July 19, 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]:

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

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: $D_i \in \{1, 2\}$, 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 $X_{i2} \neq X_{i1} - \mathbb{E}[X_{i1} \mid W_i, X_{i2}]$, 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

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

  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, leadign to the *contamination weight* $\lambda_{12}(W_i)$ non-zero.

A numerical example

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.