Causal Inference Workshop

Week 8 - Limitations of identification strategies, pre-estimation, estimation, and post-estimation steps

Causal Inference Workshop

March 22, 2024

Anna Papp, ap3907@columbia.edu - SDEV 9280

Workshop outline

A. Causal inference fundamentals
   - Modeling assumptions matter too
   - Conceptual framework (potential outcomes framework)

B. Design stage: common identification strategies
   - IV + RDD [coding]
   - DiD, DiDiD, Event Studies, New TWFE Lit [coding]
   - Synthetic Control / Synthetic DiD [coding]

C. Analysis stage: strengthening inferences
   - Limitations of identification strategies, pre-estimation steps
   - Estimation [controls] and post-estimation steps [supporting assumptions]

D. Other topics in causal inference and sustainable development
   - Inference (randomization inference, bootstrapping)
   - Weather data regressions, other common/fun SDev topics [coding]
   - Remote sensing data, other common/fun SDev topics

Causal inference roadmap

What is causal inference?

- Process by which we use data to make claims about causal relationships

- *Potential outcomes* [framework]
  - Causal effect is difference between two potential outcomes

- *Identification* [application/implementation]
  - Identifying assumptions needed for a statistical estimate to have causal interpretation
  - Removing selection bias in regressions
  - E.g., RD, IV, ...

- *Estimation* [application/implementation] [today]
  - (Usually) use linear regression model

Outline

Limitations of identification strategies

- We can recover an **unbiased** estimator of a causal effect iff an **identifying/independence assumption** holds:
  - if IA holds ($(Y_i^0, Y_i^1) \perp\!\!\!\perp D_i$) $\rightarrow$ estimate ATT
  - if ~~IA~~, but CIA ($(Y_i^0, Y_i^1) \perp\!\!\!\perp D_i | X_i$) $\rightarrow$ can estimate ATT in each stratum (and then combine)
  - if ~~CIA~~, need relevant exogenous source of variation in D_i (e.g., $(Y_i^0, Y_i^1) \perp\!\!\!\perp Z_i$; $Z_i \perp\!\!\!\perp D_i$) $\rightarrow$ estimate a LATE

- What **identification strategy** buys us is overcoming selection bias (*unbiased* estimator)

  $\rightarrow$ estimated $\hat{\beta} = \ldots$ is *never* the true effect, it is one realization of an unbiased distribution! (it could be anywhere in the distribution)

- This is limited in at least three main ways

Limitations of identification strategies

- This is limited in at least three main ways
  1. **Accurate estimation still requires proper modeling**
  2. **Unbiasedness is no panacea**
  3. **The ATT is no panacea**

1. Accurate estimation still requires proper modeling

- In observational studies, we have at best a CIA
  → if CIA + know the correct function form $f()$ w.r.t. **confounders** X, the regression of Y on $\{D, f(X)\}$ gives an unbiased $\hat{ATT}$

- We don't ever know this $f()$ for sure (especially if distribution of X different across treatment groups, there are many Xs, etc. etc.)

- So we don't want to have to rely on $f()$
  - With imperfect overlap w.r.t. **confounders**, the model will have to extrapolate, so inferences will depend more on the specification of $f()$ (vs. direct support from data)
  - Avoid areas of imperfect overlap w.r.t **confounders** in the data!

→ **Choose a sensible $f()$ and restructure the data to have overlap wr.t. to confounders!**

2. Unbiasedness is no panacea

- Unbiased estimator $\hat{\theta}$: its distribution, $f_{\hat{\theta}}$, (over possible trials for the given sample size) is correctly centered around the true value of the estimand θ
  - Does **not** guarantee that any one realization is close to the center value!

→ **Adjust as much as possible for potential imbalance between groups!**

→ **Consider another property: efficiency** (reducing the width of $f_{\hat{\theta}}$)!

3. The ATT is no panacea

- We obtain an estimate on the ATT, but what knowledge are we generating from that?
- Reduced forms generally motivated by “RCT = gold standard”
  - In an RCT, *D* is an intervention; the average effect of that intervention might very well be the knowledge desired
- But in some other (non-intervention) contexts, estimating the magnitude of an effect but not identifying underlying mechanism may be less informative

So what can be done to address these limitations?

- Of course, need identification strategy

- But post-design, given a fixed dataset, there are still steps that can be done to generate more insightful inferences

- Specifically:
  1. **Pre-estimation**: restructuring data to improve overlap and balance w.r.t. confounders
  2. **At estimation stage**: include the right covariates, allow for TE heterogeneity
  3. **Post-estimation**: check assumptions, consider external validity

Outline

Pre-estimation: Restructuring data

- Causal inference requires that the treated units are comparable to the control units w.r.t. the confounders X

- Two forms of departures from comparability:
  - **Incomplete overlap:** the *support* of the distribution of X differs across groups; some observations have no counterfactuals

    → model forced to extrapolate when $\exists$ no overlap; inferences based on modeling assumptions

  - **Imbalance:** the *shape* of the distribution of X differs across groups

    → simple difference of group averages might not be a reliable estimate of the ATT

- If the two groups have sufficient overlap and balance, even if we misspecify $f(X)$, should get reasonable $\widehat{TE}$ (Gelman et al. 2020)
  - We want our full sample to be representative of the treatment group
    → Match groups to have balance/overlap w.r.t confounders (make sample look like RCT)

Pre-estimation: Restructuring data to balance observed confounders

Matching gives more overlap & balance, **not** identification, & is **not** a method of estimation!

Matching can be used in two ways:

- **In place of regression**: matching as estimation method (Angrist and Pischke 2008)
  - Regression with covariates OR covariate-matching are two ways to balance covariates
  - Comparisons for cells with the same covariate values, then compute difference in means

- **On top of regression:** matching as preprocessing method (Gelman et al. 2020)
  - Restructure sample *before* statistical analysis
  - Nonparametric preprocessing to reduce reliance on parametric assumptions (Ho et al. 2007)

Pre-estimation: Restructuring data, common distance metrics

Goal is to match each *treated* unit to its closest control unit w.r.t. confounders X

- Easy with a continuous X or even a binary X_1 and continuous X_2

- What about with many many confounders? Want to define a univariate **distance metric** between observations as a function of the Xs
  1. **Mahalanobis distance:** define distance metric that can include multiple dimensions of "closeness" between observations

$$d_{ij} = \sqrt{(X_i - X_j)' \Sigma_X^{-1} (X_i - X_j)}$$

     where Σ_X is the sample covariance matrix
  2. **Propensity score:** reduce dimensionality to 1 by computing a unit's *predicted probability of getting treated*, $\hat{p}_i = P(D_i = 1|X_i)$
     - Propensity score model: logistic regression of D_i on confounders, predict $\hat{p}_i$
     - Match each treated unit to nearest control units using $\hat{p}_i$

Outline

Estimation: Controls

Why do we *need* and why might we *like* certain controls?

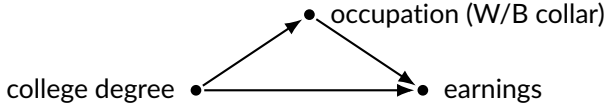
1. Required/forbidden controls for **bias**
2. Optional good/bad controls for **efficiency**

Estimation: Required/forbidden controls (bias)

For TE estimator to be unbiased, the identification strategy commands us to:

+ Do adjust for all confounders (in DAG language, to block *all* back-door paths, by adjusting for one variable along each path)

- Do not adjust for post-treatment variables that may be affected by the treatment ("intermediate outcomes") and that are also correlated with Y
  → **bad controls!** (see Chapter 3.2.3 of Angrist and Pischke (2008))
  - Example: effects of college degree on earnings



→ college degree changes the composition of the pool of white collar workers (even if college degree is randomly assigned!)

Estimation: Required/forbidden controls (bias)

For TE estimator to be unbiased, the identification strategy commands us to:

+ Do adjust for all confounders (in DAG language, to block *all* back-door paths, by adjusting for one variable along each path)

- Do not adjust for post-treatment variables that may be affected by the treatment ("intermediate outcomes") and that are also correlated with Y
  → **bad controls!** (see Chapter 3.2.3 of Angrist and Pischke (2008))

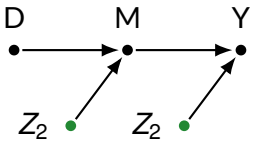
- Helpful to think about *timing*; variables measured before the variable of interest was determined are generally good controls (variables that cannot be outcomes)

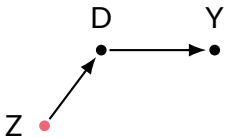
- *However*, bad controls not always useless! given some assumptions and data availability, may help us find bounds to causal effect of interest (see Chapter 3.2.3 of Angrist and Pischke (2008))

Estimation: Optional good/bad controls (efficiency)

Separately from the identification strategy, what other covariates should we adjust for?

- Among variables which do not interfere with identification:
  - \+ Adjusting for pre-treatment covariates that are strong determinants of Y may increase the efficiency of $\hat{\beta}$, by reducing the residual variance (unexplained variation in Y), reducing standard errors
  - \- Adjusting for determinants of treatment (D) will reduce the variation of D and therefore reduce the precision of $\hat{\beta}$ in finite samples





Estimation: Controls

Gelman et al. 2020 pg. 368 on the benefits of controls in terms of bias vs. precision:

> *"Under a clean randomization, adjusting for pre-treatment predictors in this way does not change what we are estimating. However, if the predictor has a strong association with the outcome it can help to bring each estimate closer (on average) to the truth* **[precision]**, *and if the randomization was less than pristine, the addition of predictors to the equation may help us adjust for systematically unbalanced characteristics across groups* **[bias]**. *Thus, this strategy has the potential to adjust for both random and systematic differences between the treatment and control groups (that is, to reduce both variance and bias), as long as these differences are characterized by differences in the pre-test."*

Outline

Post-estimation: Supporting assumptions and predictions

1. Diagnosis tests of modeling assumptions
2. Falsification tests of identifying assumptions
   - We can never directly *test* identifying assumptions (prove that they hold), but can do "falsification" or "placebo" analyses that either increase or decrease our confidence in them → *internal validity* of the study
   - *"Falsification tests set you up for failure: do them."* - Doug

Post-estimation: Model diagnostics

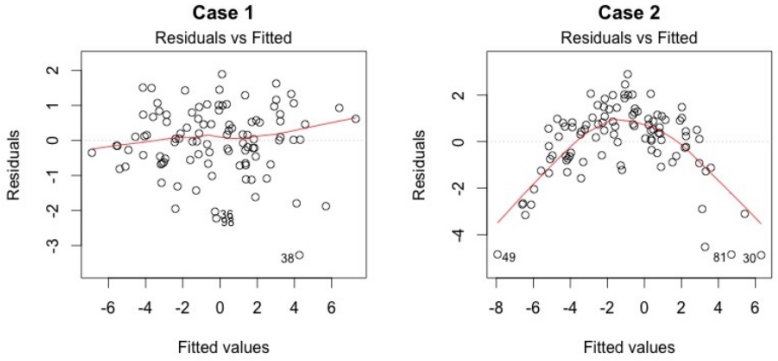
Recall the assumptions of the Classical Linear Regression Model (CLRM):

Notation:	System of n equations	Matrix
Model:	$y_i = x_i'\beta + e_i \quad (i = 1, ..., n)$	$y = X\beta + e$
Assumptions		
A1. linearity	model is linear in β	model is linear in β
A2. identification	$\rho x_k, x_l \approx 1$	$X_{N\times K}$ has rank K
A3. exogeneity	$\mathbb{E}[e_i\|X] = 0$	$\mathbb{E}[e_i\|X] = 0_{N\times 1}$
A4. spherical errors	$e_i\|X \overset{\text{iid}}{\sim} (0, \sigma^2)$	$\mathbb{V}[e\|X] = \sigma^2 I_N$
-independent errors	$cov[e_i, e_j\|X] = 0$	
-homoskedastic errors	$\mathbb{V}[e_i\|X] = \sigma^2 \quad \sigma_i^2$	
A5. normal errors	$e_i\|X \sim \mathcal{N}(0, \sigma^2)$	$e_i\|X \sim \mathcal{N}(0_{N\times 1}, \sigma^2 I_N)$

Post-estimation: Model diagnostics

"Residuals vs. Fitted" plot: *Is there an unmodeled non-linear pattern?*

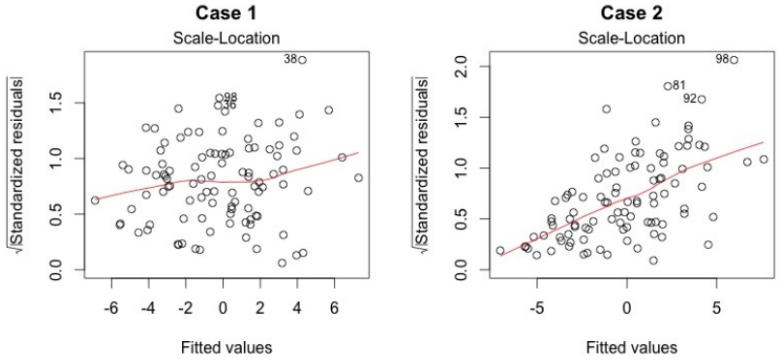
- Plot residuals vs. fitted values of $\hat{y}_i$
- Are the residuals spread rather equally around a horizontal line, without distinct patterns? If not, non-linear relationship is not explained by the model



Post-estimation: Model diagnostics

"Scale-Location" plot: *Are the residuals homoskedastic?*

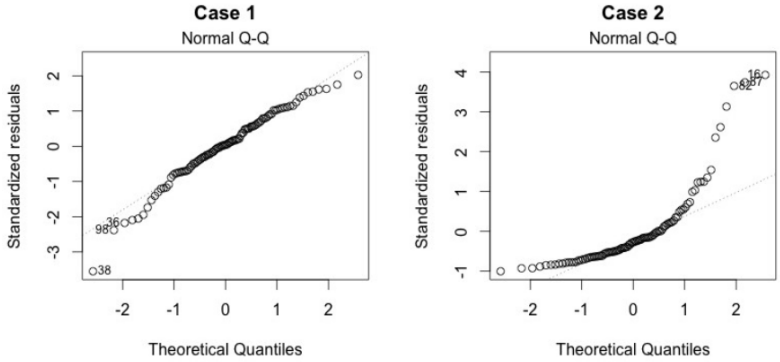
- Plot square root of the abs. value of standardized residuals $\sqrt{|r_i|}$ vs. fitted values of $\hat{y}_i$
- Is the vertical spread of points uniform along x? If not, suggests heteroskedasticity



Post-estimation: Model diagnostics

"Normal $\mathbb{Q}$-$\mathbb{Q}$" plot: *Are the residuals normally distributed?*

- Plot quantiles of the residuals against theoretical quantiles of normal distribution
- If residuals are approximately normally distributed, should see a roughly straight line



Post-estimation: Model diagnostics

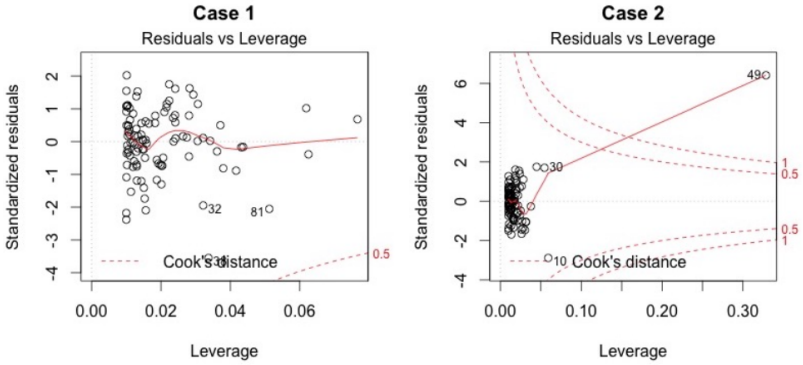
"Residuals vs. Leverage" plot: *Are there influential observations?*

- Different types of points:
  - **Outliers: observations with unusual outcomes**: they may or may not have a lot of influence on the regression line
  - **High-leverage points**: observations with unusual predictor values of x_i; leverage measures how sensitive a fitted $\hat{y}_i$ is to change in the true y_i
  - **Influential points**: observations whose removal from the data would cause a large change in the estimated regression line
- Cook's distance d_i measures the effect of omitting that observation; points with $d_i > 1$ generally considered influential

Post-estimation: Model diagnostics

"Residuals vs. Leverage" plot: *Are there influential observations?*

- Cook's distance d_i measures the effect of omitting that observation; points with $d_i > 1$ generally considered influential



Post-estimation: Show balance

- Causal inference rests upon the assumption that treatment and control groups are comparable to some extent
  - In RCT, identifying assumption is random assignment → explanatory variables should be the same across treatment and control groups (*in expectation*)
  - Even in observational settings, good idea to show a balance table to document, for each confounder X, the difference in distribution across treatment status

- Show how sample means differ between groups
  - Normalized difference $\Delta X = (\overline{X}_1 - \overline{X}_0)/\sqrt{S_0^2 + S_1^2}$, where S_0^2, S_1^2 are the sample variances of X in the control/treatment groups
  - According to Gelman et al. 2020, plot:
    - Standardized difference for continuous X: $\Delta X = (\overline{X}_1 - \overline{X}_0)/S_1$
    - Absolute difference in means for binary X: $\Delta X = (\overline{X}_1 - \overline{X}_0)$

Post-estimation: Falsification/placebo tests

- Show that the specification does not find an effect when one “should not” exist
- Look at outcome which should not be affected under the identifying assumption
  → if the analysis picks up effect, suggests that identifying assumption is violated
- Note that falsification tests $\neq$ robustness checks! Robustness checks estimate alternative specifications that test the same hypothesis

Post-estimation: IV falsification tests

Two main identifying assumptions can be tested:

- **Relevance** (Z is strongly related to sorting into treatment D)
  - Directly observable in first stage

- **Exclusion restriction** (Z isn't correlated with Y through some pathway other than D)
  - Can test in a falsification test
  - Ideal falsification test is to estimate the reduced form effect of Z on Y in some other situation where Z can't affect D (e.g., alternative population that can't be affected by treatment, but would be by potential confounders)

    → If there is an affect, that means Z affects Y through another channel than D

Post-estimation: RD falsification tests

Two main identifying assumptions can be tested:

- **Continuity or "local randomization"** (all other factors determining Y evolve smoothly w.r.t. Z)
  - Do other covariates jump at the cutoff c? Estimate same model, replace Y by covariates
  - Concern that units might be sorting on the running variables; if that was the case, we would expect some bunching of units at the cutoff (McCrary 2008 density test)

- **Relevance** (discontinuity in the dependence of D on Z)
  - For falsification tests, look at whether jumps occur at placebo cutoffs $\tilde{c}$

Post-estimation: DiD falsification tests

Two main identifying assumptions can be tested:

- **Same counterfactual trends across groups**
  - Compare trends in the pre-period
  - Use alternative outcome that shouldn't be affected by the treatment, but would be affected by potential confounders
  - Use alternative control group, check that you get the same $\hat{\beta}$
  - Move the event to points earlier in time and check that you get 0 effect

- **Same group composition over time**
  - Panel data satisfy it by definition
  - With repeated cross-sectional data, run covariate balance regressions

Post-estimation: SC falsification tests

- Move the event to points earlier in time, estimate on this earlier placebo date and check that there are zero effects

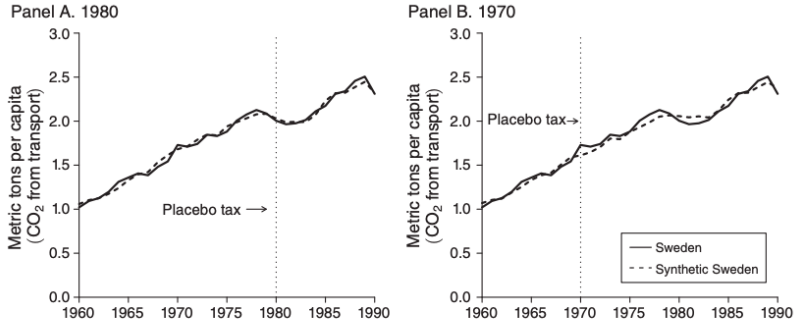


FIGURE 6. PLACEBO IN-TIME TESTS

Questions? Comments?

Thank you!

References

Heavily based on Claire Palandri's 2022 version of the Causal Inference Workshop.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press. ISBN: 978-1-4008-2982-8. http://muse.jhu.edu/book/64829.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories.* Analytical Methods for Social Research. Cambridge University Press.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15 (3): 199–236. https://doi.org/10.1093/pan/mpl013.

McCrary, Justin. 2008. "Manipulation of the running variable in the regression discontinuity design: A density test." The regression discontinuity design: Theory and applications, *Journal of Econometrics* 142 (2): 698–714. ISSN: 0304-4076. https://doi.org/https://doi.org/10.1016/j.jeconom.2007.05.005. https://www.sciencedirect.com/science/article/pii/S0304407607001133.