Day 2: On Weights and Clusters

Peter Hull

Design-Based Regression Inference
Fall 2024

Outline

1. Heterogeneous Treatment Effects

2. Clustered Standard Errors

Whose Treatment Effect is it Anyway?

- On Monday we contrasted design vs. outcome-model strategies in a constant-effect world (i.e. with a causal model of $y_i = \beta x_i + \varepsilon_i$)
    - Of course the real world is messier: more realistic is $y_i = \beta_i x_i + \varepsilon_i$ (or more complicated forms of effect heterogeneity)

Whose Treatment Effect is it Anyway?

- On Monday we contrasted design vs. outcome-model strategies in a constant-effect world (i.e. with a causal model of $y_i = \beta x_i + \varepsilon_i$)
  - Of course the real world is messier: more realistic is $y_i = \beta_i x_i + \varepsilon_i$ (or more complicated forms of effect heterogeneity)
  - Can think about what the regression/IV estimand equals in such models

Whose Treatment Effect is it Anyway?

- On Monday we contrasted design vs. outcome-model strategies in a constant-effect world (i.e. with a causal model of $y_i = \beta x_i + \varepsilon_i$)
  - Of course the real world is messier: more realistic is $y_i = \beta_i x_i + \varepsilon_i$ (or more complicated forms of effect heterogeneity)
  - Can think about what the regression/IV estimand equals in such models

- Today we'll see another difference: how design-based vs. model-based regression/IV weigh together heterogeneous effects
  - Bottom line: design avoids recent concerns over "negative weights"...

Whose Treatment Effect is it Anyway?

- On Monday we contrasted design vs. outcome-model strategies in a constant-effect world (i.e. with a causal model of $y_i = \beta x_i + \varepsilon_i$)
  - Of course the real world is messier: more realistic is $y_i = \beta_i x_i + \varepsilon_i$ (or more complicated forms of effect heterogeneity)
  - Can think about what the regression/IV estimand equals in such models
- Today we'll see another difference: how design-based vs. model-based regression/IV weigh together heterogeneous effects
  - Bottom line: design avoids recent concerns over "negative weights"...
  - ... at least as long as you don't have multiple treatments!

Primer 1: Angrist (1998)

- Let $x_i \in \{0,1\}$; general causal model: $y_i = \underbrace{(y_i(1) - y_i(0))}_{\beta_i} x_i + \underbrace{y_i(0)}_{\varepsilon_i}$
  - Design: $x_i \mid w, y(0), y(1) \sim F_x(w_i)$ with linear $E[x_i \mid w_i]$

Primer 1: Angrist (1998)

- Let $x_i \in \{0,1\}$; general causal model: $y_i = \underbrace{(y_i(1) - y_i(0))}_{\beta_i} x_i + \underbrace{y_i(0)}_{\varepsilon_i}$
  - Design: $x_i \mid w, y(0), y(1) \sim F_x(w_i)$ with linear $E[x_i \mid w_i]$

- What does the w_i-controlled regression *actually* identify?

Primer 1: Angrist (1998)

- Let $x_i \in \{0,1\}$; general causal model: $y_i = \underbrace{(y_i(1) - y_i(0))}_{\beta_i} x_i + \underbrace{y_i(0)}_{\varepsilon_i}$
  - Design: $x_i \mid w, y(0), y(1) \sim F_x(w_i)$ with linear $E[x_i \mid w_i]$
- What does the w_i-controlled regression *actually* identify?

$$\beta = \frac{Cov(\tilde{x}_i, y_i)}{Var(\tilde{x}_i)} =$$

Primer 1: Angrist (1998)

- Let $x_i \in \{0,1\}$; general causal model: $y_i = \underbrace{(y_i(1) - y_i(0))}_{\beta_i} x_i + \underbrace{y_i(0)}_{\varepsilon_i}$
  - Design: $x_i \mid w, y(0), y(1) \sim F_x(w_i)$ with linear $E[x_i \mid w_i]$
- What does the w_i-controlled regression *actually* identify?

$$\beta = \frac{Cov(\tilde{x}_i, y_i)}{Var(\tilde{x}_i)} = \frac{Cov(\tilde{x}_i, x_i\beta_i)}{Var(\tilde{x}_i)} + \frac{Cov(\tilde{x}_i, \varepsilon_i)}{Var(\tilde{x}_i)} =$$

Primer 1: Angrist (1998)

- Let $x_i \in \{0,1\}$; general causal model: $y_i = \underbrace{(y_i(1) - y_i(0))}_{\beta_i} x_i + \underbrace{y_i(0)}_{\varepsilon_i}$
  - Design: $x_i \mid w, y(0), y(1) \sim F_x(w_i)$ with linear $E[x_i \mid w_i]$
- What does the w_i-controlled regression *actually* identify?

$$\beta = \frac{Cov(\tilde{x}_i, y_i)}{Var(\tilde{x}_i)} = \frac{Cov(\tilde{x}_i, x_i\beta_i)}{Var(\tilde{x}_i)} + \frac{Cov(\tilde{x}_i, \varepsilon_i)}{Var(\tilde{x}_i)} = \frac{E[\tilde{x}_i x_i \beta_i]}{E[\tilde{x}_i^2]}$$

Primer 1: Angrist (1998)

- Let $x_i \in \{0,1\}$; general causal model: $y_i = \underbrace{(y_i(1) - y_i(0))}_{\beta_i} x_i + \underbrace{y_i(0)}_{\varepsilon_i}$
  - Design: $x_i \mid w, y(0), y(1) \sim F_x(w_i)$ with linear $E[x_i \mid w_i]$

- What does the w_i-controlled regression *actually* identify?

$$\beta = \frac{Cov(\tilde{x}_i, y_i)}{Var(\tilde{x}_i)} = \frac{Cov(\tilde{x}_i, x_i\beta_i)}{Var(\tilde{x}_i)} + \frac{Cov(\tilde{x}_i, \varepsilon_i)}{Var(\tilde{x}_i)} = \frac{E[\tilde{x}_i x_i \beta_i]}{E[\tilde{x}_i^2]}$$

- Further, $E[\tilde{x}_i x_i \beta_i] = E[E[\tilde{x}_i x_i \mid w, y(0), y(1)]\beta_i] =$

Primer 1: Angrist (1998)

- Let $x_i \in \{0,1\}$; general causal model: $y_i = \underbrace{(y_i(1) - y_i(0))}_{\beta_i} x_i + \underbrace{y_i(0)}_{\varepsilon_i}$
  - Design: $x_i \mid w, y(0), y(1) \sim F_x(w_i)$ with linear $E[x_i \mid w_i]$

- What does the w_i-controlled regression *actually* identify?

$$\beta = \frac{Cov(\tilde{x}_i, y_i)}{Var(\tilde{x}_i)} = \frac{Cov(\tilde{x}_i, x_i\beta_i)}{Var(\tilde{x}_i)} + \frac{Cov(\tilde{x}_i, \varepsilon_i)}{Var(\tilde{x}_i)} = \frac{E[\tilde{x}_i x_i \beta_i]}{E[\tilde{x}_i^2]}$$

- Further, $E[\tilde{x}_i x_i \beta_i] = E[E[\tilde{x}_i x_i \mid w, y(0), y(1)]\beta_i] = E[Var(x_i \mid w)\beta_i]$ and $E[\tilde{x}_i^2] =$

Primer 1: Angrist (1998)

- Let $x_i \in \{0,1\}$; general causal model: $y_i = \underbrace{(y_i(1) - y_i(0))}_{\beta_i} x_i + \underbrace{y_i(0)}_{\varepsilon_i}$
  - Design: $x_i \mid w, y(0), y(1) \sim F_x(w_i)$ with linear $E[x_i \mid w_i]$
- What does the w_i-controlled regression *actually* identify?

$$\beta = \frac{Cov(\tilde{x}_i, y_i)}{Var(\tilde{x}_i)} = \frac{Cov(\tilde{x}_i, x_i\beta_i)}{Var(\tilde{x}_i)} + \frac{Cov(\tilde{x}_i, \varepsilon_i)}{Var(\tilde{x}_i)} = \frac{E[\tilde{x}_i x_i \beta_i]}{E[\tilde{x}_i^2]}$$

- Further, $E[\tilde{x}_i x_i \beta_i] = E[E[\tilde{x}_i x_i \mid w, y(0), y(1)]\beta_i] = E[Var(x_i \mid w)\beta_i]$ and $E[\tilde{x}_i^2] = E[Var(x_i \mid w)]$

Primer 1: Angrist (1998)

- Let $x_i \in \{0, 1\}$; general causal model: $y_i = \underbrace{(y_i(1) - y_i(0))}_{\beta_i} x_i + \underbrace{y_i(0)}_{\varepsilon_i}$
  - Design: $x_i \mid w, y(0), y(1) \sim F_x(w_i)$ with linear $E[x_i \mid w_i]$
- What does the w_i-controlled regression *actually* identify?

$$\beta = \frac{Cov(\tilde{x}_i, y_i)}{Var(\tilde{x}_i)} = \frac{Cov(\tilde{x}_i, x_i\beta_i)}{Var(\tilde{x}_i)} + \frac{Cov(\tilde{x}_i, \varepsilon_i)}{Var(\tilde{x}_i)} = \frac{E[\tilde{x}_i x_i \beta_i]}{E[\tilde{x}_i^2]}$$

- Further, $E[\tilde{x}_i x_i \beta_i] = E[E[\tilde{x}_i x_i \mid w, y(0), y(1)]\beta_i] = E[Var(x_i \mid w)\beta_i]$ and $E[\tilde{x}_i^2] = E[Var(x_i \mid w)]$
- Hence the regression a proper (convex) weighted avg. of the β_i:

$$\beta = \frac{E[Var(x_i \mid w)\beta_i]}{E[Var(x_i \mid w)]}$$

More weight put on observations with more treatment variability

Primer 2: TWFE with Staggered Adoption

- Now suppose we have a panel: $y_{it} = (y_{it}(1) - y_{it}(0))x_{it} + y_{it}(0)$
  - Units (non-randomly) adopt treatment over time: $x_{it} = \mathbf{1}[t \geq g_i]$ where $g_i \in \{1, \ldots, T\} \cup \infty$ gives adoption time ($g_i = \infty$ for never treated)

Primer 2: TWFE with Staggered Adoption

- Now suppose we have a panel: $y_{it} = (y_{it}(1) - y_{it}(0))x_{it} + y_{it}(0)$
  - Units (non-randomly) adopt treatment over time: $x_{it} = \mathbf{1}[t \geq g_i]$ where $g_i \in \{1, \ldots, T\} \cup \infty$ gives adoption time ($g_i = \infty$ for never treated)

- We assume parallel trends in $y_{it}(0)$ and run TWFE:

$$y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$$

If we start with a constant FX model, we'd be done!

Primer 2: TWFE with Staggered Adoption

- Now suppose we have a panel: $y_{it} = (y_{it}(1) - y_{it}(0))x_{it} + y_{it}(0)$
  - Units (non-randomly) adopt treatment over time: $x_{it} = \mathbf{1}[t \geq g_i]$ where $g_i \in \{1, \ldots, T\} \cup \infty$ gives adoption time ($g_i = \infty$ for never treated)

- We assume parallel trends in $y_{it}(0)$ and run TWFE:

$$y_{it} = \beta x_{it} + \alpha_i + \tau_t + \nu_{it}$$

  If we start with a constant FX model, we'd be done!

- But notice something a bit weird here: we can run this regression even if there are no never-treated units ...
  - How, then, is the regression using parallel trends in $y_{it}(0)$?

Simple Staggered Adoption

- Consider $T = 2$ and two groups: *always-treated* units (with $g_i = 1$; $x_{i1} = x_{i2} = 1$) and *switchers* (with $g_i = 2$; $x_{i1} = 0$, $x_{i2} = 1$)

Simple Staggered Adoption

- Consider $T = 2$ and two groups: *always-treated* units (with $g_i = 1$; $x_{i1} = x_{i2} = 1$) and *switchers* (with $g_i = 2$; $x_{i1} = 0$, $x_{i2} = 1$)
  - We can use the usual two-period trick: $\Delta y_i = \tau + \beta^{OLS} \Delta x_i + \Delta \nu_i$, so $\beta^{OLS} = E[\Delta y_i \mid g_i = 2] - E[\Delta y_i \mid g_i = 1]$

Simple Staggered Adoption

- Consider $T = 2$ and two groups: *always-treated* units (with $g_i = 1$; $x_{i1} = x_{i2} = 1$) and *switchers* (with $g_i = 2$; $x_{i1} = 0$, $x_{i2} = 1$)
  - We can use the usual two-period trick: $\Delta y_i = \tau + \beta^{OLS} \Delta x_i + \Delta \nu_i$, so $\beta^{OLS} = E[\Delta y_i \mid g_i = 2] - E[\Delta y_i \mid g_i = 1]$

- Under PT, $E[y_{i2}(0) - y_{i1}(0) \mid g_i = 1] = E[y_{i2}(0) - y_{i1}(0) \mid g_i = 2]$ so:

$$\beta = E[y_{i2}(1) - y_{i1}(0) \mid g_i = 2] - E[y_{i2}(1) - y_{i1}(1) \mid g_i = 1]$$

Simple Staggered Adoption

- Consider $T = 2$ and two groups: *always-treated* units (with $g_i = 1$; $x_{i1} = x_{i2} = 1$) and *switchers* (with $g_i = 2$; $x_{i1} = 0$, $x_{i2} = 1$)
    - We can use the usual two-period trick: $\Delta y_i = \tau + \beta^{OLS} \Delta x_i + \Delta v_i$, so $\beta^{OLS} = E[\Delta y_i \mid g_i = 2] - E[\Delta y_i \mid g_i = 1]$

- Under PT, $E[y_{i2}(0) - y_{i1}(0) \mid g_i = 1] = E[y_{i2}(0) - y_{i1}(0) \mid g_i = 2]$ so:

$$\begin{aligned}
\beta =& E[y_{i2}(1) - y_{i1}(0) \mid g_i = 2] - E[y_{i2}(1) - y_{i1}(1) \mid g_i = 1] \\
=& E[y_{i2}(1) - y_{i2}(0) \mid g_i = 2] + E[y_{i2}(0) - y_{i1}(0) \mid g_i = 2] \\
& - E[y_{i2}(1) - y_{i2}(0) \mid g_i = 1] + E[y_{i1}(1) - y_{i1}(0) \mid g_i = 1] \\
& - E[y_{i2}(0) - y_{i1}(0) \mid g_i = 1]
\end{aligned}$$

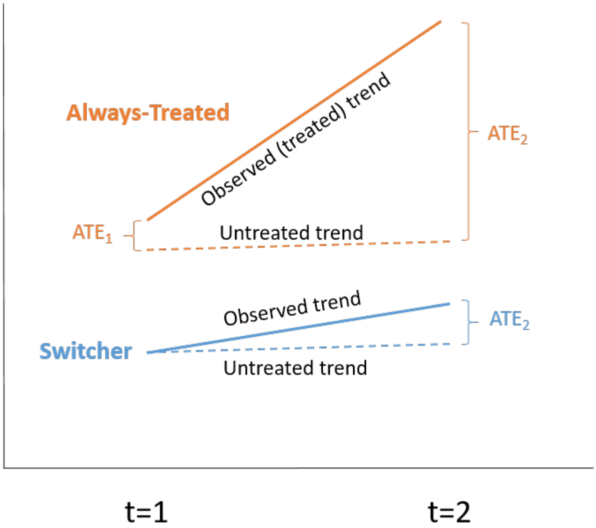
Simple Staggered Adoption

- Consider $T = 2$ and two groups: *always-treated* units (with $g_i = 1$; $x_{i1} = x_{i2} = 1$) and *switchers* (with $g_i = 2$; $x_{i1} = 0$, $x_{i2} = 1$)
  - We can use the usual two-period trick: $\Delta y_i = \tau + \beta^{OLS} \Delta x_i + \Delta v_i$, so $\beta^{OLS} = E[\Delta y_i \mid g_i = 2] - E[\Delta y_i \mid g_i = 1]$

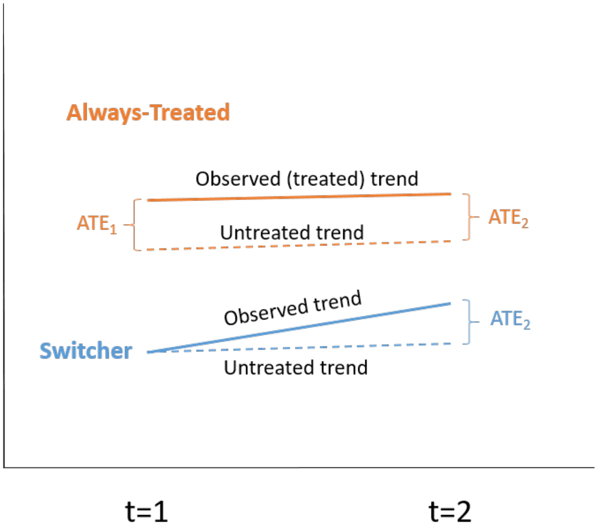
- Under PT, $E[y_{i2}(0) - y_{i1}(0) \mid g_i = 1] = E[y_{i2}(0) - y_{i1}(0) \mid g_i = 2]$ so:

$$
\begin{aligned}
\beta =& E[y_{i2}(1) - y_{i1}(0) \mid g_i = 2] - E[y_{i2}(1) - y_{i1}(1) \mid g_i = 1] \\
=& E[y_{i2}(1) - y_{i2}(0) \mid g_i = 2] + E[y_{i2}(0) - y_{i1}(0) \mid g_i = 2] \\
& - E[y_{i2}(1) - y_{i2}(0) \mid g_i = 1] + E[y_{i1}(1) - y_{i1}(0) \mid g_i = 1] \\
& - E[y_{i2}(0) - y_{i1}(0) \mid g_i = 1] \\
=& \underbrace{E[y_{i2}(1) - y_{i2}(0) \mid g_i = 2]}_{\text{ATE for switchers}} \\
& - (\underbrace{E[y_{i2}(1) - y_{i2}(0) \mid g_i = 1] - E[y_{i1}(1) - y_{i1}(0) \mid g_i = 1]}_{\text{Change in ATE for always-treated}})
\end{aligned}
$$

"Forbidden Comparisons," Illustrated



No Problem Under Constant Effects



Why So Negative?

- We can write the previous expression as a *non-convex* weighted average of $\beta_{it} = y_{it}(1) - y_{it}(0)$:

$$\begin{aligned}\beta =& E[\beta_{i2} \mid g_i = 2] + E[\beta_{i2} \mid g_i = 1] - E[\beta_{i1} \mid g_i = 1] \\ =& \frac{E[\psi_{it}\beta_{it}]}{E[\psi_{it}]} \quad \text{for } \psi_{it} = \begin{cases} +, & \text{if t=2} \\ 0, & \text{if t=1, g=2} \\ -, & \text{if t=1, g=1} \end{cases}\end{aligned}$$

Why So Negative?

- We can write the previous expression as a *non-convex* weighted average of $\beta_{it} = y_{it}(1) - y_{it}(0)$:

$$\beta = E[\beta_{i2} \mid g_i = 2] + E[\beta_{i2} \mid g_i = 1] - E[\beta_{i1} \mid g_i = 1]$$
$$= \frac{E[\psi_{it}\beta_{it}]}{E[\psi_{it}]} \quad \text{for } \psi_{it} = \begin{cases} +, & \text{if t=2} \\ 0, & \text{if t=1, g=2} \\ -, & \text{if t=1, g=1} \end{cases}$$

We'll term the ψ_{it} "ex-post" weights, for reasons you'll see shortly

Why So Negative?

- We can write the previous expression as a *non-convex* weighted average of $\beta_{it} = y_{it}(1) - y_{it}(0)$:

$$\begin{aligned}\beta =& E[\beta_{i2} \mid g_i = 2] + E[\beta_{i2} \mid g_i = 1] - E[\beta_{i1} \mid g_i = 1] \\ =& \frac{E[\psi_{it}\beta_{it}]}{E[\psi_{it}]} \quad \text{for } \psi_{it} = \begin{cases} +, & \text{if t=2} \\ 0, & \text{if t=1, g=2} \\ -, & \text{if t=1, g=1} \end{cases}\end{aligned}$$

  We'll term the ψ_{it} "ex-post" weights, for reasons you'll see shortly
- Why is negativity of ψ_{it} a concern? The potential for *sign reversals*:
  - Even if all TEs are positive β^{OLS} could end up negative (or vice versa)

Why So Negative?

- We can write the previous expression as a *non-convex* weighted average of $\beta_{it} = y_{it}(1) - y_{it}(0)$:

$$\begin{aligned}\beta =& E[\beta_{i2} \mid g_i = 2] + E[\beta_{i2} \mid g_i = 1] - E[\beta_{i1} \mid g_i = 1] \\ =& \frac{E[\psi_{it}\beta_{it}]}{E[\psi_{it}]} \quad \text{for } \psi_{it} = \begin{cases} +, & \text{if t=2} \\ 0, & \text{if t=1, g=2} \\ -, & \text{if t=1, g=1} \end{cases}\end{aligned}$$

  We'll term the ψ_{it} "ex-post" weights, for reasons you'll see shortly

- Why is negativity of ψ_{it} a concern? The potential for *sign reversals*:
  - Even if all TEs are positive β^{OLS} could end up negative (or vice versa)
  - The recent TWFE literature points this issue out in many settings and proposes alternative specifications / procedures to address it

Why So Negative?

- We can write the previous expression as a *non-convex* weighted average of $\beta_{it} = y_{it}(1) - y_{it}(0)$:

$$\begin{aligned}\beta =& E[\beta_{i2} \mid g_i = 2] + E[\beta_{i2} \mid g_i = 1] - E[\beta_{i1} \mid g_i = 1] \\ =& \frac{E[\psi_{it}\beta_{it}]}{E[\psi_{it}]} \quad \text{for } \psi_{it} = \begin{cases} +, & \text{if t=2} \\ 0, & \text{if t=1, g=2} \\ -, & \text{if t=1, g=1} \end{cases}\end{aligned}$$

  We'll term the ψ_{it} "ex-post" weights, for reasons you'll see shortly

- Why is negativity of ψ_{it} a concern? The potential for *sign reversals*:
  - Even if all TEs are positive β^{OLS} could end up negative (or vice versa)
  - The recent TWFE literature points this issue out in many settings and proposes alternative specifications / procedures to address it

- It turns out that such ψ_i also arise in design-based specifications, and they can also be negative
  - But sign reversals are impossible in design-based specs: then we can also write $\beta = E[\phi_i \beta_i]/E[\phi_i]$ for "ex-ante" ϕ_i which are non-negative

Simple Setup

- Suppose a researcher estimates by OLS:

$$y_i = \beta x_i + w_i'\gamma + e_i$$

for some outcome y_i, treatment x_i, and vector of controls w_i

Simple Setup

- Suppose a researcher estimates by OLS:

$$y_i = \beta x_i + w_i' \gamma + e_i$$

for some outcome y_i, treatment x_i, and vector of controls w_i

- To interpret β, we consider a linear-effect causal model:

$$y_i = \beta_i x_i + \varepsilon_i$$

with heterogeneous effects β_i and untreated potential outcomes ε_i

Simple Setup

- Suppose a researcher estimates by OLS:

$$y_i = \beta x_i + w_i'\gamma + e_i$$

for some outcome y_i, treatment x_i, and vector of controls w_i

- To interpret β, we consider a linear-effect causal model:

$$y_i = \beta_i x_i + \varepsilon_i$$

with heterogeneous effects β_i and untreated potential outcomes ε_i

- In large enough samples, OLS consistently estimates:

$$\beta = \frac{E[\tilde{x}_i y_i]}{E[\tilde{x}_i^2]} = \frac{E[\tilde{x}_i x_i \beta]}{E[\tilde{x}_i^2]} + \frac{E[\tilde{x}_i \varepsilon_i]}{E[\tilde{x}_i^2]}$$

where $\tilde{x}_i$ are residuals from the population regression of x_i on w_i

Two Paths to Avoiding Omitted Variables Bias

- $E[\tilde{x}_i \varepsilon_i] = 0$ under either one of two assumptions:

Two Paths to Avoiding Omitted Variables Bias

- $E[\tilde{x}_i \varepsilon_i] = 0$ under either one of two assumptions:

  ASSUMPTION 1: $E[\varepsilon_i \mid x_i, w_i] = w_i'\gamma$

  - Untreated potential outcomes are linear in controls, given treatment
  - E.g. parallel trends, where i indexes unit-period pairs in a panel and w_i includes unit and time dummies

Two Paths to Avoiding Omitted Variables Bias

- $E[\tilde{x}_i \varepsilon_i] = 0$ under either one of two assumptions:

ASSUMPTION 1: $E[\varepsilon_i \mid x_i, w_i] = w_i' \gamma$

  - Untreated potential outcomes are linear in controls, given treatment
  - E.g. parallel trends, where i indexes unit-period pairs in a panel and w_i includes unit and time dummies

ASSUMPTION 2: $E[x_i \mid \varepsilon_i, \beta_i, w_i] = w_i' \lambda$

  - Treatment is conditionally mean-independent of potential outcomes, with a linear *expected treatment* $E[x_i \mid w_i]$ (e.g. the propensity score)
  - E.g. a stratified experiment, where x_i is randomly assigned within strata dummied out in w_i
  - Note we're conditioning on *both* ε_i and β_i, ruling out "selection on gains" (will relax with IV version soon)

Two Paths to Avoiding Omitted Variables Bias

- $E[\tilde{x}_i \varepsilon_i] = 0$ under either one of two assumptions:

  ASSUMPTION 1: $E[\varepsilon_i \mid x_i, w_i] = w_i' \gamma$

  - Untreated potential outcomes are linear in controls, given treatment
  - E.g. parallel trends, where i indexes unit-period pairs in a panel and w_i includes unit and time dummies

  ASSUMPTION 2: $E[x_i \mid \varepsilon_i, \beta_i, w_i] = w_i' \lambda$

  - Treatment is conditionally mean-independent of potential outcomes, with a linear *expected treatment* $E[x_i \mid w_i]$ (e.g. the propensity score)
  - E.g. a stratified experiment, where x_i is randomly assigned within strata dummied out in w_i
  - Note we're conditioning on *both* ε_i and β_i, ruling out "selection on gains" (will relax with IV version soon)

- The second assumption yields a design-based OLS specification

  - Stronger (sufficient) condition: $x_i \mid (\varepsilon_i, \beta_i, w_i) \overset{iid}{\sim} F_x(w_i)$

Ex-Post Weights

- Since $E[\tilde{x}_i \varepsilon_i] = 0$, the OLS estimand has an average-effect representation under either assumption:

$$\beta = \frac{E[\psi_i \beta_i]}{E[\psi_i]}, \qquad \psi_i = \tilde{x}_i x_i$$

Ex-Post Weights

- Since $E[\tilde{x}_i \varepsilon_i] = 0$, the OLS estimand has an average-effect representation under either assumption:

$$\beta = \frac{E[\psi_i \beta_i]}{E[\psi_i]}, \qquad \psi_i = \tilde{x}_i x_i$$

- But the ex-post weights ψ_i are generally non-convex: $E[\tilde{x}_i] = 0$, so $\tilde{x}_i$ must take on both positive and negative values

Ex-Post Weights

- Since $E[\tilde{x}_i \varepsilon_i] = 0$, the OLS estimand has an average-effect representation under either assumption:

$$\beta = \frac{E[\psi_i \beta_i]}{E[\psi_i]}, \qquad \psi_i = \tilde{x}_i x_i$$

- But the ex-post weights ψ_i are generally non-convex: $E[\tilde{x}_i] = 0$, so $\tilde{x}_i$ must take on both positive and negative values
    - E.g. if $x_i > 0$ then i with low values of x_i (the effective control group) will always receive negative ex-post weight

Ex-Post Weights

- Since $E[\tilde{x}_i \varepsilon_i] = 0$, the OLS estimand has an average-effect representation under either assumption:

$$\beta = \frac{E[\psi_i \beta_i]}{E[\psi_i]}, \qquad \psi_i = \tilde{x}_i x_i$$

- But the ex-post weights ψ_i are generally non-convex: $E[\tilde{x}_i] = 0$, so $\tilde{x}_i$ must take on both positive and negative values
  - E.g. if $x_i > 0$ then i with low values of x_i (the effective control group) will always receive negative ex-post weight
  - This can lead to sign reversals: e.g. $\beta < 0$, despite $\beta_i > 0$

Ex-Post Weights

- Since $E[\tilde{x}_i \varepsilon_i] = 0$, the OLS estimand has an average-effect representation under either assumption:

$$\beta = \frac{E[\psi_i \beta_i]}{E[\psi_i]}, \qquad \psi_i = \tilde{x}_i x_i$$

- But the ex-post weights ψ_i are generally non-convex: $E[\tilde{x}_i] = 0$, so $\tilde{x}_i$ must take on both positive and negative values
  - E.g. if $x_i > 0$ then i with low values of x_i (the effective control group) will always receive negative ex-post weight
  - This can lead to sign reversals: e.g. $\beta < 0$, despite $\beta_i > 0$
- The ex-post weights are the end of the story for β under Assumption 1. But in design-based specifications we can take one more step
  - In experiments, who is in the effective control group is *random*. Before treatment is drawn, everyone expects the same weight!

Ex-Ante Weights

- Using the law of iterated expectations, we can also write:

$$\beta = \frac{E[E[\psi_i \mid w_i, \beta_i]\beta_i]}{E[E[\psi_i \mid w_i, \beta_i]]} \equiv \frac{E[\phi_i \beta_i]}{E[\phi_i]}$$

for ex-ante weights $\phi_i = E[\tilde{x}_i x_i \mid w_i, \beta_i]$

Ex-Ante Weights

- Using the law of iterated expectations, we can also write:

$$\beta = \frac{E[E[\psi_i \mid w_i, \beta_i]\beta_i]}{E[E[\psi_i \mid w_i, \beta_i]]} \equiv \frac{E[\phi_i \beta_i]}{E[\phi_i]}$$

  for ex-ante weights $\phi_i = E[\tilde{x}_i x_i \mid w_i, \beta_i]$

  - Under Assumption 1, this need not help: i.e. if treatment is deterministic in the unit/time FE in w_i, then $\phi_i = \psi_i$

Ex-Ante Weights

- Using the law of iterated expectations, we can also write:

$$\beta = \frac{E[E[\psi_i \mid w_i, \beta_i]\beta_i]}{E[E[\psi_i \mid w_i, \beta_i]]} \equiv \frac{E[\phi_i \beta_i]}{E[\phi_i]}$$

  for ex-ante weights $\phi_i = E[\tilde{x}_i x_i \mid w_i, \beta_i]$

  - Under Assumption 1, this need not help: i.e. if treatment is deterministic in the unit/time FE in w_i, then $\phi_i = \psi_i$

- But under Assumption 2, $\phi_i = Var(x_i \mid w_i, \beta_i)$ which is non-negative!

Ex-Ante Weights

- Using the law of iterated expectations, we can also write:

$$\beta = \frac{E[E[\psi_i \mid w_i, \beta_i]\beta_i]}{E[E[\psi_i \mid w_i, \beta_i]]} \equiv \frac{E[\phi_i \beta_i]}{E[\phi_i]}$$

  for ex-ante weights $\phi_i = E[\tilde{x}_i x_i \mid w_i, \beta_i]$

  - Under Assumption 1, this need not help: i.e. if treatment is deterministic in the unit/time FE in w_i, then $\phi_i = \psi_i$

- But under Assumption 2, $\phi_i = Var(x_i \mid w_i, \beta_i)$ which is non-negative!
  - $E[\tilde{x}_i x_i \mid w_i, \beta_i] = E[\tilde{x}_i^2 \mid w_i, \beta_i] + E[\tilde{x}_i \mid w_i, \beta_i] w_i' \lambda = Var(x_i \mid w_i, \beta_i) + 0$

Ex-Ante Weights

- Using the law of iterated expectations, we can also write:

$$\beta = \frac{E[E[\psi_i \mid w_i, \beta_i]\beta_i]}{E[E[\psi_i \mid w_i, \beta_i]]} \equiv \frac{E[\phi_i \beta_i]}{E[\phi_i]}$$

  for ex-ante weights $\phi_i = E[\tilde{x}_i x_i \mid w_i, \beta_i]$

  - Under Assumption 1, this need not help: i.e. if treatment is deterministic in the unit/time FE in w_i, then $\phi_i = \psi_i$

- But under Assumption 2, $\phi_i = Var(x_i \mid w_i, \beta_i)$ which is non-negative!
  - $E[\tilde{x}_i x_i \mid w_i, \beta_i] = E[\tilde{x}_i^2 \mid w_i, \beta_i] + E[\tilde{x}_i \mid w_i, \beta_i] w_i' \lambda = Var(x_i \mid w_i, \beta_i) + 0$

- Hence: sign reversals cannot occur in design-based OLS specifications

Interpretation

- Even if we formulate a design-based regression in terms of constant effects, the estimand is still reasonable under heterogeneous effects
  - Not necessarily true for outcome models (makes sense: we were just modeling ε_i! But additional models on β_i need not help)

Interpretation

- Even if we formulate a design-based regression in terms of constant effects, the estimand is still reasonable under heterogeneous effects
  - Not necessarily true for outcome models (makes sense: we were just modeling ε_i! But additional models on β_i need not help)
- With the stronger design assumption of $x_i \mid (\varepsilon_i, \beta_i, w_i) \overset{iid}{\sim} G(w_i)$, the ex ante weights become identified: $\phi_i = Var(x_i \mid w_i, \beta_i) = Var(x_i \mid w_i)$
  - C.f. earlier results in Angrist (1998), Angrist and Krueger (1999), etc

Interpretation

- Even if we formulate a design-based regression in terms of constant effects, the estimand is still reasonable under heterogeneous effects
    - Not necessarily true for outcome models (makes sense: we were just modeling ε_i! But additional models on β_i need not help)
- With the stronger design assumption of $x_i \mid (\varepsilon_i, \beta_i, w_i) \overset{iid}{\sim} G(w_i)$, the ex ante weights become identified: $\phi_i = Var(x_i \mid w_i, \beta_i) = Var(x_i \mid w_i)$
    - C.f. earlier results in Angrist (1998), Angrist and Krueger (1999), etc
    - Could inverse-weight by $\widehat{Var}(x_i \mid w_i)$ to estimate unweighted $E[\beta_i]$

Interpretation

- Even if we formulate a design-based regression in terms of constant effects, the estimand is still reasonable under heterogeneous effects
  - Not necessarily true for outcome models (makes sense: we were just modeling ε_i! But additional models on β_i need not help)
- With the stronger design assumption of $x_i \mid (\varepsilon_i, \beta_i, w_i) \overset{iid}{\sim} G(w_i)$, the ex ante weights become identified: $\phi_i = Var(x_i \mid w_i, \beta_i) = Var(x_i \mid w_i)$
  - C.f. earlier results in Angrist (1998), Angrist and Krueger (1999), etc
  - Could inverse-weight by $\widehat{Var}(x_i \mid w_i)$ to estimate unweighted $E[\beta_i]$
- Of course, the ϕ_i-weighted estimand may not be most of interest!
  - If $Cov(\phi_i, \beta_i) \approx 0$, we'll still get something close to $E[\beta_i]$
  - Otherwise, ϕ_i-weighting has desirable efficiency properties (Goldsmith-Pinkham et al. 2024)
  - Large class of alternative propensity-score-based estimators for other estimands under the stronger design assumption

General Setting

- Borusyak and Hull (2024) extend ex ante / ex post weights to:
  1. A more general causal model: potential outcomes $y_i(x)$ and $y_i = y_i(x_i)$
  2. IV: design-based assumption is then $E[z_i \mid y_i(\cdot), w_i] = w_i'\lambda$

  Note implicit exclusion restriction: $y_i(\cdot)$ is indexed by x, not z

General Setting

- Borusyak and Hull (2024) extend ex ante / ex post weights to:
  1. A more general causal model: potential outcomes $y_i(x)$ and $y_i = y_i(x_i)$
  2. IV: design-based assumption is then $E[z_i \mid y_i(\cdot), w_i] = w_i'\lambda$

  Note implicit exclusion restriction: $y_i(\cdot)$ is indexed by x, not z

- For convex ex-ante weights in IV we require first-stage *monotonicity*: that x_i is non-decreasing in z_i for all units regardless of $y_i(\cdot)$
  - C.f. earlier results in Imbens and Angrist ('94, '95), Angrist et al. ('00)

General Setting

- Borusyak and Hull (2024) extend ex ante / ex post weights to:
  1. A more general causal model: potential outcomes $y_i(x)$ and $y_i = y_i(x_i)$
  2. IV: design-based assumption is then $E[z_i \mid y_i(\cdot), w_i] = w_i'\lambda$

  Note implicit exclusion restriction: $y_i(\cdot)$ is indexed by x, not z

- For convex ex-ante weights in IV we require first-stage *monotonicity*: that x_i is non-decreasing in z_i for all units regardless of $y_i(\cdot)$
  - C.f. earlier results in Imbens and Angrist ('94, '95), Angrist et al. ('00)
  - Ex post weights are still potentially non-convex under monotonicity

General Setting

- Borusyak and Hull (2024) extend ex ante / ex post weights to:
  1. A more general causal model: potential outcomes $y_i(x)$ and $y_i = y_i(x_i)$
  2. IV: design-based assumption is then $E[z_i \mid y_i(\cdot), w_i] = w_i'\lambda$

  Note implicit exclusion restriction: $y_i(\cdot)$ is indexed by x, not z

- For convex ex-ante weights in IV we require first-stage *monotonicity*: that x_i is non-decreasing in z_i for all units regardless of $y_i(\cdot)$
  - C.f. earlier results in Imbens and Angrist ('94, '95), Angrist et al. ('00)
  - Ex post weights are still potentially non-convex under monotonicity

- Framework is general, allowing for "formula" IVs (e.g. shift-share) where the first stage relationship need not be causal
  - We'll see more about this in tomorrow's class

Special Case: IV with Linear Heterogeneity

- Suppose $y_i = \beta_i x_i + \varepsilon_i$ (without loss of generality for binary x_i). Then:

$$\beta = \frac{Cov(\tilde{z}_i, y_i)}{Cov(\tilde{z}_i, x_i)} =$$

Special Case: IV with Linear Heterogeneity

- Suppose $y_i = \beta_i x_i + \varepsilon_i$ (without loss of generality for binary x_i). Then:

$$\beta = \frac{Cov(\tilde{z}_i, y_i)}{Cov(\tilde{z}_i, x_i)} = \frac{Cov(\tilde{z}_i, \beta_i x_i + \varepsilon_i)}{Cov(\tilde{z}_i, x_i)} =$$

Special Case: IV with Linear Heterogeneity

- Suppose $y_i = \beta_i x_i + \varepsilon_i$ (without loss of generality for binary x_i). Then:

$$\beta = \frac{Cov(\tilde{z}_i, y_i)}{Cov(\tilde{z}_i, x_i)} = \frac{Cov(\tilde{z}_i, \beta_i x_i + \varepsilon_i)}{Cov(\tilde{z}_i, x_i)} = \frac{E[\tilde{z}_i x_i \beta_i]}{E[\tilde{z}_i x_i]}$$

$$=$$

Special Case: IV with Linear Heterogeneity

- Suppose $y_i = \beta_i x_i + \varepsilon_i$ (without loss of generality for binary x_i). Then:

$$
\begin{aligned}
\beta =& \frac{Cov(\tilde{z}_i, y_i)}{Cov(\tilde{z}_i, x_i)} = \frac{Cov(\tilde{z}_i, \beta_i x_i + \varepsilon_i)}{Cov(\tilde{z}_i, x_i)} = \frac{E[\tilde{z}_i x_i \beta_i]}{E[\tilde{z}_i x_i]} \\
=& \frac{E[E[\tilde{z}_i x_i \mid w, \beta]\beta_i]}{E[E[\tilde{z}_i x_i \mid w, \beta]]} =
\end{aligned}
$$

Special Case: IV with Linear Heterogeneity

- Suppose $y_i = \beta_i x_i + \varepsilon_i$ (without loss of generality for binary x_i). Then:

$$
\begin{aligned}
\beta &= \frac{Cov(\tilde{z}_i, y_i)}{Cov(\tilde{z}_i, x_i)} = \frac{Cov(\tilde{z}_i, \beta_i x_i + \varepsilon_i)}{Cov(\tilde{z}_i, x_i)} = \frac{E[\tilde{z}_i x_i \beta_i]}{E[\tilde{z}_i x_i]} \\
&= \frac{E[E[\tilde{z}_i x_i \mid w, \beta]\beta_i]}{E[E[\tilde{z}_i x_i \mid w, \beta]]} = \frac{E[Cov(z_i, x_i \mid w, \beta)\beta_i]}{E[Cov(z_i, x_i \mid w, \beta)]} \\
&=
\end{aligned}
$$

Special Case: IV with Linear Heterogeneity

- Suppose $y_i = \beta_i x_i + \varepsilon_i$ (without loss of generality for binary x_i). Then:

$$
\begin{aligned}
\beta &= \frac{Cov(\tilde{z}_i, y_i)}{Cov(\tilde{z}_i, x_i)} = \frac{Cov(\tilde{z}_i, \beta_i x_i + \varepsilon_i)}{Cov(\tilde{z}_i, x_i)} = \frac{E[\tilde{z}_i x_i \beta_i]}{E[\tilde{z}_i x_i]} \\
&= \frac{E[E[\tilde{z}_i x_i \mid w, \beta]\beta_i]}{E[E[\tilde{z}_i x_i \mid w, \beta]]} = \frac{E[Cov(z_i, x_i \mid w, \beta)\beta_i]}{E[Cov(z_i, x_i \mid w, \beta)]} \\
&= \frac{E[\sigma_i \pi_i \beta_i]}{E[\sigma_i \pi_i]}
\end{aligned}
$$

where:

$$
\begin{aligned}
\sigma_i &= Var(z_i \mid w, \beta) && \Longrightarrow \text{ more weight where } z_i \text{ varies more} \\
\pi_i &= \frac{Cov(z_i, x_i \mid w, \beta)}{Var(z_i \mid w, \beta)} && \Longrightarrow \text{ more weight where the first stage is larger}
\end{aligned}
$$

Special Case: IV with Linear Heterogeneity

- Suppose $y_i = \beta_i x_i + \varepsilon_i$ (without loss of generality for binary x_i). Then:

$$\begin{aligned}
\beta &= \frac{Cov(\tilde{z}_i, y_i)}{Cov(\tilde{z}_i, x_i)} = \frac{Cov(\tilde{z}_i, \beta_i x_i + \varepsilon_i)}{Cov(\tilde{z}_i, x_i)} = \frac{E[\tilde{z}_i x_i \beta_i]}{E[\tilde{z}_i x_i]} \\
&= \frac{E[E[\tilde{z}_i x_i \mid w, \beta]\beta_i]}{E[E[\tilde{z}_i x_i \mid w, \beta]]} = \frac{E[Cov(z_i, x_i \mid w, \beta)\beta_i]}{E[Cov(z_i, x_i \mid w, \beta)]} \\
&= \frac{E[\sigma_i \pi_i \beta_i]}{E[\sigma_i \pi_i]}
\end{aligned}$$

where:

$$\begin{aligned}
\sigma_i &= Var(z_i \mid w, \beta) && \Longrightarrow \text{ more weight where } z_i \text{ varies more} \\
\pi_i &= \frac{Cov(z_i, x_i \mid w, \beta)}{Var(z_i \mid w, \beta)} && \Longrightarrow \text{ more weight where the first stage is larger}
\end{aligned}$$

- Reduces to Angrist '98 result if $z_i = x_i$ is fully independently assigned

Special Case: IV with Linear Heterogeneity

- Suppose $y_i = \beta_i x_i + \varepsilon_i$ (without loss of generality for binary x_i). Then:

$$\begin{aligned}\beta &= \frac{Cov(\tilde{z}_i, y_i)}{Cov(\tilde{z}_i, x_i)} = \frac{Cov(\tilde{z}_i, \beta_i x_i + \varepsilon_i)}{Cov(\tilde{z}_i, x_i)} = \frac{E[\tilde{z}_i x_i \beta_i]}{E[\tilde{z}_i x_i]} \\ &= \frac{E[E[\tilde{z}_i x_i \mid w, \beta]\beta_i]}{E[E[\tilde{z}_i x_i \mid w, \beta]]} = \frac{E[Cov(z_i, x_i \mid w, \beta)\beta_i]}{E[Cov(z_i, x_i \mid w, \beta)]} \\ &= \frac{E[\sigma_i \pi_i \beta_i]}{E[\sigma_i \pi_i]}\end{aligned}$$

  where:

$$\begin{aligned}\sigma_i &= Var(z_i \mid w, \beta) && \implies \text{more weight where } z_i \text{ varies more} \\ \pi_i &= \frac{Cov(z_i, x_i \mid w, \beta)}{Var(z_i \mid w, \beta)} && \implies \text{more weight where the first stage is larger}\end{aligned}$$

- Reduces to Angrist '98 result if $z_i = x_i$ is fully independently assigned
- Reduces to Imbens-Angrist LATE result if the first stage is causal

Multiple Treatments: Contamination Bias

- Goldsmith-Pinkham et al. (2024) generalize things in a different direction: hetFX weighting for regressions w/multiple treatments
  - Unfortunately the picture is a bit less rosy for design here

Multiple Treatments: Contamination Bias

- Goldsmith-Pinkham et al. (2024) generalize things in a different direction: hetFX weighting for regressions w/multiple treatments
  - Unfortunately the picture is a bit less rosy for design here

- The coefficient on treatment j estimates the sum of two terms:
  1. A weighted average of treatment j's effects, with convex weights in design-based specifications ✓

Multiple Treatments: Contamination Bias

- Goldsmith-Pinkham et al. (2024) generalize things in a different direction: hetFX weighting for regressions w/multiple treatments
  - Unfortunately the picture is a bit less rosy for design here

- The coefficient on treatment j estimates the sum of two terms:
  1. A weighted average of treatment j's effects, with convex weights in design-based specifications ✓
  2. A non-convex combination of effects from other treatments k ("contamination bias") X

Multiple Treatments: Contamination Bias

- Goldsmith-Pinkham et al. (2024) generalize things in a different direction: hetFX weighting for regressions w/multiple treatments
  - Unfortunately the picture is a bit less rosy for design here

- The coefficient on treatment j estimates the sum of two terms:
  1. A weighted average of treatment j's effects, with convex weights in design-based specifications ✓
  2. A non-convex combination of effects from other treatments k ("contamination bias") X

  See also Sun and Abraham (2021) for earlier finding in event studies

Multiple Treatments: Contamination Bias

- Goldsmith-Pinkham et al. (2024) generalize things in a different direction: hetFX weighting for regressions w/multiple treatments
  - Unfortunately the picture is a bit less rosy for design here

- The coefficient on treatment j estimates the sum of two terms:
  1. A weighted average of treatment j's effects, with convex weights in design-based specifications ✓
  2. A non-convex combination of effects from other treatments k ("contamination bias") X

  See also Sun and Abraham (2021) for earlier finding in event studies

- We derive alternative estimators which avoid contamination bias while maintaining some nice properties of OLS weighting
  - Ultimately, becomes an empirical question of how important bias is

General Problem

- Goldsmith-Pinkham et al. (2024) consider a partially linear regression:

$$y_i = \sum_k x_{ik}\beta_k + g(w_i) + u_i$$

for mutually exclusive $x_{ik} \in \{0,1\}$ (usual regression: $g(w_i) = w_i'\gamma$)

General Problem

- Goldsmith-Pinkham et al. (2024) consider a partially linear regression:

$$y_i = \sum_k x_{ik}\beta_k + g(w_i) + u_i$$

for mutually exclusive $x_{ik} \in \{0,1\}$ (usual regression: $g(w_i) = w_i'\gamma$)

  - Assume "exogeneity": $E[y_i(k) \mid x_i, w_i] = E[y_i(k) \mid w_i]$ for all k
  - Suppose $g(\cdot)$ is flexible enough to span either $E[y_i(0) \mid w_i]$ (e.g. parallel trends) or $p_k = E[x_{ik} \mid w_i]$ for all k (i.e. design)

General Problem

- Goldsmith-Pinkham et al. (2024) consider a partially linear regression:

$$y_i = \sum_k x_{ik}\beta_k + g(w_i) + u_i$$

for mutually exclusive $x_{ik} \in \{0,1\}$ (usual regression: $g(w_i) = w_i'\gamma$)

  - Assume "exogeneity": $E[y_i(k) \mid x_i, w_i] = E[y_i(k) \mid w_i]$ for all k
  - Suppose $g(\cdot)$ is flexible enough to span either $E[y_i(0) \mid w_i]$ (e.g. parallel trends) or $p_k = E[x_{ik} \mid w_i]$ for all k (i.e. design)

- They show each regression coefficient β_k can be decomposed:

$$\beta_k = E[\lambda_{kk}(w_i)\tau_k(w_i)] + \sum_{\ell \neq k} E[\lambda_{k\ell}(w_i)\tau_\ell(w_i)]$$

for $\tau_k(w_i) = E[y_i(k) - y_i(0) \mid w_i]$, $\lambda_{kk} = \frac{E[\tilde{x}_{ik}x_{ik}|w_i]}{E[\tilde{x}_{ik}^2]}$, $\lambda_{k\ell} = \frac{E[\tilde{x}_{ik}x_{i\ell}|w_i]}{E[\tilde{x}_{ik}^2]}$; $\tilde{x}_{ik}$ is the residual from regressing x_{ik} on $g(w_i)$ *and* all other $x_{i,-k}$

  - $E[\lambda_{kk}(w_i)] = 1$, $E[\lambda_{k\ell}(w_i)] = 0$. Further $\lambda_{kk}(w_i) \geq 0$ if $g(\cdot)$ spans p_k

Unpacking The Result

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

- $E[\lambda_{kk}(w_i)] = 1$, $E[\lambda_{k\ell}(w_i)] = 0$. If [(*) $g(\cdot)$ spans p_k], $\lambda_{kk}(w_i) \geq 0$

Unpacking The Result

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

- $E[\lambda_{kk}(w_i)] = 1$, $E[\lambda_{k\ell}(w_i)] = 0$. If [(*) $g(\cdot)$ spans p_k], $\lambda_{kk}(w_i) \geq 0$
  - (*) corresponds to a "design-based" regression: No negative own-treatment weights (generalizing Angrist '98 further)

Unpacking The Result

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

- $E[\lambda_{kk}(w_i)] = 1$, $E[\lambda_{k\ell}(w_i)] = 0$. If [(*) $g(\cdot)$ spans p_k], $\lambda_{kk}(w_i) \geq 0$
  - (*) corresponds to a "design-based" regression: No negative own-treatment weights (generalizing Angrist '98 further)
  - Unless $\lambda_{k\ell} = 0$ identically, there's potential for contamination bias

Unpacking The Result

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

- $E[\lambda_{kk}(w_i)] = 1$, $E[\lambda_{k\ell}(w_i)] = 0$. If [(*) $g(\cdot)$ spans p_k], $\lambda_{kk}(w_i) \geq 0$
  - (*) corresponds to a "design-based" regression: No negative own-treatment weights (generalizing Angrist '98 further)
  - Unless $\lambda_{k\ell} = 0$ identically, there's potential for contamination bias

- Intuition: FWL partials both $g(w_i)$ and $x_{i,-k}$ out of x_{ik} to estimate β_k
  - The trick to Angrist '98 was that this auxilliary regression identified a CEF (the p-score). But here $E[x_{ik} \mid w_i, x_{i,-k}]$ is inherently nonlinear

Unpacking The Result

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

- $E[\lambda_{kk}(w_i)] = 1$, $E[\lambda_{k\ell}(w_i)] = 0$. If [(*) $g(\cdot)$ spans p_k], $\lambda_{kk}(w_i) \geq 0$
  - (*) corresponds to a "design-based" regression: No negative own-treatment weights (generalizing Angrist '98 further)
  - Unless $\lambda_{k\ell} = 0$ identically, there's potential for contamination bias

- Intuition: FWL partials both $g(w_i)$ and $x_{i,-k}$ out of x_{ik} to estimate β_k
  - The trick to Angrist '98 was that this auxilliary regression identified a CEF (the p-score). But here $E[x_{ik} \mid w_i, x_{i,-k}]$ is inherently nonlinear
  - FWL residual $\tilde{x}_{ik}$ is thus not mean-zero given $(w_i, x_{i,-k})$, so it "picks up" effects of other treatments x_{ik} given w_i

Is This a Problem?

- In principle, contamination bias applies to a large number of settings:
  1. RCTs with multiple treatments and randomization strata
  2. Selection-on-obs with multiple treatments (e.g. "value-added" models)
  3. TWFE with multiple treatments (e.g. "mover" regressions)
  4. IV with multiple instruments (e.g. "examiner/judge" IVs)
  5. Descriptive regressions on multiple variables (e.g. disparity analyses)

Is This a Problem?

- In principle, contamination bias applies to a large number of settings:
  1. RCTs with multiple treatments and randomization strata
  2. Selection-on-obs with multiple treatments (e.g. “value-added” models)
  3. TWFE with multiple treatments (e.g. “mover” regressions)
  4. IV with multiple instruments (e.g. “examiner/judge” IVs)
  5. Descriptive regressions on multiple variables (e.g. disparity analyses)

- But again, the severity of the issue is an empirical matter
  - Since the CB weights average to zero, if they’re uncorrelated with effect heterogeneity there’s no issue
  - The weights are identified; we can estimate them to diagnose bias

Solutions

- Contamination bias comes from the FWL auxilliary regression not controlling "flexibly enough" for $(w_i, x_{i,-k})$ ... but we can fix that:

$$y_i = \sum_k x_{ik}\beta_k + g(w_i) + \sum_k x_{ik}(q_k(w_i) - E[q_k(w_i)]) + u_i$$

The blue term captures non-linearities in (w_i, x_i)

Solutions

- Contamination bias comes from the FWL auxilliary regression not controlling "flexibly enough" for $(w_i, x_{i,-k})$ ... but we can fix that:

$$y_i = \sum_k x_{ik}\beta_k + g(w_i) + \sum_k x_{ik}(q_k(w_i) - E[q_k(w_i)]) + u_i$$

  The blue term captures non-linearities in (w_i, x_i)

  - When $x_i \mid w_i$ is as-good-as-randomly assigned, β_k identifies the ATE of treatment k (Imbens and Wooldridge, 2009)
  - See our *multe* Stata package for automating this + other CB checks

Solutions

- Contamination bias comes from the FWL auxilliary regression not controlling "flexibly enough" for $(w_i, x_{i,-k})$ ... but we can fix that:

$$y_i = \sum_k x_{ik}\beta_k + g(w_i) + \sum_k x_{ik}(q_k(w_i) - E[q_k(w_i)]) + u_i$$

  The blue term captures non-linearities in (w_i, x_i)
  - When $x_i \mid w_i$ is as-good-as-randomly assigned, β_k identifies the ATE of treatment k (Imbens and Wooldridge, 2009)
  - See our *multe* Stata package for automating this + other CB checks

- This works in principle, but in practice can fail / be really noisy
  - Key challenge: limited overlap ($p_k(w_i)$ may be close to zero or one)
  - If CB is limited, an uninteracted regression is likely more efficient...

Solutions (Cont.)

- We could of course instead just focus on one treatment at a time:

$$y_i = x_{ik}\beta_k + g(w_i) + u_i,$$

just using observations where either $x_{ik} = 1$ or $x_{i0} = 1$

  - We know β_k has no negative weights, and is likely precise because of the $Var(x_{ik} \mid w_i)$ weighting regression uses

Solutions (Cont.)

- We could of course instead just focus on one treatment at a time:

$$y_i = x_{ik}\beta_k + g(w_i) + u_i,$$

just using observations where either $x_{ik} = 1$ or $x_{i0} = 1$

  - We know β_k has no negative weights, and is likely precise because of the $Var(x_{ik} \mid w_i)$ weighting regression uses
  - But the weights are k specific, so β_k is hard to compare across k

Solutions (Cont.)

- We could of course instead just focus on one treatment at a time:

$$y_i = x_{ik}\beta_k + g(w_i) + u_i,$$

just using observations where either $x_{ik} = 1$ or $x_{i0} = 1$

  - We know β_k has no negative weights, and is likely precise because of the $Var(x_{ik} \mid w_i)$ weighting regression uses
  - But the weights are k specific, so β_k is hard to compare across k

- Goldsmith-Pinkham et al. (2024) derive an alternative estimator, which uses *common* variance weights across treatments
  - Formally, we show this weighting scheme attains a semiparametric efficiency bound while still avoiding contaimination bias

Solutions (Cont.)

- We could of course instead just focus on one treatment at a time:

$$y_i = x_{ik}\beta_k + g(w_i) + u_i,$$

just using observations where either $x_{ik} = 1$ or $x_{i0} = 1$

  - We know β_k has no negative weights, and is likely precise because of the $Var(x_{ik} \mid w_i)$ weighting regression uses
  - But the weights are k specific, so β_k is hard to compare across k

- Goldsmith-Pinkham et al. (2024) derive an alternative estimator, which uses *common* variance weights across treatments
  - Formally, we show this weighting scheme attains a semiparametric efficiency bound while still avoiding contaimination bias

- As before, whether any of these alternatives give a different answer than OLS is an empirical matter...

Example: Project STAR

- Krueger (1999) studies the STAR RCT, which randomized students in public elementary to one of three classroom types:
  1. Regular-sized (20-25 students) – Control
  2. Small (13-17 students) – Treatment 1
  3. Regular-sized with a teaching aide – Treatment 2

Example: Project STAR

- Krueger (1999) studies the STAR RCT, which randomized students in public elementary to one of three classroom types:
  1. Regular-sized (20-25 students) – Control
  2. Small (13-17 students) – Treatment 1
  3. Regular-sized with a teaching aide – Treatment 2

- Kids were randomized within schools, so the propensity of assignment to each treatment varied by school
  - Krueger thus estimates: $TestScore_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_{school(i)} + \varepsilon_i$

Example: Project STAR

- Krueger (1999) studies the STAR RCT, which randomized students in public elementary to one of three classroom types:
  1. Regular-sized (20-25 students) – Control
  2. Small (13-17 students) – Treatment 1
  3. Regular-sized with a teaching aide – Treatment 2

- Kids were randomized within schools, so the propensity of assignment to each treatment varied by school
  - Krueger thus estimates: $TestScore_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_{school(i)} + \varepsilon_i$

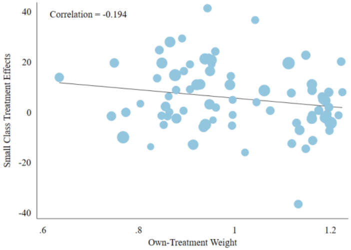
- We find significant *potential* for contamination bias: lots of treatment effect heterogeneity and variation in contamination weights
  - But actual contamination bias is minimal: $Corr(effects, weights) \approx 0$

Project STAR, Revisited

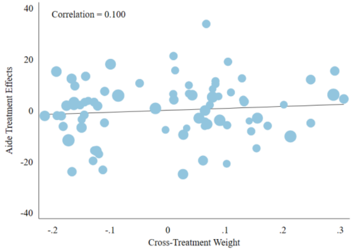
	A. Contamination Bias Estimates				
	Regression Coefficient	Own Effect	Bias	Worst-Case Bias	
				Negative	Positive
	(1)	(2)	(3)	(4)	(5)
Small Class Size	5.357	5.202	0.155	-1.654	1.670
	(0.778)	(0.778)	(0.160)	(0.185)	(0.187)
Teaching Aide	0.177	0.360	-0.183	-1.529	1.530
	(0.720)	(0.714)	(0.149)	(0.176)	(0.177)

	B. Treatment Effect Estimates		
	Unweighted	Efficiently-Weighted	
	(ATE)	One-at-a-time	Common
	(1)	(2)	(3)
Small Class Size	5.561	5.295	5.563
	(0.763)	(0.775)	(0.764)
	[0.744]	[0.743]	[0.742]
Teaching Aide	0.070	0.263	-0.003
	(0.708)	(0.715)	(0.712)
	[0.694]	[0.691]	[0.695]

STAR Regression Weights vs. Treatment Effects

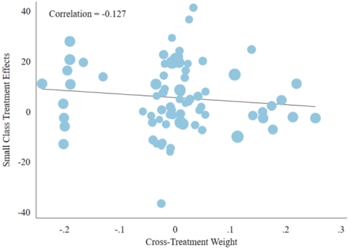


Panel A: Small Class
Own-Treatment Weight



Panel B: Aide
Cross-Treatment Weight

Panel C: Aide
Own-Treatment Weight



Panel D: Small Class
Cross-Treatment Weight

Does Contamination Bias Ever Matter?

- On the "advice" of a referee, we added eight empirical applications:
  - Five stratified RCTs, like STAR
  - Three observational apps (analyses of multiple racial disparities)

Does Contamination Bias Ever Matter?

- On the "advice" of a referee, we added eight empirical applications:
  - Five stratified RCTs, like STAR
  - Three observational apps (analyses of multiple racial disparities)
- Key finding: virtually no contamination bias in the experiments, but significant bias in 2/3rds of the observational regressions
  - Intuitively, experimental strata are unlikely to strongly predict TE heterogeneity (variation driven by experimenter constraints, etc.)

Does Contamination Bias Ever Matter?

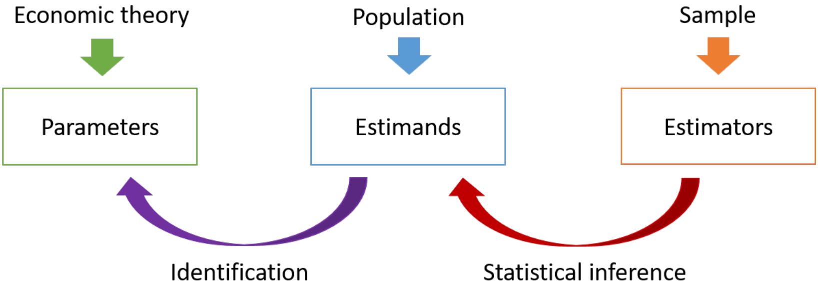
- On the "advice" of a referee, we added eight empirical applications:
  - Five stratified RCTs, like STAR
  - Three observational apps (analyses of multiple racial disparities)
- Key finding: virtually no contamination bias in the experiments, but significant bias in 2/3rds of the observational regressions
  - Intuitively, experimental strata are unlikely to strongly predict TE heterogeneity (variation driven by experimenter constraints, etc.)
- Practical takeaway: bias diagnostics can be useful, especially in observational analyses (use our *multe* package!)

Outline

1. Heterogeneous Treatment Effects✓

2. Clustered Standard Errors

Journey to the Red Arrow...



OLS Asymptotics: Review

- Where do SEs come from? OLS $\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$ can be rewritten:

$$\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{\mathbf{x}'\mathbf{x}}{N}\right)^{-1} \left(\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{N}}\right)$$

where $\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\varepsilon}$ stacks observations of the population regression

OLS Asymptotics: Review

- Where do SEs come from? OLS $\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$ can be rewritten:

$$\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{\mathbf{x}'\mathbf{x}}{N}\right)^{-1}\left(\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{N}}\right)$$

where $\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\varepsilon}$ stacks observations of the population regression

  - Under rather mild conditions (a LLN), $\frac{\mathbf{x}'\mathbf{x}}{N} \xrightarrow{p} E\left[\frac{1}{N}\sum_i x_i x_i'\right]$

OLS Asymptotics: Review

- Where do SEs come from? OLS $\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$ can be rewritten:

$$\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{\mathbf{x}'\mathbf{x}}{N}\right)^{-1} \left(\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{N}}\right)$$

where $\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\varepsilon}$ stacks observations of the population regression

  - Under rather mild conditions (a LLN), $\frac{\mathbf{x}'\mathbf{x}}{N} \xrightarrow{p} E\left[\frac{1}{N}\sum_i x_i x_i'\right]$
  - W/slightly stronger conditions (a CLT), $\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{N}} \Rightarrow \mathrm{N}(0, Var(\frac{1}{\sqrt{N}}\sum_i x_i \varepsilon_i))$

OLS Asymptotics: Review

- Where do SEs come from? OLS $\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$ can be rewritten:

$$\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{\mathbf{x}'\mathbf{x}}{N}\right)^{-1}\left(\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{N}}\right)$$

  where $\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\varepsilon}$ stacks observations of the population regression
  - Under rather mild conditions (a LLN), $\frac{\mathbf{x}'\mathbf{x}}{N} \xrightarrow{p} E\left[\frac{1}{N}\sum_i x_i x_i'\right]$
  - W/slightly stronger conditions (a CLT), $\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{N}} \Rightarrow \mathrm{N}(0, Var(\frac{1}{\sqrt{N}}\sum_i x_i \varepsilon_i))$

- This gives our general asymptotic approximation for OLS: $\hat{\beta} \approx \beta^*$ for

$$\beta^* \sim \mathrm{N}(\beta, V/N), \quad V = E\left[\frac{1}{N}\sum_i x_i x_i'\right]^{-1} Var\left(\frac{1}{\sqrt{N}}\sum_i x_i \varepsilon_i\right) E\left[\frac{1}{N}\sum_i x_i x_i'\right]^{-1}$$

OLS Asymptotics: Review

- Where do SEs come from? OLS $\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$ can be rewritten:

$$\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{\mathbf{x}'\mathbf{x}}{N}\right)^{-1} \left(\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{N}}\right)$$

  where $\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\varepsilon}$ stacks observations of the population regression
  - Under rather mild conditions (a LLN), $\frac{\mathbf{x}'\mathbf{x}}{N} \xrightarrow{p} E\left[\frac{1}{N}\sum_i x_i x_i'\right]$
  - W/slightly stronger conditions (a CLT), $\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{N}} \Rightarrow \mathrm{N}(0, Var(\frac{1}{\sqrt{N}}\sum_i x_i \varepsilon_i))$

- This gives our general asymptotic approximation for OLS: $\hat{\beta} \approx \beta^*$ for

$$\beta^* \sim \mathrm{N}(\beta, V/N), \quad V = E\left[\frac{1}{N}\sum_i x_i x_i'\right]^{-1} Var\left(\frac{1}{\sqrt{N}}\sum_i x_i \varepsilon_i\right) E\left[\frac{1}{N}\sum_i x_i x_i'\right]^{-1}$$

- SEs come from $\hat{V} = \left(\frac{1}{N}\sum_i x_i x_i'\right)^{-1} \widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i x_i \varepsilon_i\right) \left(\frac{1}{N}\sum_i x_i x_i'\right)^{-1}$

Getting to the Meat of the "Sandwich Estimator," $\hat{V}$

- Key q: how do we form the variance estimate $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i x_i \varepsilon_i\right)$?

Getting to the Meat of the "Sandwich Estimator," $\hat{V}$

- Key q: how do we form the variance estimate $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right)$?

- In *iid* data, we know $Var\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right) =$

Getting to the Meat of the "Sandwich Estimator," $\hat{V}$

- Key q: how do we form the variance estimate $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right)$?

- In *iid* data, we know $Var\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right) = \frac{1}{N}\sum_i Var(x_i\varepsilon_i) =$

Getting to the Meat of the "Sandwich Estimator," $\hat{V}$

- Key q: how do we form the variance estimate $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right)$?

- In *iid* data, we know $Var\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right) = \frac{1}{N}\sum_i Var(x_i\varepsilon_i) = E[x_i x_i'\varepsilon_i^2]$

Getting to the Meat of the "Sandwich Estimator," $\hat{V}$

- Key q: how do we form the variance estimate $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right)$?

- In *iid* data, we know $Var\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right) = \frac{1}{N}\sum_i Var(x_i\varepsilon_i) = E[x_ix_i'\varepsilon_i^2]$
  - This suggests $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i^2\right) = \frac{1}{N}\sum_i x_ix_i'\hat{\varepsilon}_i^2$ for $\hat{\varepsilon}_i = y_i - x\hat{\beta}$, which leads to our usual heteroskedasticity-robust estimator

Getting to the Meat of the "Sandwich Estimator," $\hat{V}$

- Key q: how do we form the variance estimate $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right)$?

- In *iid* data, we know $Var\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right) = \frac{1}{N}\sum_i Var(x_i\varepsilon_i) = E[x_i x_i' \varepsilon_i^2]$
  - This suggests $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i^2\right) = \frac{1}{N}\sum_i x_i x_i' \hat{\varepsilon}_i^2$ for $\hat{\varepsilon}_i = y_i - x\hat{\beta}$, which leads to our usual heteroskedasticity-robust estimator

- The motivation for alternative estimators comes from the possibility that $x_i\varepsilon_i$ and $x_j\varepsilon_j$ may be correlated for $i \neq j$
  - Generally, $Var\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right) = \frac{1}{N}\sum_i Var(x_i\varepsilon_i) + 2\sum_{i,j\neq i} Cov(x_i\varepsilon_i, x_j\varepsilon_j)$

Getting to the Meat of the "Sandwich Estimator," $\hat{V}$

- Key q: how do we form the variance estimate $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right)$?

- In *iid* data, we know $Var\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right) = \frac{1}{N}\sum_i Var(x_i\varepsilon_i) = E[x_i x_i' \varepsilon_i^2]$
  - This suggests $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i^2\right) = \frac{1}{N}\sum_i x_i x_i' \hat{\varepsilon}_i^2$ for $\hat{\varepsilon}_i = y_i - x\hat{\beta}$, which leads to our usual heteroskedasticity-robust estimator

- The motivation for alternative estimators comes from the possibility that $x_i\varepsilon_i$ and $x_j\varepsilon_j$ may be correlated for $i \neq j$
  - Generally, $Var\left(\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i\right) = \frac{1}{N}\sum_i Var(x_i\varepsilon_i) + 2\sum_{i,j\neq i} Cov(x_i\varepsilon_i, x_j\varepsilon_j)$
  - But we can't allow for arbitrary cross-sectional correlations, since then we couldn't guarantee $\frac{1}{\sqrt{N}}\sum_i x_i\varepsilon_i$ converges ...

Getting to the Meat of the “Sandwich Estimator,” $\hat{V}$

- Key q: how do we form the variance estimate $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i x_i \varepsilon_i\right)$?

- In *iid* data, we know $Var\left(\frac{1}{\sqrt{N}}\sum_i x_i \varepsilon_i\right) = \frac{1}{N}\sum_i Var(x_i \varepsilon_i) = E[x_i x_i' \varepsilon_i^2]$
    - This suggests $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i x_i \varepsilon_i^2\right) = \frac{1}{N}\sum_i x_i x_i' \hat{\varepsilon}_i^2$ for $\hat{\varepsilon}_i = y_i - x\hat{\beta}$, which leads to our usual heteroskedasticity-robust estimator

- The motivation for alternative estimators comes from the possibility that $x_i \varepsilon_i$ and $x_j \varepsilon_j$ may be correlated for $i \neq j$
    - Generally, $Var\left(\frac{1}{\sqrt{N}}\sum_i x_i \varepsilon_i\right) = \frac{1}{N}\sum_i Var(x_i \varepsilon_i) + 2\sum_{i,j\neq i} Cov(x_i \varepsilon_i, x_j \varepsilon_j)$
    - But we can't allow for arbitrary cross-sectional correlations, since then we couldn't guarantee $\frac{1}{\sqrt{N}}\sum_i x_i \varepsilon_i$ converges ...
    - We need to zero out some covariances to make progress

Cluster-Robust Estimators

- Suppose we can partition observations into clusters, $c(i) \in 1, \ldots, C$
  - To ease notation, suppose equal sizes: $|i : c(i) = c| = N/C \equiv T$

Cluster-Robust Estimators

- Suppose we can partition observations into clusters, $c(i) \in 1, \ldots, C$
  - To ease notation, suppose equal sizes: $|i : c(i) = c| = N/C \equiv T$
  - With $N = CT$, OLS can be rewritten: $\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{\mathbf{x}'\mathbf{x}}{N}\right)^{-1} \cdot \left(\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{CT}}\right)$

Cluster-Robust Estimators

- Suppose we can partition observations into clusters, $c(i) \in 1, \ldots, C$
  - To ease notation, suppose equal sizes: $|i : c(i) = c| = N/C \equiv T$
  - With $N = CT$, OLS can be rewritten: $\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{\mathbf{x}'\mathbf{x}}{N}\right)^{-1} \cdot \left(\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{CT}}\right)$
- Define $q_c = \frac{1}{\sqrt{T}} \sum_{i:c(i)=c} x_i \varepsilon_i$ and note that $\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{CT}} = \frac{1}{\sqrt{C}} \sum_c q_c$

Cluster-Robust Estimators

- Suppose we can partition observations into clusters, $c(i) \in 1, \ldots, C$
  - To ease notation, suppose equal sizes: $|i : c(i) = c| = N/C \equiv T$
  - With $N = CT$, OLS can be rewritten: $\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{\mathbf{x}'\mathbf{x}}{N}\right)^{-1} \cdot \left(\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{CT}}\right)$
- Define $q_c = \frac{1}{\sqrt{T}} \sum_{i:c(i)=c} x_i \varepsilon_i$ and note that $\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{CT}} = \frac{1}{\sqrt{C}} \sum_c q_c$
  - If the q_c clusters are *iid*, a CLT applies: $\frac{1}{\sqrt{C}} \sum_c q_c \Rightarrow \mathrm{N}(0, Var(q_c))$
  - E.g. in a balanced panel, could have *iid* series $(x_{c1}\varepsilon_{c1} \ldots, x_{cT}\varepsilon_{cT})$

Cluster-Robust Estimators

- Suppose we can partition observations into clusters, $c(i) \in 1, \ldots, C$
  - To ease notation, suppose equal sizes: $|i : c(i) = c| = N/C \equiv T$
  - With $N = CT$, OLS can be rewritten: $\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{\mathbf{x}'\mathbf{x}}{N}\right)^{-1} \cdot \left(\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{CT}}\right)$
- Define $q_c = \frac{1}{\sqrt{T}} \sum_{i:c(i)=c} x_i \varepsilon_i$ and note that $\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{CT}} = \frac{1}{\sqrt{C}} \sum_c q_c$
  - If the q_c clusters are *iid*, a CLT applies: $\frac{1}{\sqrt{C}} \sum_c q_c \Rightarrow \mathrm{N}(0, Var(q_c))$
  - E.g. in a balanced panel, could have *iid* series $(x_{c1}\varepsilon_{c1} \ldots, x_{cT}\varepsilon_{cT})$
- This gives us a new "clustered" variance estimate to plug into $\hat{V}$:

$$\widehat{Var}\left(\frac{1}{\sqrt{N}} \sum_i x_i \varepsilon_i\right) = \frac{1}{C} \sum_c \hat{q}_c^2, \text{ for } \hat{q} = \frac{1}{\sqrt{T}} \sum_{i:c(i)=c} x_i \hat{\varepsilon}_i$$

Cluster-Robust Estimators

- Suppose we can partition observations into clusters, $c(i) \in 1, \ldots, C$
  - To ease notation, suppose equal sizes: $|i : c(i) = c| = N/C \equiv T$
  - With $N = CT$, OLS can be rewritten: $\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{\mathbf{x}'\mathbf{x}}{N}\right)^{-1} \cdot \left(\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{CT}}\right)$

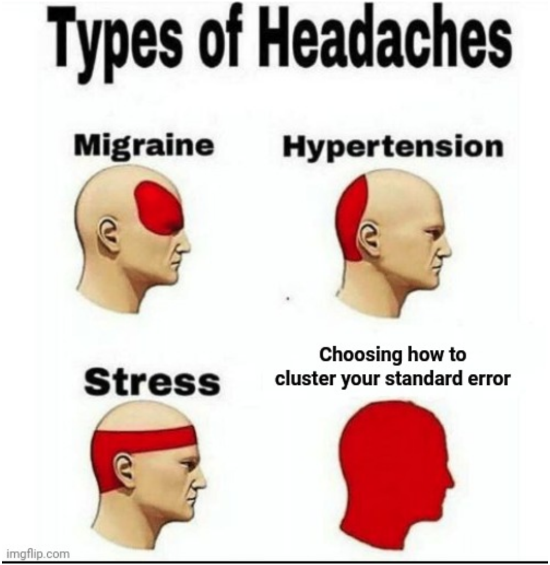
- Define $q_c = \frac{1}{\sqrt{T}} \sum_{i:c(i)=c} x_i \varepsilon_i$ and note that $\frac{\mathbf{x}'\boldsymbol{\varepsilon}}{\sqrt{CT}} = \frac{1}{\sqrt{C}} \sum_c q_c$
  - If the q_c clusters are *iid*, a CLT applies: $\frac{1}{\sqrt{C}} \sum_c q_c \Rightarrow \mathrm{N}(0, Var(q_c))$
  - E.g. in a balanced panel, could have *iid* series $(x_{c1}\varepsilon_{c1} \ldots, x_{cT}\varepsilon_{cT})$

- This gives us a new "clustered" variance estimate to plug into $\hat{V}$:

$$\widehat{Var}\left(\frac{1}{\sqrt{N}} \sum_i x_i \varepsilon_i\right) = \frac{1}{C} \sum_c \hat{q}_c^2, \text{ for } \hat{q} = \frac{1}{\sqrt{T}} \sum_{i:c(i)=c} x_i \hat{\varepsilon}_i$$

- This is what's going on under the hood when you ", *cluster(c)*"!

Easy, Right?



Source: Khoa Vu

Design Can Help!

- At an (unhelpfully) high level, the previous results tell us when to cluster i and j together: when we think $Cov(x_i\varepsilon_i, x_j\varepsilon_j) \neq 0$

Design Can Help!

- At an (unhelpfully) high level, the previous results tell us when to cluster i and j together: when we think $Cov(x_i\varepsilon_i, x_j\varepsilon_j) \neq 0$
- With design, however, this may not be too hard to figure out:
  - Suppose $(x_1, \ldots, x_N) \mid (\varepsilon_1, \ldots, \varepsilon_N)$ is mean-zero with $x_i \perp\!\!\!\perp x_j$ whenever $c(i) \neq c(j)$ (e.g. village-level RCT with $c(i)$ giving i's village)

Design Can Help!

- At an (unhelpfully) high level, the previous results tell us when to cluster i and j together: when we think $Cov(x_i\varepsilon_i, x_j\varepsilon_j) \neq 0$

- With design, however, this may not be too hard to figure out:
  - Suppose $(x_1, \ldots, x_N) \mid (\varepsilon_1, \ldots, \varepsilon_N)$ is mean-zero with $x_i \perp\!\!\!\perp x_j$ whenever $c(i) \neq c(j)$ (e.g. village-level RCT with $c(i)$ giving i's village)
  - Then whenever $c(i) \neq c(j)$:

$$Cov(x_i\varepsilon_i, x_j\varepsilon_j) = E[x_i x_j' \varepsilon_i \varepsilon_j] =$$

Design Can Help!

- At an (unhelpfully) high level, the previous results tell us when to cluster i and j together: when we think $Cov(x_i\varepsilon_i, x_j\varepsilon_j) \neq 0$

- With design, however, this may not be too hard to figure out:
  - Suppose $(x_1,\ldots,x_N) \mid (\varepsilon_1,\ldots,\varepsilon_N)$ is mean-zero with $x_i \perp\!\!\!\perp x_j$ whenever $c(i) \neq c(j)$ (e.g. village-level RCT with $c(i)$ giving i's village)
  - Then whenever $c(i) \neq c(j)$:

$$Cov(x_i\varepsilon_i, x_j\varepsilon_j) = E[x_i x_j' \varepsilon_i \varepsilon_j] = E[E[x_i x_j' \mid \varepsilon_i, \varepsilon_j]\varepsilon_i \varepsilon_j] =$$

Design Can Help!

- At an (unhelpfully) high level, the previous results tell us when to cluster i and j together: when we think $Cov(x_i\varepsilon_i, x_j\varepsilon_j) \neq 0$
- With design, however, this may not be too hard to figure out:
  - Suppose $(x_1, \ldots, x_N) \mid (\varepsilon_1, \ldots, \varepsilon_N)$ is mean-zero with $x_i \perp\!\!\!\perp x_j$ whenever $c(i) \neq c(j)$ (e.g. village-level RCT with $c(i)$ giving i's village)
  - Then whenever $c(i) \neq c(j)$:

$$Cov(x_i\varepsilon_i, x_j\varepsilon_j) = E[x_i x_j' \varepsilon_i \varepsilon_j] = E[E[x_i x_j' \mid \varepsilon_i, \varepsilon_j]\varepsilon_i \varepsilon_j] = 0$$

Design Can Help!

- At an (unhelpfully) high level, the previous results tell us when to cluster i and j together: when we think $Cov(x_i\varepsilon_i, x_j\varepsilon_j) \neq 0$
- With design, however, this may not be too hard to figure out:
  - Suppose $(x_1, \ldots, x_N) \mid (\varepsilon_1, \ldots, \varepsilon_N)$ is mean-zero with $x_i \perp\!\!\!\perp x_j$ whenever $c(i) \neq c(j)$ (e.g. village-level RCT with $c(i)$ giving i's village)
  - Then whenever $c(i) \neq c(j)$:

$$Cov(x_i\varepsilon_i, x_j\varepsilon_j) = E[x_i x_j' \varepsilon_i \varepsilon_j] = E[E[x_i x_j' \mid \varepsilon_i, \varepsilon_j]\varepsilon_i \varepsilon_j] = 0$$

  - So we only need to cluster by $c(i)$: the design tells us what to do!

Design Can Help!

- At an (unhelpfully) high level, the previous results tell us when to cluster i and j together: when we think $Cov(x_i\varepsilon_i, x_j\varepsilon_j) \neq 0$
- With design, however, this may not be too hard to figure out:
  - Suppose $(x_1, \ldots, x_N) \mid (\varepsilon_1, \ldots, \varepsilon_N)$ is mean-zero with $x_i \perp\!\!\!\perp x_j$ whenever $c(i) \neq c(j)$ (e.g. village-level RCT with $c(i)$ giving i's village)
  - Then whenever $c(i) \neq c(j)$:

$$Cov(x_i\varepsilon_i, x_j\varepsilon_j) = E[x_i x_j' \varepsilon_i \varepsilon_j] = E[E[x_i x_j' \mid \varepsilon_i, \varepsilon_j]\varepsilon_i \varepsilon_j] = 0$$

  - So we only need to cluster by $c(i)$: the design tells us what to do!
- This leads to the popular (and sometimes misused) heuristic: cluster at the level of treatment / identifying variation
  - See Abadie et al. (2023) for a more complete version of this argument

Where Intuition Can Fall Short: Paired Randomization

- Suppose (as is often done) we pair individuals up by some baseline characteristics, then in each pair c we randomly treat one individual
  - Treatment is at the individual level... so should we just ", r" ?

Where Intuition Can Fall Short: Paired Randomization

- Suppose (as is often done) we pair individuals up by some baseline characteristics, then in each pair c we randomly treat one individual
  - Treatment is at the individual level... so should we just ", r" ?
- de Chaisemartin and Ramirez-Cuellar (2022) show the answer is no: non-clustered SEs will generally be downward-biased (maybe badly)
  - Under constant effects, $E[\hat{V}] = V/2$; severe over-rejection!

Where Intuition Can Fall Short: Paired Randomization

- Suppose (as is often done) we pair individuals up by some baseline characteristics, then in each pair c we randomly treat one individual
  - Treatment is at the individual level... so should we just ", r" ?
- de Chaisemartin and Ramirez-Cuellar (2022) show the answer is no: non-clustered SEs will generally be downward-biased (maybe badly)
  - Under constant effects, $E[\hat{V}] = V/2$; severe over-rejection!
- Paired randomization makes x_i and x_j *negatively* correlated in pairs
  - Clustering by pair solves this; treatment assignment is *iid* across pairs