CONTAMINATION BIAS IN LINEAR REGRESSIONS

Paul Goldsmith-Pinkham, Peter Hull, Michal Kolesár

Joseph Paul

April, 2024

R&R @ American Economic Review

KEY TAKEAWAYS

- Regressions with multiple treatments and flexible controls generally fail to estimate a convex weighted average of heterogeneous treatment effects
- This is due to the coefficients of each treatment being biased or ‘contaminated’ by the effect of the other treatments.
- Not adjusting for this contamination bias results in economically meaningful bias in studies with multiple treatments.

KEY TAKEAWAYS

- Regressions with multiple treatments and flexible controls generally fail to estimate a convex weighted average of heterogeneous treatment effects
- This is due to the coefficients of each treatment being biased or ‘contaminated’ by the effect of the other treatments.
- Not adjusting for this contamination bias results in economically meaningful bias in studies with multiple treatments.

MOTIVATION AND EXAMPLE (PROJECT STAR - KRUEGER (1999))

$$Y_i = \alpha + \beta D_i + \gamma W_i + e_i$$

where

- $D_i \in \{0, 1\}$ - Single Treatment - *Small Class Size*
- $W_i \in \{0, 1\}$ - Single Control (Strata) - *Different Schools*
- e_i - Uncorrelated Residual

Using potential outcome notation, let $Y_i(d)$ denote test scores when $D_i = d$.

- Individual treatment effect $\tau_{1i} = Y_i(1) - Y_i(0)$
- By assumption $(Y_i(0), Y_i(1)) \perp D_i \mid W_i$

MOTIVATION AND EXAMPLE (PROJECT STAR - KRUEGER (1999))

$$Y_i = \alpha + \beta D_i + \gamma W_i + e_i$$

where

- $D_i \in \{0, 1\}$ - Single Treatment - *Small Class Size*
- $W_i \in \{0, 1\}$ - Single Control (Strata) - *Different Schools*
- e_i - Uncorrelated Residual

Using potential outcome notation, let $Y_i(d)$ denote test scores when $D_i = d$.

- Individual treatment effect $\tau_{1i} = Y_i(1) - Y_i(0)$
- By assumption $(Y_i(0), Y_i(1)) \perp D_i \mid W_i$

ANGRIST (1998)

Angrist (1998) showed that the coefficient β identifies a convexly weighted average of within strata ATE:

$$\beta = \phi\tau_1(0) + (1 - \phi)\tau_1(1)$$

where

- $\phi = \frac{var(D_i|W_i{=}0)Pr(W_i{=}0)}{\sum_{w=0}^{1} var(D_i|W_i{=}w)Pr(W_i{=}w)}$
- $\tau_1(w) = \mathbb{E}[Y_i(1) - Y_i(0) \mid W_i = w]$ is the ATE within strata $w \in \{0, 1\}$

Thus β identifies a weighted average of strata effects across the two groups.

DERIVATION USING FRISCH-WAUGH-LOWELL THEOREM

Using FWL theorem, we can write our multivariate regression as a univariate regression:

$$
\begin{aligned}
\text{Let } M_W &= I - W_i(W_i'W_i)^{-1}W_i' \\
M_{W_i}Y_i &= M_{W_i}D_i\beta + M_{W_i}e_i \\
\implies Y_i &= \tilde{D}_i\beta + e_i \\
\tilde{D}_iY &= \tilde{D}_i\tilde{D}_i\beta + \tilde{D}_ie_i \\
\implies \beta &= \frac{\mathbb{E}\tilde{D}_iY_i}{\mathbb{E}\tilde{D}_i^2}
\end{aligned}
$$

DERIVATION (CONT.)

Note that:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

$$\text{and } Y_i(1) = Y_i(0) + \tau$$

Substituting back in:

$$\begin{aligned} Y_i &= D_i(Y_i(0) + \tau_i) + (1 - D_i) Y_i(0) \\ &= D_i Y_i(0) + D_i \tau_i + Y_i(0) - D_i Y_i(0) \\ &= Y_i(0) + D_i \tau_i \end{aligned}$$

DERIVATION (CONT.)

Substituting into β:

$$\beta = \frac{\mathbb{E}[\tilde{D}_i(Y_i(0) + D_i\tau_{1i})]}{\mathbb{E}[\tilde{D}_i^2]}$$

$$= \frac{\mathbb{E}[\tilde{D}_i Y_i(0)]}{\mathbb{E}[\tilde{D}_i^2]} + \frac{\mathbb{E}[\tilde{D}_i D_i \tau_{1i}]}{\mathbb{E}[\tilde{D}_i^2]}$$

Using LIE and conditional random assignment:

$$\mathbb{E}[\tilde{D}_i Y_i(0)] = \mathbb{E}[\mathbb{E}[\tilde{D}_i Y_i(0) \mid W_i]] = \mathbb{E}[\mathbb{E}[\tilde{D}_i \mid W_i]\mathbb{E}[Y_i(0) \mid W_i]] = 0$$

Therefore,

$$\beta = \frac{\mathbb{E}[\tilde{D}_i Y_i(0)]}{\mathbb{E}[\tilde{D}_i^2]} + \frac{\mathbb{E}[\tilde{D}_i D_i \tau_{1i}]}{\mathbb{E}[\tilde{D}_{i2}]} = \frac{\mathbb{E}[\tilde{D}_i D_i \tau_{1i}]}{\mathbb{E}[\tilde{D}_{i2}]}$$

DERIVATION (CONT.)

Substituting into β:

$$\beta = \frac{\mathbb{E}[\tilde{D}_i(Y_i(0) + D_i\tau_{1i})]}{\mathbb{E}[\tilde{D}_i^2]}$$

$$= \frac{\mathbb{E}[\tilde{D}_i Y_i(0)]}{\mathbb{E}[\tilde{D}_i^2]} + \frac{\mathbb{E}[\tilde{D}_i D_i \tau_{1i}]}{\mathbb{E}[\tilde{D}_i^2]}$$

Using LIE and conditional random assignment:

$$\mathbb{E}[\tilde{D}_i Y_i(0)] = \mathbb{E}[\mathbb{E}[\tilde{D}_i Y_i(0) \mid W_i]] = \mathbb{E}[\mathbb{E}[\tilde{D}_i \mid W_i]\mathbb{E}[Y_i(0) \mid W_i]] = 0$$

Therefore,

$$\beta = \frac{\mathbb{E}[\tilde{D}_i Y_i(0)]}{\mathbb{E}[\tilde{D}_i^2]} + \frac{\mathbb{E}[\tilde{D}_i D_i \tau_{1i}]}{\mathbb{E}[\tilde{D}_{i2}]} = \frac{\mathbb{E}[\tilde{D}_i D_i \tau_{1i}]}{\mathbb{E}[\tilde{D}_{i2}]}$$

DERIVATION (CONT.)

Substituting into β:

$$
\begin{aligned}
\beta &= \frac{\mathbb{E}[\tilde{D}_i(Y_i(0) + D_i\tau_{1i})]}{\mathbb{E}[\tilde{D}_i^2]} \\
&= \frac{\mathbb{E}[\tilde{D}_i Y_i(0)]}{\mathbb{E}[\tilde{D}_i^2]} + \frac{\mathbb{E}[\tilde{D}_i D_i \tau_{1i}]}{\mathbb{E}[\tilde{D}_i^2]}
\end{aligned}
$$

Using LIE and conditional random assignment:

$$\mathbb{E}[\tilde{D}_i Y_i(0)] = \mathbb{E}[\mathbb{E}[\tilde{D}_i Y_i(0) \mid W_i]] = \mathbb{E}[\mathbb{E}[\tilde{D}_i \mid W_i]\mathbb{E}[Y_i(0) \mid W_i]] = 0$$

Therefore,

$$\beta = \frac{\mathbb{E}[\tilde{D}_i Y_i(0)]}{\mathbb{E}[\tilde{D}_i^2]} + \frac{\mathbb{E}[\tilde{D}_i D_i \tau_{1i}]}{\mathbb{E}[\tilde{D}_{i^2}]} = \frac{\mathbb{E}[\tilde{D}_i D_i \tau_{1i}]}{\mathbb{E}[\tilde{D}_{i^2}]}$$

DERIVATION (CONT.)

Again, using LIE:

$$\begin{aligned}
\mathbb{E}[\tilde{D}_i D_i \tau_{1i}] &= \mathbb{E}[\mathbb{E}[\tilde{D}_i D_i \tau_{1i} \mid W_i]] \\
&= \mathbb{E}[\mathbb{E}[\tilde{D}_i D_i \mid W_i]\mathbb{E}[\tau_{1i} \mid W_i]] \\
&= \mathbb{E}[var(D_i \mid W_i)\tau_{1i}(W_i)]
\end{aligned}$$

Note that $\mathbb{E}[\tilde{D}_i^2] = \mathbb{E}[\mathbb{E}[\tilde{D}_i^2 \mid W_i]] = \mathbb{E}[var(D_i \mid W)]$, which gives us:

$$\begin{aligned}
\beta = \frac{\mathbb{E}[\tilde{D}_i D_i \tau_{1i}]}{\mathbb{E}[\tilde{D}_i^2]} &= \frac{\mathbb{E}[var(D_i \mid W_i)\tau_{1i}(W_i)]}{\mathbb{E}[var(D_i \mid W)]} \\
&= \phi\tau_1(0) + (1 - \phi)\tau_1(1)
\end{aligned}$$

DERIVATION (CONT.)

Again, using LIE:

$$\begin{aligned}
\mathbb{E}[\tilde{D}_i D_i \tau_{1i}] &= \mathbb{E}[\mathbb{E}[\tilde{D}_i D_i \tau_{1i} \mid W_i]] \\
&= \mathbb{E}[\mathbb{E}[\tilde{D}_i D_i \mid W_i]\mathbb{E}[\tau_{1i} \mid W_i]] \\
&= \mathbb{E}[var(D_i \mid W_i)\tau_{1i}(W_i)]
\end{aligned}$$

Note that $\mathbb{E}[\tilde{D}_i^2] = \mathbb{E}[\mathbb{E}[\tilde{D}_i^2 \mid W_i]] = \mathbb{E}[var(D_i \mid W)]$, which gives us:

$$\begin{aligned}
\beta = \frac{\mathbb{E}[\tilde{D}_i D_i \tau_{1i}]}{\mathbb{E}[\tilde{D}_i^2]} &= \frac{\mathbb{E}[var(D_i \mid W_i)\tau_{1i}(W_i)]}{\mathbb{E}[var(D_i \mid W)]} \\
&= \phi\tau_1(0) + (1 - \phi)\tau_1(1)
\end{aligned}$$

CONTAMINATION BIAS WITH TWO RANDOMISED TREATMENTS

Consider an example with:

- A Control Group: $D_i = 0$
- A treatment that reduces class sizes: $D_i = 1$
- A treatment that introduces full time teaching aids: $D_i = 2$

Let $X_i = (X_{1i}, X_{2i})'$, where $X_{ki} = \mathbb{1}\{D_i = k\}$ indicates assignment to treatments $k = 1, 2$. We include a constant and school indicators W_i.

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma W_i + e_i$$

With potential outcomes: $Y_i = Y_i(0) + \tau_{1i} X_{1i} + \tau_{2i} X_{2i}$

$$\tau_{1i} = Y_i(1) - Y_i(0), \ \tau_{2i} = Y_i(2) - Y_i(0)$$

And again assuming conditional random assignment: $(Y_i(0), Y_i(1), Y_i(2)) \perp X_i \mid W_i$

DERIVATION WITH TWO TREATMENTS USING FWL

Denote $Q_i := (\mathbf{1}, X_{2i}, W_i)$, then:

$$
\begin{aligned}
M_{Q_i} Y_i &= \beta M_{Q_i} X_{1i} + M_{Q_i} e_i \\
\implies \quad Y_i &= \beta \tilde{\tilde{X}}_{1i} + e_i \\
\implies \quad \beta &= \frac{\mathbb{E}[\tilde{\tilde{X}}_{1i} Y_i]}{\mathbb{E}[\tilde{\tilde{X}}_{1i}^2]}
\end{aligned}
$$

Substituting potential outcomes:

$$
\begin{aligned}
\beta &= \frac{\mathbb{E}[\tilde{\tilde{X}}_{1i}(Y_i(0) + \tau_{1i} X_{1i} + \tau_{2i} X_{2i})]}{\mathbb{E}[\tilde{\tilde{X}}_{1i}^2]} \\
&= \frac{\mathbb{E}[\tilde{\tilde{X}}_{1i} Y_i(0)]}{\mathbb{E}[\tilde{\tilde{X}}_{1i}^2]} + \frac{\mathbb{E}[\tilde{\tilde{X}}_{1i} X_{1i} \tau_{1i}]}{\mathbb{E}[\tilde{\tilde{X}}_{1i}^2]} + \frac{\mathbb{E}[\tilde{\tilde{X}}_{1i} X_{2i} \tau_{2i}]}{\mathbb{E}[\tilde{\tilde{X}}_{1i}^2]}
\end{aligned}
$$

KEY DIFFERENCE WITH TWO TREATMENTS

$$\beta = \frac{\mathbb{E}[\tilde{\tilde{X}}_{1i} Y_i(0)]}{\mathbb{E}[\tilde{\tilde{X}}_{1i}^2]} + \frac{\mathbb{E}[\tilde{\tilde{X}}_{1i} X_{1i} \tau_{1i}]}{\mathbb{E}[\tilde{\tilde{X}}_{1i}^2]} + \frac{\mathbb{E}[\tilde{\tilde{X}}_{1i} X_{2i} \tau_{2i}]}{\mathbb{E}[\tilde{\tilde{X}}_{1i}^2]}$$

We still have $\mathbb{E}[\tilde{\tilde{X}}_{1i} Y_i(0)] = 0$ since the auxiliary regression residuals are uncorrelated with potential outcomes.

However, we do not generally have $\mathbb{E}[\tilde{\tilde{X}}_{i1} X_{i2} \tau_{i2}] = 0$.

The key difference here is that $\tilde{\tilde{X}}_{1i}$ is uncorrelated with W_i, X_{2i}, but it is not mean independent.

KEY DIFFERENCE WITH TWO TREATMENTS

$$\beta = \frac{\mathbb{E}[\tilde{\tilde{X}}_{1i} Y_i(0)]}{\mathbb{E}[\tilde{\tilde{X}}_{1i}^2]} + \frac{\mathbb{E}[\tilde{\tilde{X}}_{1i} X_{1i} \tau_{1i}]}{\mathbb{E}[\tilde{\tilde{X}}_{1i}^2]} + \frac{\mathbb{E}[\tilde{\tilde{X}}_{1i} X_{2i} \tau_{2i}]}{\mathbb{E}[\tilde{\tilde{X}}_{1i}^2]}$$

We still have $\mathbb{E}[\tilde{\tilde{X}}_{1i} Y_i(0)] = 0$ since the auxiliary regression residuals are uncorrelated with potential outcomes.

However, we do not generally have $\mathbb{E}[\tilde{\tilde{X}}_{i1} X_{i2} \tau_{i2}] = 0$.

The key difference here is that $\tilde{\tilde{X}}_{1i}$ is uncorrelated with W_i, X_{2i}, but it is not mean independent.

CONTAMINATION BIAS TERM

This is because the dependence between X_{1i} and X_{2i} is non-linear as they are mutually exclusive treatments:

- If $X_{2i} = 1$, then $X_{1i} = 0$
- If $X_{2i} = 0$, then $Pr(X_{1i} = 1)$ depends on W_i

Thus, $\tilde{\tilde{X}}_{1i} \neq X_{1i} - \mathbb{E}[X_{1i} \mid W_i, X_{2i}]$.

Because $\tilde{\tilde{X}}_{i1}$ does not coincide with the conditionally demeaned X_{i1}, we cannot generally reduce the expression to only involve the effects of the first treatment, τ_{i1}.

“CONTAMINATED” ESTIMATE

Instead, β_1 simplifies to:

$$\beta_1 = \mathbb{E}[\lambda_{11}(W_i)\tau_1(W_i)] + \mathbb{E}[\lambda_{12}(W_i)\tau_2(W_i)]$$

where:

- $\lambda_{11}(W_i) = \frac{\mathbb{E}[\tilde{\tilde{X}}_{1i}X_{1i}|W_i]}{\mathbb{E}\tilde{\tilde{X}}_{1i}^2}$ is non-negative and averages to one
- $\lambda_{12}(W_i) = \frac{\mathbb{E}[\tilde{\tilde{X}}_{1i}X_{2i}|W_i]}{E[\tilde{\tilde{X}}_{1i}^2]}$ is the contamination bias term and is generally non-zero

The second term includes conditional effects of the other treatment $\tau_2(W_i) = \mathbb{E}[Y_i(2) - Y_i(0) \mid W_i]$, causing the bias.

UNDERSTANDING THE CONTAMINATION BIAS

The contamination bias term arises because the residualised treatment $\tilde{\tilde{X}}_{1i}$ is not conditionally independent of the second treatment X_{2i} within strata, despite being uncorrelated with X_{2i} by construction.

This can be understood by interpreting $\tilde{\tilde{X}}_{1i}$ as the result of a two-step residualization process:

First, demean treatments within strata:

$$\tilde{X}_{1i} = X_{1i} - \mathbb{E}[X_{1i} \mid W_i] = X_{1i} - p_1(W_i)$$
$$\tilde{X}_{2i} = X_{2i} - \mathbb{E}[X_{2i} \mid W_i] = X_{2i} - p_2(W_i)$$

where $p_j(W_i)$ gives the propensity score for treatment j within strata.

UNDERSTANDING THE CONTAMINATION BIAS

The contamination bias term arises because the residualised treatment $\tilde{\tilde{X}}_{1i}$ is not conditionally independent of the second treatment X_{2i} within strata, despite being uncorrelated with X_{2i} by construction.

This can be understood by interpreting $\tilde{\tilde{X}}_{1i}$ as the result of a two-step residualization process:

First, demean treatments within strata:

$$\tilde{X}_{1i} = X_{1i} - \mathbb{E}[X_{1i} \mid W_i] = X_{1i} - p_1(W_i)$$
$$\tilde{X}_{2i} = X_{2i} - \mathbb{E}[X_{2i} \mid W_i] = X_{2i} - p_2(W_i)$$

where $p_j(W_i)$ gives the propensity score for treatment j within strata.

UNDERSTANDING THE CONTAMINATION BIAS (CONT.)

$$\tilde{X}_{1i} = X_{1i} - \mathbb{E}[X_{1i} \mid W_i] = X_{1i} - p_1(W_i)$$
$$\tilde{X}_{2i} = X_{2i} - \mathbb{E}[X_{2i} \mid W_i] = X_{2i} - p_2(W_i)$$

Second, regress $\tilde{X}_{1i}$ on $\tilde{X}_{2i}$ to generate the residuals $\tilde{\tilde{X}}_{1i}$.

When the propensity scores differ across strata ($p_j(0) \neq p_j(1)$), the relationship between these residuals varies by school, and the line of best-fit averages across this relationship.

As a result, the line of best fit does not isolate the conditional (within strata) variation in X_{1i}: the remaining variation of $\tilde{\tilde{X}}_{1i}$ will tend to predict X_{2i} within schools, making the contamination weight $\lambda_{12}(W_i)$ non-zero.

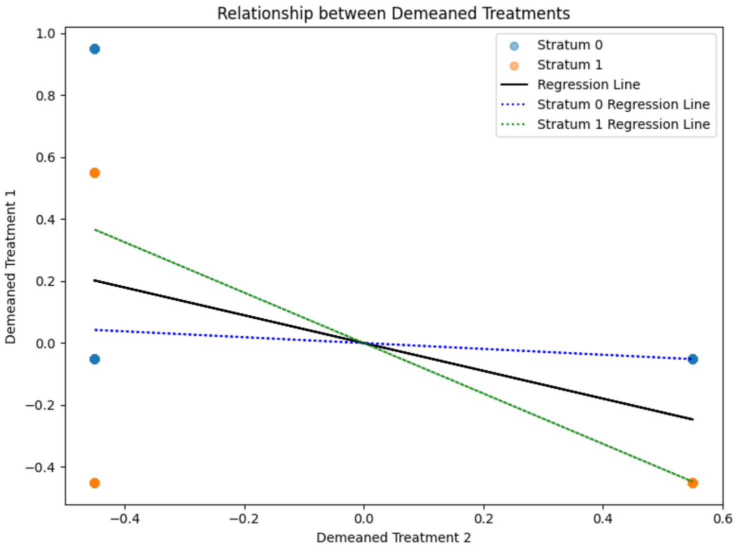
UNDERSTANDING THE CONTAMINATION BIAS (CONT.)

$$\tilde{X}_{1i} = X_{1i} - \mathbb{E}[X_{1i} \mid W_i] = X_{1i} - p_1(W_i)$$
$$\tilde{X}_{2i} = X_{2i} - \mathbb{E}[X_{2i} \mid W_i] = X_{2i} - p_2(W_i)$$

Second, regress $\tilde{X}_{1i}$ on $\tilde{X}_{2i}$ to generate the residuals $\tilde{\tilde{X}}_{1i}$.

When the propensity scores differ across strata ($p_j(0) \neq p_j(1)$), the relationship between these residuals varies by school, and the line of best-fit averages across this relationship.

As a result, the line of best fit does not isolate the conditional (within strata) variation in X_{1i}: the remaining variation of $\tilde{\tilde{X}}_{1i}$ will tend to predict X_{2i} within schools, making the contamination weight $\lambda_{12}(W_i)$ non-zero.



- OLS assumes the black line, such that the variation in X_{i1} after resdualsing linearly for X_{i1} and W_i tends to predict the X_{i2} treatment.
- The treated X_{i1} units are picking up "treated" X_{i2} units

ILLUSTRATION

Suppose that:

- In school 0 ($W_i = 0$):
  - 5% assigned to small classroom treatment ($X_{1i} = 1$)
  - 45% assigned to full-time aid treatment ($X_{2i} = 1$)
  - Remaining assigned to control
- In school 1 ($W_i = 1$):
  - 45% assigned to both treatments
- Schools have the same number of students: $Pr(W_i = 1) = 0.5$

Assume:

- $\tau_1(W_i) = 0$
- $\tau_2(0) = 0$ and $\tau_2(1) = 1$

ILLUSTRATION - CODE

```
# Generate strata indicators
strata = np.random.choice([0, 1], size=n, p=[0.5, 0.5])
group_1 = np.random.choice([0, 1, 2], size=n, p=[0.5, 0.05, 0.45])
group_2 = np.random.choice([0, 1, 2], size=n, p=[0.10, 0.45, 0.45])

D[strata == 0] = group_1[strata == 0]
D[strata == 1] = group_2[strata == 1]

treatment1 = np.where(D == 1, 1, 0)
treatment2 = np.where(D == 2, 1, 0)

# Generate heterogeneous treatment effects
y0 = np.random.normal(0, 1, n)  # Potential outcome under control
y1 = y0                         # Potential outcome under treatment 1
y2_0 = y0                       # Potential outcome under treatment 2 for strata 0
y2_1 = y0 + 1                   # Potential outcome under treatment 2 for strata 1

# Generate observed outcomes
y = y0 + treatment1 * (y1 - y0) + treatment2 * (y2_0 - y0) * (1 - strata) + treatment2 * (y2_1 - y0) * strata +
     np.random.normal(0, 1, n)

# Estimate the treatment effect
X = np.array([np.ones_like(y), treatment1, treatment2, strata]).T
reg(X, y, print=True)
```

ILLUSTRATION - RESULTS

$$\beta_0 = -0.118(\text{SE: } 0.027)$$

!! $\beta_1 = -\mathbf{0.471}(\text{SE: } 0.046)$!!

$$\beta_2 = 0.285(\text{SE: } 0.035)$$

$$\beta_3 = 0.652(\text{ SE: } 0.034)$$

CONTAMINATION BIAS SUMMARY

- Unlike with a binary *D*, the estimates of β_1 and β_2 are not convex estimates of $\tau_1(W_i)$ and $\tau_2(W_i)$, but are contaminated by the other treatment effects.
- Why?
  - Controlling for W_i is analogous to controlling for the propensity score $Pr(D_i = 1 \mid W_i)$
  - But with two treatment, X_{1i} is a function of both the conditioning variables W_i *and* X_{2i}
  - Therefore, the propensity score will *not* be correctly estimated. We are measuring the "overall" propensity score, not within a given stratum.

CONTAMINATION BIAS SUMMARY

- Unlike with a binary D, the estimates of β_1 and β_2 are not convex estimates of $\tau_1(W_i)$ and $\tau_2(W_i)$, but are contaminated by the other treatment effects.
- Why?
  - Controlling for W_i is analogous to controlling for the propensity score $Pr(D_i = 1 \mid W_i)$
  - But with two treatment, X_{1i} is a function of both the conditioning variables W_i *and* X_{2i}
  - Therefore, the propensity score will *not* be correctly estimated. We are measuring the “overall” propensity score, not within a given stratum.

SOLUTION - INTERACTIVE REGRESSION

The interactive regression model takes the form:

$$Y_i = X_i\beta + q_0(W_i) + \sum_{k=1}^{K} X_{ik}\left(q_k(W_i) - \mathbb{E}[q_k(W_i)]\right) + \dot{U}_i$$

where $\beta, \{q_k\}_{k=0}^{K} \in \mathcal{G}^{K+1}$ are minimizers of $\mathbb{E}[\dot{U}_i]$.

For linear functions:

$$Y_i = \alpha_0 + \sum_{k=1}^{K} X_{ik}\tau_k + W_i'\alpha_{W,0} + \sum_{k=1}^{K} X_{ik}(W_i - \bar{W})'\gamma_{W,k} + \dot{U}_i$$

which can be estimated by OLS, where $\bar{W}$ is the sample mean of W_i (Imbens & Wooldridge, 2009). This can be easily extended to basis functions such as polynomials or splines for q_k.

SOLUTION - INTERACTIVE REGRESSION

The interactive regression model takes the form:

$$Y_i = X_i\beta + q_0(W_i) + \sum_{k=1}^{K} X_{ik}\left(q_k(W_i) - \mathbb{E}[q_k(W_i)]\right) + \dot{U}_i$$

where $\beta, \{q_k\}_{k=0}^{K} \in \mathcal{G}^{K+1}$ are minimizers of $\mathbb{E}[\dot{U}_i]$.

For linear functions:

$$Y_i = \alpha_0 + \sum_{k=1}^{K} X_{ik}\tau_k + W_i'\alpha_{W,0} + \sum_{k=1}^{K} X_{ik}(W_i - \bar{W})'\gamma_{W,k} + \dot{U}_i$$

which can be estimated by OLS, where $\bar{W}$ is the sample mean of W_i (Imbens & Wooldridge, 2009). This can be easily extended to basis functions such as polynomials or splines for q_k.

INTERACTIVE REGRESSION - CODE

```
Z = np.concatenate((X, W.T, interactions), axis=1)

reg(Z, y, print=True)

```

$$\beta 0 = -0.005 (SE : 375063.470)$$

$$\beta 1 = \mathbf{0.003} \quad (SE : 0.058)$$

$$\beta 2 = \mathbf{0.480} \quad (SE : 0.040)$$

$$\beta 3 = 0.026 \quad (SE : 761704.854)$$

$$\beta 4 = 0.029 \quad (SE : 761704.854)$$

$$\beta 5 = -0.099 \quad (SE : 0.116)$$

$$\beta 6 = 0.963 \quad (SE : 0.081)$$

DECOMPOSING CONTAMINATION

We can decompose the OLS estimate of $\hat{\beta}$ from the uninteracted regression

$$Y_i = \alpha + \sum_k X_{ik}\beta_k + W_i'\gamma + U_i$$

Contamination bias weights are identified by the linear regression of X_i on the residuals $\tilde{X}_i$. Specifically $\lambda_{kl}(W_i)$ is given by (k, l)th element of

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

where $\dot{X}_i$ is the sample residuals from an OLS regression of X_i on W_i.

DECOMPOSING CONTAMINATION (CONT.)

$\hat{\beta}$ from the uninteracted regression model is equivalent to

$$\hat{\beta} = \sum_{i=1}^{n} \text{diag}(\hat{\Lambda}_i)\hat{\tau}(W_i) + \sum_{i=1}^{n} [\hat{\Lambda}_i - \text{diag}(\hat{\Lambda}_i)]\hat{\tau}(W_i)$$

The first term estimates the own-treatment effect components, while the second term estimates the contamination bias components.

$$\hat{\lambda}_{kl}(w) = \frac{\sum_i \mathbb{1}\{W_i = w\}\hat{\Lambda}_{i,kl}}{\sum_i \mathbb{1}\{W_i = w\}}$$

$$\beta_1 = \mathbb{E}[\lambda_{11}(W_i)\tau_1(W_i)] + \sum_{k=2}^{K} \mathbb{E}[\lambda_{1k}(W_i)\tau_k(W_i)]$$

APPLICATION - PROJECT STAR

- The project STAR as studied by Kruger (1999) randomised $11,600$ students in 79 public school to one of three types of classes:
  1. regular sized (20-25 students)
  2. small (13-17 students)
  3. with extra teaching aid
- Proportion of students randomised to the small class size and teaching aide treatment varied across schools
- Y_i is the average percentile of student i's math, reading and word recognition at the end of kindergarten.

APPLICATION - RESULTS

Column 1 are estimates of kindergarten treatment effects in the uninteracted regression model.

Column 2 are the estimates the own-treatment effect from the above decomposition.

Column 3 are the estimates from the interacted regression model.

A. Treatment effect estimates

	$\hat{\beta}$ (1)	Own (2)	ATE (3)	EW (4)	CW (5)
Small	5.357	5.202	5.561	5.295	5.577
	(0.778)	(0.778)	(0.763)	(0.775)	(0.764)
			[0.744]	[0.743]	[0.742]
Aide	0.177	0.360	0.070	0.263	0.011
	(0.720)	(0.714)	(0.708)	(0.715)	(0.712)
			[0.694]	[0.691]	[0.695]
Number of controls	77				
Sample size	5,868				

B. Contamination bias estimates

		Worst-Case Bias	
	Bias (1)	Negative (2)	Positive (3)
Small class size	0.155	−1.654	1.670
	(0.160)	(0.185)	(0.187)
Teaching aide	−0.183	−1.529	1.530
	(0.149)	(0.176)	(0.177)

Notes: Panel A gives estimates of small class and teaching aide treatment effects for the Project STAR kindergarten analysis. Col. 1 reports estimates from a partially linear model in eq. (21), col. 2 reports the own-treatment component of the decomposition in eq. (23), col. 3 reports the interacted regression estimates based on eq. (17), col. 4 reports estimates based on the EW scheme using one-treatment-at-a-time regressions in eq. (24), and col 5 uses

APPLICATION - RESULTS

Column 1 in panel B reports the contamination bias, which appears minimal.

They show this is due to weak correlation between the contamination weights and the treatment effects.

A. Treatment effect estimates

	$\hat{\beta}$ (1)	Own (2)	ATE (3)	EW (4)	CW (5)
Small	5.357	5.202	5.561	5.295	5.577
	(0.778)	(0.778)	(0.763)	(0.775)	(0.764)
			[0.744]	[0.743]	[0.742]
Aide	0.177	0.360	0.070	0.263	0.011
	(0.720)	(0.714)	(0.708)	(0.715)	(0.712)
			[0.694]	[0.691]	[0.695]
Number of controls	77				
Sample size	5,868				

B. Contamination bias estimates

		Worst-Case Bias	
	Bias (1)	Negative (2)	Positive (3)
Small class size	0.155	−1.654	1.670
	(0.160)	(0.185)	(0.187)
Teaching aide	−0.183	−1.529	1.530
	(0.149)	(0.176)	(0.177)

Notes: Panel A gives estimates of small class and teaching aide treatment effects for the Project STAR kindergarten analysis. Col. 1 reports estimates from a partially linear model in eq. (21), col. 2 reports the own-treatment component of the decomposition in eq. (23), col. 3 reports the interacted regression estimates based on eq. (17), col. 4 reports estimates based on the EW scheme using one-treatment-at-a-time regressions in eq. (24), and col 5 uses

AER?

- Widely Applicable
- Rigorous characterisation of the problem
- Provides useful practical guidance and tools for measuring and avoiding contamination bias

However

- The empirical application isn’t the strongest

AER?

- Widely Applicable
- Rigorous characterisation of the problem
- Provides useful practical guidance and tools for measuring and avoiding contamination bias

However

- The empirical application isn't the strongest