_{k=0}^{K} \frac{c_k^2 \sigma_k^2(\mathbf{W}_i)}{p_k(\mathbf{W}_i)}\right)^{-1}\right]^{-1}$$

which is just the **harmonic** mean of $\sum_{k=0}^{K} \frac{c_k^2 \sigma_k^2(\mathbf{W}_i)}{p_k(\mathbf{W}_i)}$, in contrast to the efficient bound for the unweighted contract

$$\mathbb{E}\left[\left(\sum_{k=0}^{K} \frac{c_k^2 \sigma_k^2(\mathbf{W}_i)}{p_k(\mathbf{W}_i)}\right)\right]$$

which is the **arithmetic** mean of $\sum_{k=0}^{K} \frac{c_k^2 \sigma_k^2(\mathbf{W}_i)}{p_k(\mathbf{W}_i)}$.

3.2.1 One-at-a-time Comparisons

Corollary 3.2: Optimal Weights: One-at-a-time Comparisons

For some $k \geq 1$, let $\mathbf{c}^k$ be a vector with elements $c_j^k = 1$ if $j = k$, $c_j^k = -1$ if $j = 0$ and $c_j^k = 0$ otherwise. Suppose the conditional variance of relevant potential outcomes is homoskedastic $\sigma_k^2(\mathbf{W}_i) = \sigma_0^2(\mathbf{W}_i) = \sigma^2$, then following the 2 steps, variance-minizing weighting scheme is given by $\lambda^*_{\mathbf{c}^k} = \lambda^k$, where

$$\lambda^k(\mathbf{W}_i) = \frac{p_0(\mathbf{W}_i)p_k(\mathbf{W}_i)}{p_0(\mathbf{W}_i) + p_k(\mathbf{W}_i)} \tag{17}$$

with the semiparametric efficiency bound given by

$$\mathcal{V}_{\lambda^k,\mathbf{c}^k} = \sigma^2\mathbb{E}\left[\frac{p_0(\mathbf{W}_i)p_k(\mathbf{W}_i)}{p_0(\mathbf{W}_i) + p_k(\mathbf{W}_i)}\right]^{-1} \tag{18}$$

where $p_0(\mathbf{W}_i) = \Pr(D_i = 0 \mid \mathbf{W}_i) = 1 - \sum_{k=1}^{K} p_k(\mathbf{W}_i)$.

notice that when fit the partial linear model (8) on the subsample $D_i \in \{0, k\}$ (control and treatment arm k), the propensity score is given by

$$\Pr\left(D_i = k \mid \mathbf{W}_i, D_i \in \{0, k\}\right) = \frac{p_k(\mathbf{W}_i)}{p_0(\mathbf{W}_i) + p_k(\mathbf{W}_i)}$$

hence, the partial linear model with an additive covariate adjustment can be used to estimate the effect of any given treatment k, provided that g is sufficiently flexible.

Remarks the one-at-a-time regressions is

- **easy to implement**: it does not require explicity estimating the propensity score
- **causal interpretation**: the regression coefficients are causally interpretable as weighted average of conditional treatment effects $\tau_k(\mathbf{W}_i)$ as long as $p_k/(p_0 + p_k) \in \mathcal{G}$
- **treatment-specific**: the weight λ^k is treatment specific, hence cannot be compared horizontally. When the control group is *arbitrarily* chosen, this issue would be more salient.

3.2.2 Simultaneous Comparisons

Consider estimates of a vector $\boldsymbol{\beta}_{lambda^C}$ of K coefficients, with elements

$$\beta_{\lambda^C,k} = \frac{\sum_{i=1}^N \lambda^C(\mathbf{W}_i)\tau_k(\mathbf{W}_i)}{\sum_{i=1}^N \lambda^C(\mathbf{W}_i)}$$

where the weights λ^C are common across all treatment arms. The optimal weight minimizes Eq. (15) with $c_k^2 = 2/(K+1)$, that is

$$\mathcal{V}_{\lambda,\mathbf{c}} = \frac{1}{\mathbb{E}\left[\lambda(\mathbf{W}_i)\right]^2}\mathbb{E}\left[\sum_{k=0}^{K} \frac{\lambda(\mathbf{W}_i)^2 \left(\frac{2}{K+1}\right)^2 \sigma_k^2(\mathbf{W}_i)}{p_k(\mathbf{W}_i)}\right]$$

which gives

Corollary 3.3: Optimal Weights: Simultaneous Comparisons

Let F denote the uniform distribution over the possible contrast vectors, suppose that $\sigma_k^2(\mathbf{W}_i) = \sigma^2$ for all k. Then the weight scheme minimizing $\int \mathcal{V}_{\lambda,\mathbf{c}} \mathrm{d}F(\mathbf{c})$ is given by

$$\lambda^C(\mathbf{W}_i) = \frac{1}{\sum_{k=0}^K p_k(\mathbf{W}_i)^{-1}} \tag{19}$$

The optimal weights λ^C captures the intuition that

- **higher** weights one covariate strata where the treatmetns are **evenly distributed**
- **lower** weights one covariate strata where the treatmetns are **weakly overlapping**

again, these weights are non-negative.

3.3 Estimating Weighted Average Effects with Unknown Propensity Scores

If the propensity scores $p(\mathbf{W}_i)$ are known, $\boldsymbol{\beta}_{\lambda^C}$ can be estimated by a weighted regression of Y_i onto $\mathbf{X}_i$ and a constant, with each observation weighted by $\lambda^C(\mathbf{W}_i)/p_{D_i}(\mathbf{W}_i)$, as shown above. However, since the propensity scores are unknown, we instead use an feasible estimation of the weights

$$\hat{\lambda}^C(\mathbf{W}_i)/\hat{p}_{D_i}(\mathbf{W}_i)$$

where $\hat{p}_k(\mathbf{W}_i)$ is a feasible estimate of the propensity score, $\hat{\lambda}^C(\mathbf{W}_i) = \frac{1}{\sum_k^K=0 \frac{1}{\hat{p}_k(\mathbf{W}_i)}}$. When $\mathcal{G}$ is finite-dimensional, we may run

$$\hat{p}_k(\mathbf{W}_i) = \arg\min_{\tilde{p}\in\mathcal{G}} \sum_{i=1}^{N} \left(X_{ik} - \tilde{p}(\mathbf{W}_i)\right)^2$$

and the resulting estimator is

$$\hat{\beta}_{\hat{\lambda}^C,k} = \frac{\sum_{i=1}^N \frac{\hat{\lambda}^C(\mathbf{W}_i)}{\hat{p}_k(\mathbf{W}_i)} X_{ik} Y_i}{\sum_{i=1}^N \frac{\hat{\lambda}^C(\mathbf{W}_i)}{\hat{p}_k(\mathbf{W}_i)} X_{ik}} - \frac{\sum_{i=1}^N \frac{\hat{\lambda}^C(\mathbf{W}_i)}{\hat{p}_0(\mathbf{W}_i)} X_{i0} Y_i}{\sum_{i=1}^N \frac{\hat{\lambda}^C(\mathbf{W}_i)}{\hat{p}_0(\mathbf{W}_i)} X_{i0}} \tag{20}$$

and the estimator $\hat{\boldsymbol{\beta}}_{\hat{\lambda}^C}$ is efficient that it achieves the semiparametric efficiency bound

Proposition 3.4: Efficiency of Estimator $\hat{\boldsymbol{\beta}}_{\hat{\lambda}^C}$

Suppose Assumption 1 holds in an i.i.d. sample of size N, with known non-degenerate propensity scores $p_k(\mathbf{W}_i)$. Let

$$\beta^*_{\lambda^C,k} = \mathbb{E}\left[\lambda^C(\mathbf{W}_i)\tau_k(\mathbf{W}_i)\right]/\mathbb{E}\left[\lambda^C(\mathbf{W}_i)\right] \qquad \alpha^*_k = \beta^*_{\lambda^C,k} + \mathbb{E}\left[\lambda^C(\mathbf{W}_i)\mu_0(\mathbf{W}_i)\right]/\mathbb{E}\left[\lambda^C(\mathbf{W}_i)\right]$$

and suppose that the fourth moments of $\lambda^C(\mathbf{W}_i)$ and $\mu(\mathbf{W}_i)$ are bounded, and that

$$p_k \in \mathcal{G} \qquad \left(\mu_k(\mathbf{W}_i) - \alpha^*_k\right)\frac{\lambda^C(\mathbf{W}_i)^2}{p_{k'}(\mathbf{W}_i)^2} \in \mathcal{G} \qquad \left(\mu_k(\mathbf{W}_i) - \alpha^*_k\right)\frac{\lambda^C(\mathbf{W}_i)^2}{p_k(\mathbf{W}_i)^2} \in \mathcal{G}, \qquad \forall k, k'$$

Then, provided it is asymptotically linear and regular, $\hat{\boldsymbol{\beta}}_{\hat{\lambda}^C}$ achieves the semiparametric efficiency bound for estimating $\boldsymbol{\beta}_{\lambda^C}$, with **diagonal** elements of its asymptotic variance of

$$\frac{1}{\mathbb{E}\left[\lambda^C(\mathbf{W}_i)\right]^2}\mathbb{E}\left[\frac{\lambda^C(\mathbf{W}_i)^2\sigma_0^2(\mathbf{W}_i)}{p_0(\mathbf{W}_i)} + \frac{\lambda^C(\mathbf{W}_i)^2\sigma_k^2(\mathbf{W}_i)}{p_k(\mathbf{W}_i)} + \lambda^C(\mathbf{W}_i)^2\left(\tau_k(\mathbf{W}_i) - \beta^*_{\lambda^C,k}\right)^2\left(\sum_{k'=0}^{K}\frac{\lambda^C(\mathbf{W}_i)^2}{p_k(\mathbf{W}_i)^3} - 1\right)\right]$$

Remarks this efficiency result

- does **NOT** rely on homoskedasticity, but the weighting $\lambda^C(\mathbf{W}_i)$ might not be optimal under heterogeneity.
- the asymptotic variance of the estimator $\hat{\boldsymbol{\beta}}_{\lambda^C}$ is larger than the infeasible estimator with the infeasible weights (with known propensity scores) $\lambda^C(\mathbf{W}_i)/p_{D_i}(\mathbf{W}_i)$, which achieves the asymptotic variance

$$\frac{1}{\mathbb{E}\left[\lambda^C(\mathbf{W}_i)\right]^2}\mathbb{E}\left[\frac{\lambda^C(\mathbf{W}_i)^2\sigma_0^2(\mathbf{W}_i)}{p_0(\mathbf{W}_i)} + \frac{\lambda^C(\mathbf{W}_i)^2\sigma_k^2(\mathbf{W}_i)}{p_k(\mathbf{W}_i)}\right]$$

- the extra variance term

$$\frac{1}{\mathbb{E}\left[\lambda^C(\mathbf{W}_i)\right]^2}\mathbb{E}\left[\lambda^C(\mathbf{W}_i)^2\left(\tau_k(\mathbf{W}_i)-\beta^*_{\lambda^C,k}\right)^2\left(\sum_{k'=0}^{K}\frac{\lambda^C(\mathbf{W}_i)^2}{p_k(\mathbf{W}_i)^3}-1\right)\right]$$

  reflects the cost of estimating the weights.

4 In Practice: Applying the Bias-Aware Estimations

Here, assume Assumption 1 and 2 both hold, s.t. all propensity scores p_k and potential outcomes conditional expectation functions μ_k are linearly spanned by the controls $\mathbf{W}_i$, consider the OLS estimator $\hat{\boldsymbol{\beta}}$ for the *uninteracted* regression

$$Y_i = \alpha + \sum_{k=1}^{K} X_{ik}\beta_k + \mathbf{W}_i'\gamma + U_i \tag{21}$$

4.1 Contamination Bias Weights

Under Prop. (2.1), we have the own-treatment and contamination bias weights as

$$\lambda_{kk}(\mathbf{W}_i) = \frac{\mathbb{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik}\mathbf{X}_{ik} \mid \mathbf{W}_k\right]}{\mathbf{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik}^2\right]} \qquad \lambda_{kl}(\mathbf{W}_i) = \frac{\mathbb{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik}\mathbf{X}_{il} \mid \mathbf{W}_k\right]}{\mathbf{E}\left[\tilde{\tilde{\mathbf{X}}}_{ik}^2\right]}$$

which can be estimate by the sample analog

$$\hat{\Lambda}_i = \left(\dot{\mathbf{X}}'\dot{\mathbf{X}}\right)^{-1}\dot{\mathbf{X}}_i\dot{\mathbf{X}}_i'$$

where $\dot{\mathbf{X}}_i$ is the sample residual from running OLS of $\dot{\mathbf{X}}_i$ on $\mathbf{W}_i$ and a constant, $\dot{\mathbf{X}}$ is a matrix collecting these sample residuals.

4.2 Estimating ATE

Assuming linearity, the kth conditional ATEs may be written as

$$\tau_k(\mathbf{W}_i) = \gamma_{0,k} + \mathbf{W}_i'\gamma_{\mathbf{W},k}$$

where $\gamma_{0,k}$ and $\gamma_{\mathbf{W},k}$ are coefficients in the interacted regression:

$$Y_i = \alpha_0 + \sum_{k=1}^{K} X_{ik}\gamma_{0,k} + \mathbf{W}_i'\alpha_{\mathbf{W},0} + \sum_{k=1}^{K} X_{ik}\mathbf{W}_i'\gamma_{\mathbf{W},k} + \dot{U}_i \tag{22}$$

OLS estimation gives the estimation $\hat{\tau}_k(\mathbf{W}_i) = \hat{\gamma}_{0,k} + \mathbf{W}_i'\hat{\gamma}_{\mathbf{W},k}$, or in a $K \times 1$ vector form, $\hat{\boldsymbol{\tau}}(\mathbf{W}_i)$.

Under Prop. (2.1)

$$\boldsymbol{\beta}_k = \mathbb{E}\left[\lambda_{kk}(\mathbf{W}_i)\tau_k(\mathbf{W}_i)\right] + \sum_{l\neq k}\mathbb{E}\left[\lambda_{kl}(\mathbf{W}_i)\tau_l(\mathbf{W}_i)\right]$$

Plug-in OLS estimations, get its sample analog

$$\hat{\boldsymbol{\beta}} = \underbrace{\sum_{i=1}^{N} \operatorname{diag}(\hat{\Lambda}_i)\hat{\boldsymbol{\tau}}(\mathbf{W}_i)}_{\text{own-treatment effect}} + \underbrace{\sum_{i=1}^{N} \left[\hat{\Lambda}_i - \operatorname{diag}(\hat{\Lambda}_i)\right] \hat{\boldsymbol{\tau}}(\mathbf{W}_i)}_{\text{contamination bias}} \tag{23}$$

and the regression weighting Λ_i can be adapted for other purposes.

Remark If we plot the estimated contamination weights

$$\hat{\lambda}_{kl}(\mathbf{w}) = \frac{\sum_{i=1}^{N} \mathbf{1}\{\mathbf{W}_i = \mathbf{w}\} \hat{\Lambda}_{i,kl}}{\sum_{i=1}^{N} \mathbf{1}\{\mathbf{W}_i = \mathbf{w}\}}, k \neq l$$

against the treatment effect estimates $\hat{\tau}_l(\mathbf{W}_i)$, we can see the sources of contamination bias.

4.3 Estimating Bias-Aware ATEs

Under linearity assumptions, several solutions can be adopted

1. Unweighted ATEs estimating

$$Y_i = \alpha_0 + \sum_{k=1}^{K} X_{ik}\tau_k + \mathbf{W}_i'\alpha_{\mathbf{W},0} + \sum_{k=1}^{K} X_{ik} \left(\mathbf{W} - \overline{\mathbf{W}}\right)' \gamma_{\mathbf{W},k} + \dot{U}_i \tag{24}$$

where $\overline{\mathbf{W}} = \frac{1}{N}\sum_i \mathbf{W}_i$ is the sample average of the covariate vector. OLS estimates give the unweighted ATEs $\tau_k = \mathbb{E}[\tau_k(\mathbf{W}_i)]$, which is equivalent to $\hat{\tau}_k = \hat{\gamma}_{0,k} + \overline{\mathbf{W}}'\hat{\gamma}_{\mathbf{W},k}$ with estimations $\hat{\gamma}_{0,k}, \hat{\gamma}_{\mathbf{W},k}$ from Eq (22)

2. weighted ATEs: one-at-a-time comparisons estimating

$$Y_i = \ddot{\alpha}_k + X_{ik}\ddot{\beta}_k + \mathbf{W}_i'\ddot{\gamma}_k + \ddot{U}_{ik} \tag{25}$$

for each of the treatments $k = 1, \cdots, K$, for observations assigned either to treatment k or the control group: $D_i \in \{0, k\}$

3. weighted ATEs: simultaneous comparisons the common weights can be estimated as

$$\hat{\lambda}^C(\mathbf{W}_i) = \left(\sum_{k=0}^{K} \hat{p}_k(\mathbf{W}_i)^{-1}\right)^{-1} \tag{26}$$

where the estimated propensity scores $\hat{p}_k(\mathbf{W}_i) = X_{ik} - \dot{X}_{ik}$. Then regress Y_i on X_i, weighting each observation by $\hat{\lambda}^C(\mathbf{W}_i)/\hat{p}_{D_i}(\mathbf{W}_i)$.

Remarks

- Method 2 and 3 yield more precise estimates than Method 1 does.
- Method 2 and 3 change the estimand to a different convex average of conditional treatment effects: covariate values $\mathbf{w}$ where $p_k(\mathbf{w})$ is close to 0 for some k will be effectively dropped.
- If the conditional treatment effects $\tau(\mathbf{W}_i)$ are approximately **independent** of the propensity scores $p(\mathbf{W}_i)$, the weighting scheme might have little effect, the contamination bias would also be small.

References

Joshua D Angrist. Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2):249–288, 1998.

Paul Goldsmith-Pinkham, Peter Hull, and Michal Kolesár. Contamination bias in linear regressions. Technical report, National Bureau of Economic Research, 2022.