Lecture 4: Multiple Regression and Causal Effects

POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

February 2, 2022

Last time:

- Agnostic regression and "honest" approximation inference.
- When we need to model but recognize that we can't model perfectly.

Today we address:

- Potential outcomes and regression.
- Regression estimates vs. ATE, ATT, or ATC.
- Implications of effect heterogeneity.
- Understanding linearity.

Potential outcomes and causal effects

- Consider a random draw, i, from P, large.
- Each draw is characterized by a covariate vector, X_i, potential outcomes that under SUTVA are characterized as Y_{di} for all $d \in \mathscr{D}$, as well as a treatment assignments, $D_i \in \mathscr{D}$.

Potential outcomes and causal effects

- Suppose $\mathscr{D} = \{0, 1\}$. Then under SUTVA, potential outcomes for an arbitrary draw from P are Y_{1i} and Y_{0i}.
- A unit level treatment effect for an arbitrary draw from P is, $\rho_i = Y_{1i} - Y_{0i}$, for which $E[\rho_i] = E[Y_{1i} - Y_{0i}] = \rho$, is the average treatment effect (ATE).
- For an arbitrary draw from P, we observe X_i and,

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{i0} = Y_{i0} + (Y_{1i} - Y_{0i}) D_i.$$

Potential outcomes and causal effects

Note that

$$
\begin{aligned}
Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\
&\qquad\qquad \text{(add and subtract } E[Y_{0i}] \text{ and } E[Y_{1i}]\text{, rearrange)} \\
&= \underbrace{E[Y_{0i}]}_{\beta_0} + \underbrace{D_i E[Y_{1i} - Y_{0i}]}_{+ \qquad D_i \beta_1} + \underbrace{D_i (Y_{1i} - E[Y_{1i}]) + (1 - D_i)(Y_{0i} - E[Y_{0i}])}_{+ \qquad \varepsilon_i}
\end{aligned}
$$

or

$$
\begin{aligned}
&= \underbrace{E[Y_{0i}]}_{\beta_0} + \underbrace{D_i E[Y_{1i} - Y_{0i}]}_{+ \qquad D_i \beta_1} \\
&\qquad + \underbrace{(Y_{0i} - E[Y_{0i}]) + D_i[(Y_{1i} - Y_{0i}) - (E[Y_{1i}] - E[Y_{0i}])]}_{+ \qquad \varepsilon_i}
\end{aligned}
$$

Potential outcomes and causal effects

- ε_i can be interpreted as (i) heterogeneity in potential outcomes, or (ii) heterogeneity in baseline potential outcomes plus effect heterogeneity.
- D_i is random (different than classical regression).
- Effect heterogeneity implies heteroskedasticity assumption needed on ε_i, because error variance differs over D_i.
- Equivalency means that we can retain much regression theory and intuitions while being "agnostic" about the nature of causal effects (e.g. we don't have to assume homogenous effects).
- Generalizations to multivalued treatments are straightforward (either dose-response functions or a bunch of binary contrasts).

Potential outcomes and causal effects

- Given (Y_i, X_i), the following allow ρ to be identified:

(random assignment) $D_i \perp\!\!\!\perp (Y_{1i}, Y_{0i})$ and $0 < Pr[D_i = 1] < 1$ for all $i \in U$

or

Potential outcomes and causal effects

- Given (Y_i, X_i), the following allow ρ to be identified:

(random assignment) $D_i \perp\!\!\!\perp (Y_{1i}, Y_{0i})$ and $0 < Pr[D_i = 1] < 1$ for all $i \in U$

or

(*conditional* r.a.*) $D_i \perp\!\!\!\perp (Y_{1i}, Y_{0i}) | X_i$ and $0 < Pr[D_i = 1 | X_i = x] < 1$ for all $x \in \mathscr{X}$

*Angrist & Pischke: "conditional independence assumption" (CIA).

Potential outcomes and causal effects

- Recall a regression model may be interpreted as follows:

$$Y_i = \underbrace{E[Y_{0i}]}_{\beta_0} + \underbrace{D_i E[Y_{1i} - Y_{0i}]}_{+ \qquad D_i\beta_1} + \underbrace{D_i(Y_{1i} - E[Y_{1i}]) + (1 - D_i)(Y_{0i} - E[Y_{0i}])}_{+ \qquad \varepsilon_i}$$

or

$$Y_i = \underbrace{E[Y_{0i}]}_{\beta_0} + \underbrace{D_i E[Y_{1i} - Y_{0i}]}_{+ \qquad D_i\beta_1} + \underbrace{(Y_{0i} - E[Y_{0i}]) + D_i[(Y_{1i} - Y_{0i}) - (E[Y_{1i}] - E[Y_{0i}])]}_{+ \qquad \varepsilon_i}$$

Potential outcomes and causal effects

- Recall a regression model may be interpreted as follows:

$$Y_i = \underbrace{E[Y_{0i}]}_{\beta_0} + \underbrace{D_i E[Y_{1i} - Y_{0i}]}_{+ \quad D_i\beta_1} + \underbrace{D_i(Y_{1i} - E[Y_{1i}]) + (1 - D_i)(Y_{0i} - E[Y_{0i}])}_{+ \quad \varepsilon_i}$$

or

$$Y_i = \underbrace{E[Y_{0i}]}_{\beta_0} + \underbrace{D_i E[Y_{1i} - Y_{0i}]}_{+ \quad D_i\beta_1} + \underbrace{(Y_{0i} - E[Y_{0i}]) + D_i[(Y_{1i} - Y_{0i}) - (E[Y_{1i}] - E[Y_{0i}])]}_{+ \quad \varepsilon_i}$$

- Under random assignment, fitting this bivariate model via OLS is unbiased.

Potential outcomes and causal effects

- Recall a regression model may be interpreted as follows:

$$Y_i = \underbrace{E[Y_{0i}]}_{\beta_0} + \underbrace{D_i E[Y_{1i} - Y_{0i}]}_{+ \quad D_i \beta_1} + \underbrace{D_i(Y_{1i} - E[Y_{1i}]) + (1 - D_i)(Y_{0i} - E[Y_{0i}])}_{+ \quad \varepsilon_i}$$

or

$$Y_i = \underbrace{E[Y_{0i}]}_{\beta_0} + \underbrace{D_i E[Y_{1i} - Y_{0i}]}_{+ \quad D_i \beta_1} + \underbrace{(Y_{0i} - E[Y_{0i}]) + D_i[(Y_{1i} - Y_{0i}) - (E[Y_{1i}] - E[Y_{0i}])]}_{+ \quad \varepsilon_i}$$

- Under random assignment, fitting this bivariate model via OLS is unbiased.
- Including X_i would be for *efficiency* when we have random assignment.

When is CIA plausible?

- Under CIA, the situation is different.
- The bivariate regression may not be unbiased or consistent. We need to “control for X_i.”

When is CIA plausible?

- Under CIA, the situation is different.
- The bivariate regression may not be unbiased or consistent. We need to “control for X_i.”
- But before going there, ask, “is CIA plausible in this case?”

When is CIA plausible?

- When you apply CIA, you should be able to answer the question,

  How could it be that two units that are *identical* in all meaningful background characteristics nonetheless receive *different* treatment?

When is CIA plausible?

- When you apply CIA, you should be able to answer the question,

  How could it be that two units that are
  identical in all meaningful background characteristics
  nonetheless receive *different* treatment?

- CIA requires that such things *can* happen. Why?
- The answer should point to something "arbitrary" with respect to the outcomes of interest, e.g.
  - A lottery-type process,
  - Administrative flukes,
  - Path dependencies in allocation of treatments, with targeting of units based on observable characteristics,
  - Leadership idiosyncrasies,
  - etc.

Generalizing potential outcomes

- Let's now turn to a more general characterization of regression under potential outcomes using a "response function."

Generalizing potential outcomes

- Let's now turn to a more general characterization of regression under potential outcomes using a "response function."
- Let $Y_{di} \equiv f_i(d)$ and average causal effect of going from $d - v$ to d be $\mathrm{E}[f_i(d) - f_i(d - v)]$.

Generalizing potential outcomes

- We can generalize the CIA to this setting,

$$D_i \perp\!\!\!\perp f_i(d) | X_i \text{ and } 0 < p_i(d) < 1 \text{ for all } d \in \mathscr{D} \tag{1}$$

Generalizing potential outcomes

- We can generalize the CIA to this setting,

$$D_i \perp\!\!\!\perp f_i(d) | X_i \text{ and } 0 < p_i(d) < 1 \text{ for all } d \in \mathscr{D} \tag{1}$$

- "Randomly assigning dosages within strata defined by X_i values."
- Conditional average causal effect: $\mathrm{E}\left[f_i(d) - f_i(d-v) | X_i\right]$.

Generalizing potential outcomes

- We can generalize the CIA to this setting,

$$D_i \perp\!\!\!\perp f_i(d)|X_i \text{ and } 0 < p_i(d) < 1 \text{ for all } d \in \mathscr{D} \tag{1}$$

- "Randomly assigning dosages within strata defined by X_i values."
- Conditional average causal effect: $\mathrm{E}[f_i(d) - f_i(d-v)|X_i]$.
- By (1),

$$\begin{aligned}
\mathrm{E}[Y_i|X_i, D_i = d] - \mathrm{E}[Y_i|X_i, D_i = d - v] \\
&= \mathrm{E}[f_i(d)|X_i, D_i = d] - \mathrm{E}[f_i(d-v)|X_i, D_i = d - v] \\
&= \mathrm{E}[f_i(d) - f_i(d-v)|X_i].
\end{aligned}$$

- Averaging over X_i,
$$\mathrm{E}_X\{\mathrm{E}[f_i(d) - f_i(d-v)|X_i]\} = \mathrm{E}[f_i(d) - f_i(d-v)].$$

Generalizing potential outcomes

- Other average causal effects are possible.
- E.g., could compute,

$$\mathrm{E}_{X|D\in\mathscr{D}'}[\mathrm{E}[f_i(D_i)-f_i(0)|X_i]|D_i\in\mathscr{D}'] = \mathrm{E}_{\in\mathscr{D}'}[f_i(D_i)-f_i(0)|D_i\in\mathscr{D}'],$$

Generalizing potential outcomes

- Other average causal effects are possible.
- E.g., could compute,

$$\mathrm{E}_{X|D\in\mathscr{D}'}[\mathrm{E}[f_i(D_i)-f_i(0)|X_i]|D_i\in\mathscr{D}'] = \mathrm{E}_{\in\mathscr{D}'}[f_i(D_i)-f_i(0)|D_i\in\mathscr{D}'],$$

- E.g., with $\mathscr{D}' = \{d : d > 0\}$ this would be the average effect of changing dose to zero for those who had been subject to some positive dose.
- Known as the "attributable risk" effect in epidemiology.

Generalizing potential outcomes

- Other average causal effects are possible.
- E.g., could compute,

$$\mathrm{E}_{X|D\in\mathscr{D}'}[\mathrm{E}[f_i(D_i)-f_i(0)|X_i]|D_i\in\mathscr{D}']=\mathrm{E}_{\in\mathscr{D}'}[f_i(D_i)-f_i(0)|D_i\in\mathscr{D}'],$$

- E.g., with $\mathscr{D}'=\{d:d>0\}$ this would be the average effect of changing dose to zero for those who had been subject to some positive dose.
- Known as the "attributable risk" effect in epidemiology.
- The point here is: be clear about the causal effects you estimating. Be clear about the *counterfactuals* they reference and for *what subpopulation* they are defined.

From general potential outcomes & CIA to regression

- Now, let's start with a simplified analysis based on assumptions of constant linear effects and linearly separable confounding .
- This aligns our analysis with classical regression, although *somewhat* more agnostic.
- Then, we will consider what happens when we loosen these assumptions (fully agnostic).

From general potential outcomes & CIA to regression

- Suppose a simplistic causal model, $f_i(d) = \alpha + \rho d + \eta_i$.
- We thus observe, $Y_i = f_i(D_i) = \alpha + \rho D_i + \eta_i$.
- η_i represents all variable determinants of $f_i(D_i)$ other than D_i.

From general potential outcomes & CIA to regression

- Suppose a simplistic causal model, $f_i(d) = \alpha + \rho d + \eta_i$.
- We thus observe, $Y_i = f_i(D_i) = \alpha + \rho D_i + \eta_i$.
- η_i represents all variable determinants of $f_i(D_i)$ other than D_i.
- Let $\eta_i = X_i'\gamma + \nu_i$, such that γ is the population OLS solution.
- Then, $\mathrm{E}[X_i \nu_i] = 0$ (they are orthogonal in the population).

From general potential outcomes & CIA to regression

- Suppose a simplistic causal model, $f_i(d) = \alpha + \rho d + \eta_i$.
- We thus observe, $Y_i = f_i(D_i) = \alpha + \rho D_i + \eta_i$.
- η_i represents all variable determinants of $f_i(D_i)$ other than D_i.
- Let $\eta_i = X_i'\gamma + \nu_i$, such that γ is the population OLS solution.
- Then, $\mathrm{E}[X_i \nu_i] = 0$ (they are orthogonal in the population).
- Assume further that linearity holds in X_i, $\mathrm{E}[\eta_i | X_i] = X_i'\gamma$.

From general potential outcomes & CIA to regression

- Suppose a simplistic causal model, $f_i(d) = \alpha + \rho d + \eta_i$.
- We thus observe, $Y_i = f_i(D_i) = \alpha + \rho D_i + \eta_i$.
- η_i represents all variable determinants of $f_i(D_i)$ other than D_i.
- Let $\eta_i = X_i'\gamma + \nu_i$, such that γ is the population OLS solution.
- Then, $\mathrm{E}[X_i \nu_i] = 0$ (they are orthogonal in the population).
- Assume further that linearity holds in X_i, $\mathrm{E}[\eta_i | X_i] = X_i'\gamma$.
- When we hold X_i fixed, the only thing that varies in η_i is ν_i.
- Therefore under this model, $D_i \perp\!\!\!\perp f_i(d) | X_i$ implies $D_i \perp\!\!\!\perp \nu_i | X_i$.

From general potential outcomes & CIA to regression

- Suppose a simplistic causal model, $f_i(d) = \alpha + \rho d + \eta_i$.
- We thus observe, $Y_i = f_i(D_i) = \alpha + \rho D_i + \eta_i$.
- η_i represents all variable determinants of $f_i(D_i)$ other than D_i.
- Let $\eta_i = X_i'\gamma + \nu_i$, such that γ is the population OLS solution.
- Then, $\mathrm{E}[X_i \nu_i] = 0$ (they are orthogonal in the population).
- Assume further that linearity holds in X_i, $\mathrm{E}[\eta_i | X_i] = X_i'\gamma$.
- When we hold X_i fixed, the only thing that varies in η_i is ν_i.
- Therefore under this model, $D_i \perp\!\!\!\perp f_i(d) | X_i$ implies $D_i \perp\!\!\!\perp \nu_i | X_i$.
- In other words, under CIA and the modeling assumptions on η_i, we have

  $$Y_i = f_i(D_i) = \alpha + \rho D_i + X_i'\gamma + \nu_i,$$

  where ν_i is uncorrelated with X_i, and ν_i is uncorrelated with D_i conditional on X_i.

From general potential outcomes & CIA to regression

- By CIA, $\mathrm{E}[f_i(d)|D_i = d, X_i] = \mathrm{E}[f_i(d)|X_i] = \alpha + \rho d + X_i'\gamma$, in which case,

$$\begin{aligned}
\mathrm{E}[f_i(d)|D_i = d, X_i] - \mathrm{E}[f_i(d)|D_i = d - v, X_i] &= \mathrm{E}[f_i(d) - f_i(d - v)|X_i] \\
&= (\alpha + \rho d + X_i'\gamma) - (\alpha + \rho(d - v) + X_i'\gamma) \\
&= \rho v
\end{aligned}$$

- (X_i disappeared because of the linearly separable confounding.)

From general potential outcomes & CIA to regression

- By CIA, $\mathrm{E}[f_i(d)|D_i = d, X_i] = \mathrm{E}[f_i(d)|X_i] = \alpha + \rho d + X_i'\gamma$, in which case,

$$\begin{aligned}\mathrm{E}[f_i(d)|D_i = d, X_i] - \mathrm{E}[f_i(d)|D_i = d - v, X_i] &= \mathrm{E}[f_i(d) - f_i(d - v)|X_i] \\ &= (\alpha + \rho d + X_i'\gamma) - (\alpha + \rho(d - v) + X_i'\gamma) \\ &= \rho v\end{aligned}$$

- (X_i disappeared because of the linearly separable confounding.)
- So ρ in the regression, $Y_i = \alpha + \rho D_i + X_i'\gamma + v_i$, estimates the *causal effect* of a unit change in d.
- Since v_i is uncorrelated with X_i and D_i, OLS is consistent for ρ.

Omitted variable bias

- Suppose we omit X_i, and just regress Y_i on D_i via OLS.

Omitted variable bias

- Suppose we omit X_i, and just regress Y_i on D_i via OLS.
- Then, the coefficient on D_i estimates, $\frac{\mathrm{Cov}(Y_i, D_i)}{\mathrm{Var}(D_i)}$, where,

$$\begin{aligned}
\mathrm{Cov}(Y_i, D_i) &= \mathrm{Cov}(\alpha + \rho D_i + X_i'\gamma + \nu_i, D_i) \\
&= \rho \mathrm{Cov}(D_i, D_i) + \mathrm{Cov}(X_{1i}\gamma_1 + \ldots + X_{Ki}\gamma_K, D_i) \\
&= \rho \mathrm{Var}(D_i) + \gamma_1 \mathrm{Cov}(X_{1i}, D_i) + \ldots + \gamma_K \mathrm{Cov}(X_{Ki}, D_i),
\end{aligned}$$

and so,

$$\frac{\mathrm{Cov}(Y_i, D_i)}{\mathrm{Var}(D_i)} = \rho + \underbrace{\gamma'\delta}_{\text{"omitted variable bias"}= (X_{ki}, Y_i) \text{ relationships } \times (X_{ki}, D_i) \text{ relationships}},$$

where δ is coefficients from regressions of $X_1, \ldots, X_K$ on D.

- By FWL, we can see what happens if we include part of X_i: $\frac{\mathrm{Cov}(\tilde{Y}_i, \tilde{D}_i)}{\mathrm{Var}(\tilde{D}_i)} = \rho + \tilde{\gamma}'\tilde{\delta}$, where $\tilde{}$ means residualize on subset of X_i.

Omitted variable bias

Let's practice applying the OVB formula:

$$\begin{tikzcd} \color{red}{X} \arrow[rr, red] \arrow[dr, red] & & Y \\ & D \arrow[ur] & \end{tikzcd}$$

"omitted variable bias"$=$ (X_{ki}, Y_i) relationships $\times(X_{ki}, D_i)$ relationships

Omitted variable bias

Let's practice applying the OVB formula:

$$\begin{tikzcd}
X \arrow[rr] \arrow[dr] & & Y \\
& D \arrow[ur] &
\end{tikzcd}$$

"omitted variable bias"$= (X_{ki}, Y_i)$ relationships $\times (X_{ki}, D_i)$ relationships

1. Effect of democratic institutions on growth, estimated via regression of growth on democratic institutions.
2. Effect of exposure to negative advertisements on turnout, estimated via regression of turnout on number of ads seen.

What is a possible omitted variable? How will this bias the estimate?

Omitted variable bias

- "Omitted variables" is a bit misleading because it could suggest that you want to include *any* variable that is correlated with treatment and outcomes.

Omitted variable bias

- "Omitted variables" is a bit misleading because it could suggest that you want to include *any* variable that is correlated with treatment and outcomes.
- This is not so ("bad control"):
  - Controlling for post-treatment variables may induce bias in the estimation of causal effects. (More on this in coming lectures.)
  - Controlling for instruments, which are only correlated with outcomes through their correlation with the treatment, may amplify bias in the estimation of causal effects. (Your homework.)
  - We will cover this in detail in lecture 5.

Omitted variable bias

- “Omitted variables” is a bit misleading because it could suggest that you want to include *any* variable that is correlated with treatment and outcomes.
- This is not so (“bad control”):
  - Controlling for post-treatment variables may induce bias in the estimation of causal effects. (More on this in coming lectures.)
  - Controlling for instruments, which are only correlated with outcomes through their correlation with the treatment, may amplify bias in the estimation of causal effects. (Your homework.)
  - We will cover this in detail in lecture 5.
- Thus, control strategies cannot be pursued mechanically with reference to “correlation between outcomes or treatments”!
- The direction of the causal arrows coming from X in the graph on the previous slide are *crucial*.
- Controlling for the wrong things can *introduce* bias.

Heterogeneity and nonlinearity

- Thus far we have simplified things by assuming constant effects ($\rho_i = \rho$ for all i) and linearity ($\mathrm{E}[\eta_i|X_i] = X_i'\gamma$).
- These are strong assumptions!

Heterogeneity and nonlinearity

- Thus far we have simplified things by assuming constant effects ($\rho_i = \rho$ for all i) and linearity ($\mathrm{E}[\eta_i | X_i] = X_i'\gamma$).
- These are strong assumptions!
- What if they are false?
- "A regression is causal when the CEF it approximates is causal."
- Even misspecified models have *causal interpretations* if they approximate causal CEFs.
- *However*, the coefficients may not estimate what you would ideally like them to estimate.
- Let's see how this works with a binary treatment.

Heterogeneity

- We are back to potential outcomes (Y_{0i}, Y_{1i}), with heterogenous treatment effects, $\rho_i = Y_{1i} - Y_{0i}$.

Heterogeneity

- We are back to potential outcomes (Y_{0i}, Y_{1i}), with heterogenous treatment effects, $\rho_i = Y_{1i} - Y_{0i}$.
- Assume CIA, $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$. Then consider the ATT,

$$
\begin{aligned}
\rho_{ATT} &= \mathrm{E}\left[Y_{1i} - Y_{0i} | D_i = 1\right] \\
&= \mathrm{E}_{X|D=1}\left\{\mathrm{E}\left[Y_{1i} - Y_{0i} | X_i, D_i = 1\right]\right\} \\
&= \mathrm{E}_{X|D=1}\{\underbrace{\mathrm{E}\left[Y_{1i} | X_i, D_i = 1\right]}_{\text{observable}} - \underbrace{\mathrm{E}\left[Y_{0i} | X_i, D_i = 1\right]}_{\text{counterfactual}}\} \\
&= \mathrm{E}_{X|D=1}\{\mathrm{E}\left[Y_{1i} | X_i, D_i = 1\right] - \underbrace{\mathrm{E}\left[Y_{0i} | X_i, {\color{red}D_i = 0}\right]}_{\text{by CIA}}\}.
\end{aligned}
$$

Heterogeneity

- We are back to potential outcomes (Y_{0i}, Y_{1i}), with heterogenous treatment effects, $\rho_i = Y_{1i} - Y_{0i}$.
- Assume CIA, $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$. Then consider the ATT,

$$\begin{aligned}
\rho_{ATT} &= \mathrm{E}\left[Y_{1i} - Y_{0i} | D_i = 1\right] \\
&= \mathrm{E}_{X|D=1}\{\mathrm{E}\left[Y_{1i} - Y_{0i} | X_i, D_i = 1\right]\} \\
&= \mathrm{E}_{X|D=1}\{\underbrace{\mathrm{E}\left[Y_{1i} | X_i, D_i = 1\right]}_{\text{observable}} - \underbrace{\mathrm{E}\left[Y_{0i} | X_i, D_i = 1\right]}_{\text{counterfactual}}\} \\
&= \mathrm{E}_{X|D=1}\{\mathrm{E}\left[Y_{1i} | X_i, D_i = 1\right] - \underbrace{\mathrm{E}\left[Y_{0i} | X_i, D_i = 0\right]}_{\text{by CIA}}\}.
\end{aligned}$$

X_i-specific effects averaged over X_i distribution for treated.

Heterogeneity

- We are back to potential outcomes (Y_{0i}, Y_{1i}), with heterogenous treatment effects, $\rho_i = Y_{1i} - Y_{0i}$.
- Assume CIA, $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$. Then consider the ATT,

$$\begin{aligned}
\rho_{ATT} &= \mathrm{E}[Y_{1i} - Y_{0i} | D_i = 1] \\
&= \mathrm{E}_{X|D=1}\{\mathrm{E}[Y_{1i} - Y_{0i} | X_i, D_i = 1]\} \\
&= \mathrm{E}_{X|D=1}\{\underbrace{\mathrm{E}[Y_{1i} | X_i, D_i = 1]}_{\text{observable}} - \underbrace{\mathrm{E}[Y_{0i} | X_i, D_i = 1]}_{\text{counterfactual}}\} \\
&= \mathrm{E}_{X|D=1}\{\mathrm{E}[Y_{1i} | X_i, D_i = 1] - \underbrace{\mathrm{E}[Y_{0i} | X_i, D_i = 0]}_{\text{by CIA}}\}.
\end{aligned}$$

  X_i-specific effects averaged over X_i distribution for treated.
- Let $\delta_X = \mathrm{E}[Y_{1i} | X_i = x, D_i = 1] - \mathrm{E}[Y_{0i} | X_i = x, D_i = 0]$.

Heterogeneity

- We are back to potential outcomes (Y_{0i}, Y_{1i}), with heterogenous treatment effects, $\rho_i = Y_{1i} - Y_{0i}$.
- Assume CIA, $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$. Then consider the ATT,

$$\begin{aligned}
\rho_{ATT} &= \mathrm{E}\left[Y_{1i} - Y_{0i} | D_i = 1\right] \\
&= \mathrm{E}_{X|D=1}\left\{\mathrm{E}\left[Y_{1i} - Y_{0i} | X_i, D_i = 1\right]\right\} \\
&= \mathrm{E}_{X|D=1}\{\underbrace{\mathrm{E}\left[Y_{1i} | X_i, D_i = 1\right]}_{\text{observable}} - \underbrace{\mathrm{E}\left[Y_{0i} | X_i, D_i = 1\right]}_{\text{counterfactual}}\} \\
&= \mathrm{E}_{X|D=1}\{\mathrm{E}\left[Y_{1i} | X_i, D_i = 1\right] - \underbrace{\mathrm{E}\left[Y_{0i} | X_i, {\color{red}D_i = 0}\right]}_{\text{by CIA}}\}.
\end{aligned}$$

X_i-specific effects averaged over X_i distribution for treated.

- Let $\delta_X = \mathrm{E}\left[Y_{1i} | X_i = x, D_i = 1\right] - \mathrm{E}\left[Y_{0i} | X_i = x, D_i = 0\right]$.
- For X_i discrete, we can construct an unbiased "matching estimator", $\hat{\rho}_{ATT}$, for which

$$\mathrm{E}\left[\hat{\rho}_{ATT}\right] = \sum_x \delta_X \Pr[X_i = x | D_i = 1] = \frac{\sum_x \delta_X \Pr[D_i = 1 | X_i = x] \Pr[X_i = x]}{\sum_x \Pr[D_i = 1 | X_i = x] \Pr[X_i = x]}.$$

Heterogeneity

- Now, suppose we use OLS to estimate,

$$Y_i = \alpha_0 + \delta_R D_i + 1(X_i = x_2)\alpha_{x_1} + ... + 1(X_i = x_L)\alpha_{x_L} + e_i,$$

where $x_1, ..., x_L$ exhausts all possible X_i values (omitting one from the specification). Thus, linearity of the CEF holds.

Heterogeneity

- We can use FWL to express the the OLS estimator for the coefficient on D_i, $\hat{\delta}_R$:

Heterogeneity

- We can use FWL to express the the OLS estimator for the coefficient on D_i, $\hat{\delta}_R$:

$$\hat{\delta}_R = \frac{\sum_{i=1}^{N} Y_i \tilde{D}_i}{\sum_{i=1}^{N} \tilde{D}_i^2} \xrightarrow{a} \frac{\text{Cov}\left(Y_i, \tilde{D}_i\right)}{\text{Var}\left(\tilde{D}_i\right)} = \frac{\sum_x \text{Cov}\left(Y_i, \tilde{D}_i | X_i = x\right) \Pr[X_i = x]}{\sum_x \text{Var}\left(\tilde{D}_i | X_i = x\right) \Pr[X_i = x]}$$

$$= \frac{\sum_x \text{Cov}\left(Y_{0i} + \rho_i D_i, \tilde{D}_i | X_i = x\right) \Pr[X_i = x]}{\sum_x \text{Var}\left(\tilde{D}_i | X_i = x\right) \Pr[X_i = x]}$$

$$= \frac{\sum_x \text{Cov}\left(\rho_i D_i, \tilde{D}_i | X_i = x\right) \Pr[X_i = x]}{\sum_x \text{Var}\left(\tilde{D}_i | X_i = x\right) \Pr[X_i = x]}$$

$$= \frac{\sum_x \text{E}\left(\rho_i D_i \tilde{D}_i | X_i = x\right) \Pr[X_i = x]}{\sum_x \text{Var}\left(\tilde{D}_i | X_i = x\right) \Pr[X_i = x]}$$

$$= \frac{\sum_x \delta_X \text{Var}\left(D_i | X_i = x\right) \Pr[X_i = x]}{\sum_x \text{Var}\left(D_i | X_i = x\right) \Pr[X_i = x]}$$

$$= \frac{\sum_x \delta_X [\Pr[D_i = 1 | X_i = x](1 - \Pr[D_i = 1 | X_i = x])] \Pr[X_i = x]}{\sum_x [\Pr[D_i = 1 | X_i = x](1 - \Pr[D_i = 1 | X_i = x])] \Pr[X_i = x]}.$$

Heterogeneity

So, let's compare:

$$\mathrm{E}\left[\hat{\rho}_{ATT}\right] = \frac{\sum_x \delta_X \Pr[D_i = 1|X_i = x] \Pr[X_i = x]}{\sum_x \Pr[D_i = 1|X_i = x] \Pr[X_i = x]}.$$

versus

$$\hat{\delta}_R \xrightarrow{a} \frac{\sum_x \delta_X [\Pr[D_i = 1|X_i = x](1 - \Pr[D_i = 1|X_i = x])] \Pr[X_i = x]}{\sum_x [\Pr[D_i = 1|X_i = x](1 - \Pr[D_i = 1|X_i = x])] \Pr[X_i = x]}$$

Heterogeneity

So, let's compare:

$$\mathrm{E}\left[\hat{\rho}_{ATT}\right] = \frac{\sum_x \delta_X \Pr[D_i = 1|X_i = x]\Pr[X_i = x]}{\sum_x \Pr[D_i = 1|X_i = x]\Pr[X_i = x]}.$$

versus

$$\hat{\delta}_R \xrightarrow{a} \frac{\sum_x \delta_X [\Pr[D_i = 1|X_i = x](1 - \Pr[D_i = 1|X_i = x])]\Pr[X_i = x]}{\sum_x [\Pr[D_i = 1|X_i = x](1 - \Pr[D_i = 1|X_i = x])]\Pr[X_i = x]}$$

- Both are weighted averages of δ_X's, but $\hat{\rho}_{ATT}$ aggregates via *population weighting* while $\hat{\delta}_R$ aggregates via *conditional variance weighting* wrt D_i.

Heterogeneity

So, let's compare:

$$E[\hat{\rho}_{ATT}] = \frac{\sum_x \delta_X \Pr[D_i = 1|X_i = x] \Pr[X_i = x]}{\sum_x \Pr[D_i = 1|X_i = x] \Pr[X_i = x]}.$$

versus

$$\hat{\delta}_R \xrightarrow{a} \frac{\sum_x \delta_X [\Pr[D_i = 1|X_i = x](1 - \Pr[D_i = 1|X_i = x])] \Pr[X_i = x]}{\sum_x [\Pr[D_i = 1|X_i = x](1 - \Pr[D_i = 1|X_i = x])] \Pr[X_i = x]}$$

- Both are weighted averages of δ_X's, but $\hat{\rho}_{ATT}$ aggregates via *population weighting* while $\hat{\delta}_R$ aggregates via *conditional variance weighting* wrt D_i.
- Population weighting is unbiased, while variance weighting is biased: it privileges X_i values for which δ_X estimates are precise.

Heterogeneity

So, let's compare:

$$\mathrm{E}\left[\hat{\rho}_{ATT}\right] = \frac{\sum_x \delta_X \Pr[D_i = 1|X_i = x] \Pr[X_i = x]}{\sum_x \Pr[D_i = 1|X_i = x] \Pr[X_i = x]}.$$

versus

$$\hat{\delta}_R \xrightarrow{a} \frac{\sum_x \delta_X [\Pr[D_i = 1|X_i = x](1 - \Pr[D_i = 1|X_i = x])] \Pr[X_i = x]}{\sum_x [\Pr[D_i = 1|X_i = x](1 - \Pr[D_i = 1|X_i = x])] \Pr[X_i = x]}$$

- Both are weighted averages of δ_X's, but $\hat{\rho}_{ATT}$ aggregates via *population weighting* while $\hat{\delta}_R$ aggregates via *conditional variance weighting* wrt D_i.
- Population weighting is unbiased, while variance weighting is biased: it privileges X_i values for which δ_X estimates are precise.
- If ρ_i's were *constant* over X_i, the precision weighting would be good from an efficiency standpoint. But this is probably irrelevant.

Heterogeneity

So, let's compare:

$$\mathrm{E}\left[\hat{\rho}_{ATT}\right] = \frac{\sum_x \delta_X \Pr[D_i = 1|X_i = x] \Pr[X_i = x]}{\sum_x \Pr[D_i = 1|X_i = x] \Pr[X_i = x]}.$$

versus

$$\hat{\delta}_R \xrightarrow{a} \frac{\sum_x \delta_X [\Pr[D_i = 1|X_i = x](1 - \Pr[D_i = 1|X_i = x])] \Pr[X_i = x]}{\sum_x [\Pr[D_i = 1|X_i = x](1 - \Pr[D_i = 1|X_i = x])] \Pr[X_i = x]}$$

- Both are weighted averages of δ_X's, but $\hat{\rho}_{ATT}$ aggregates via *population weighting* while $\hat{\delta}_R$ aggregates via *conditional variance weighting* wrt D_i.
- Population weighting is unbiased, while variance weighting is biased: it privileges X_i values for which δ_X estimates are precise.
- If ρ_i's were *constant* over X_i, the precision weighting would be good from an efficiency standpoint. But this is probably irrelevant.
- If $D_i \perp\!\!\!\perp X_i$, then both $\hat{\rho}_{ATT}$ and δ_R reduce to the same thing: weighting by number of units with $X_i = x$.

Heterogeneity

- Logic carries through to continuous treatments (MHE, 77-80; Aronow & Samii, 2013).

Heterogeneity

- Logic carries through to continuous treatments (MHE, 77-80; Aronow & Samii, 2013).
- Aronow & Samii show that for arbitrary D_i and X_i,

$$\hat{\delta}_R \xrightarrow{p} \frac{\mathrm{E}[w_i \rho_i]}{\mathrm{E}[w_i]}, \text{ where } w_i = (D_i - \mathrm{E}[D_i | X_i])^2,$$

  in which case

$$\mathrm{E}[w_i | X_i] = \mathrm{Var}[D_i | X_i].$$

Heterogeneity

- Logic carries through to continuous treatments (MHE, 77-80; Aronow & Samii, 2013).
- Aronow & Samii show that for arbitrary D_i and X_i,

$$\hat{\delta}_R \xrightarrow{p} \frac{\mathrm{E}\left[w_i \rho_i\right]}{\mathrm{E}\left[w_i\right]}, \text{ where } w_i = (D_i - \mathrm{E}\left[D_i | X_i\right])^2,$$

in which case

$$\mathrm{E}\left[w_i | X_i\right] = \mathrm{Var}\left[D_i | X_i\right].$$

- Implication: even if you start with a representative sample, regression estimates may not aggregate effects in a representative manner.
- (Results hold for MLE and random coefficient models too.)
- Shows a distinction between internal validity and generalizability.

Heterogeneity

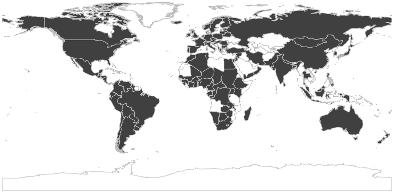

Figure 1: *On the left, the shading shows countries in the nominal sample for Jensen (2003)'s estimate of the effects of regime type on FDI. On the right, darker shading indicates that a country contributes more to the effective sample, based on the panel specification used in estimation.*

Non-linearity

- Recall that linearity in the context of regression means *linearity in the coefficients* (i.e., with respect to the finite set of polynomial terms or interactions you have included).
- There is nothing stopping you from including higher order polynomial terms or interactions, in the spirit of Weierstrauss approximation.

Non-linearity

- Recall that linearity in the context of regression means *linearity in the coefficients* (i.e., with respect to the finite set of polynomial terms or interactions you have included).
- There is nothing stopping you from including higher order polynomial terms or interactions, in the spirit of Weierstrauss approximation.
- Another type of non-linearity that we might worry about is associated with "limited" dependent variables.
- In this case, effects must be non-linear over the range of continuous variables.
- (If all regressors are binary, then there is no problem of course.)
- We will discuss this more at end of the semester.

Discussion

- Under random assignment, multiple regression is a tool for increasing efficiency without compromising consistency.

Discussion

- Under random assignment, multiple regression is a tool for increasing efficiency without compromising consistency.
- Under CIA, multiple regression estimates unbiased causal effects under linearity in X_i and constant effects.

Discussion

- Under random assignment, multiple regression is a tool for increasing efficiency without compromising consistency.
- Under CIA, multiple regression estimates unbiased causal effects under linearity in X_i and constant effects.
- When constant effects does not hold under CIA, then multiple regression estimates a conditional variance-weighted average that does not necessarily correspond to the target causal effect—i.e., it is biased and inconsistent for the ATE, ATT, ATC, etc.
- Whether or not this is a big deal depends on the extent of effect heterogeneity and how this relates to the distribution of X_i.

Discussion

- Under random assignment, multiple regression is a tool for increasing efficiency without compromising consistency.
- Under CIA, multiple regression estimates unbiased causal effects under linearity in X_i and constant effects.
- When constant effects does not hold under CIA, then multiple regression estimates a conditional variance-weighted average that does not necessarily correspond to the target causal effect—i.e., it is biased and inconsistent for the ATE, ATT, ATC, etc.
- Whether or not this is a big deal depends on the extent of effect heterogeneity and how this relates to the distribution of X_i.
- When confounding is not linear in X_i, then additional forms of bias come into play.
- Thus, under CIA, multiple regression may be biased and inconsistent for the target effect if there is effect heterogeneity or non-linear confounding.
- Such biases are what motivate matching and weighting estimators (coming soon).