Causal Inference Workshop

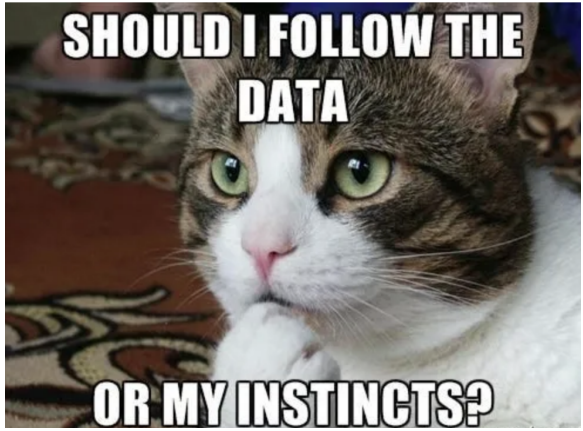
Week 1 - Modeling Fundamentals

Causal Inference Workshop

January 19, 2024

Anna Papp, ap3907@columbia.edu - SDEV 9280

Welcome to the Causal Inference Workshop!



Welcome to the Causal Inference Workshop!

- Very heavily based on Claire's awesome course
- Updated with some coding exercises for relevant topics
- Also added additional SDev-y topics (weather data, remote sensing data)
- Will also upload materials from Claire's version of the workshop which we won't cover in detail (e.g., multilevel modeling)

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?

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?

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?

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?

Workshop Logistics

- 9am - 10am Fridays (before colloquium)
  - I will be here 8am going forward, if you want to work through code / discuss anything
- If you're taking the course for credit (pass/fail) and "research tools" requirement, attendance is expected every week
  - If you're not taking it for credit, but want to get workshop-related emails, let me know!
- One assignment to turn it at the end (run and modify one of the coding exercises)

Outline

Causal inference, identification, and modeling assumptions

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]
  - (Usually) use linear regression model

Outline

Today: Modeling assumptions

- In econometrics, focus of causal inference is usually on *identification*

- But we are *never* in a perfect identification setup in observational studies
  - Model specification has influence
  - Modeling assumptions must hold (*especially* because we make causal claims based on statistical significant (e.g., p-values))
  - May want more than just unbiasedness (e.g., precision, external validity)

→ Need to think about modeling assumptions and estimator properties

The Classical Linear Regression Model

- The Classical Linear Regression Model (CLRM):

$$y|X \sim \mathcal{F}(X\beta, \sigma^2 I)$$

- Represent specific distribution induced by the data generating process (DGP)
- Regression models focus on *conditional distribution* $y|X$
- We can therefore write them as *conditional mean* $+/\times$ *error*

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + e, \qquad e \overset{\text{iid}}{\sim} \mathcal{F}(0, \sigma^2 I)$$

$$y = \mathbb{E}[y|X] + e, \qquad \mathbb{E}[y|X] = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k, \qquad e \overset{\text{iid}}{\sim} \mathcal{F}(0, \sigma^2 I)$$

Assumptions of the CLRM

- A set of assumptions that describe a DGP (CLRM)
- The assumptions (aka Gauss-Markov assumptions), by decreasing order of importance:

Notation:	System of n equations	Matrix
Model:	$y_i = x_i'\beta + e_i \quad (i = 1, ..., n)$	$y = X\beta + e$

Assumptions

Assumptions of the CLRM

- A set of assumptions that describe a DGP (CLRM)
- The assumptions (aka Gauss-Markov assumptions), by decreasing order of importance:

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

Assumptions of the CLRM

- A set of assumptions that describe a DGP (CLRM)
- The assumptions (aka Gauss-Markov assumptions), by decreasing order of importance:

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

Assumptions of the CLRM

- A set of assumptions that describe a DGP (CLRM)
- The assumptions (aka Gauss-Markov assumptions), by decreasing order of importance:

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	$\mathbb{E}[e_i \mid X] = 0$	$\mathbb{E}[e_i \mid X] = 0_{N\times 1}$

Assumptions of the CLRM

- A set of assumptions that describe a DGP (CLRM)
- The assumptions (aka Gauss-Markov assumptions), by decreasing order of importance:

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	$\mathbb{V}[e_i\|X] = \sigma^2 \quad \sigma_i^2$	

Assumptions of the CLRM

- A set of assumptions that describe a DGP (CLRM)
- The assumptions (aka Gauss-Markov assumptions), by decreasing order of importance:

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	$e_i\|X \sim \mathcal{N}(0, \sigma^2)$	$e_i\|X \sim \mathcal{N}(0_{N\times 1}, \sigma^2 I_N)$

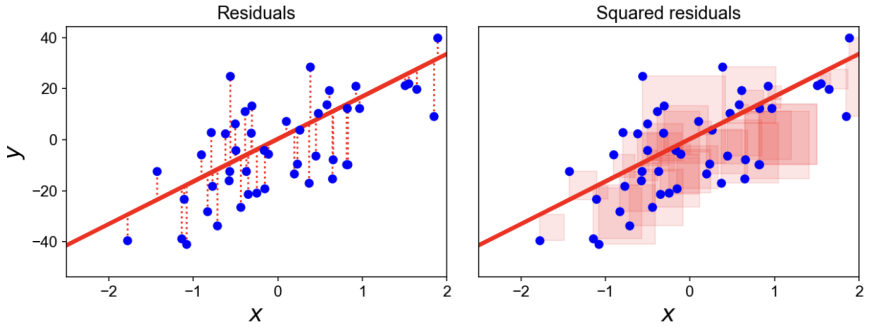
Properties of $\hat{\beta}_{OLS}$

→ if these assumptions are met, we can make the powerful claim that $\hat{\beta}_{OLS}$ is the best linear model we can use

Properties of $\hat{\beta}_{OLS}$

→ if these assumptions are met, we can make the powerful claim that $\hat{\beta}_{OLS}$ is the best linear model we can use

$$\hat{\beta}_{\text{OLS}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} r_i^2 = \sum_{j=1}^{n} (y_i - X_i'\beta)^2 = \ldots = (X'X)^{-1}X'y = \beta_0 + (X'X)^{-1}X'e$$



Properties of $\hat{\beta}_{OLS}$

→ if these assumptions are met, we can make the powerful claim that $\hat{\beta}_{OLS}$ is the best linear model we can use

$$\hat{\beta}_{\text{OLS}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} r_i^2 = \sum_{j=1}^{n} (y_i - X_i'\beta)^2 = \ldots = (X'X)^{-1}X'y = \beta_0 + (X'X)^{-1}X'e$$

Based on the assumptions, $\hat{\beta}_{OLS}$ has the following:

- finite sample properties:
  - (A1-A3) → $\hat{\beta}_{OLS}$ **unbiased**
  - (A4) → $\hat{\beta}_{OLS}$ **efficient** (Best **Linear** Unbiased Estimator (BLUE))
  - (A5) → $\hat{\beta}_{OLS}$ **efficient** ($\hat{\beta}_{MLE}$; Best Unbiased Estimator; normal)

- asymptotic properties:
  - $\hat{\beta}_{OLS}$ is consistent and is asymptomatically unbiased, normally distributed, and efficient

(Assumption 5) Normal errors

- Normal error assumption required for making inferences (computing confidence intervals or p-values)
- Without (A5), *t* and *F* tests are invalid
  - One-sample t-test for β ($\beta = 0$) assumes sampling distribution of $\hat{\beta}$ is normal (which means errors have to be normal)
- When (A5) is violated, appeal to asymptotics:
  - When *n* is large enough, Laws of Large Numbers (LLNs) and Central Limit Theorem (CLT) say that asymptotic sampling distribution of $\hat{\beta}$ is normal
  - If *n* is large, *t* and *F* tests are robust to departures from normality
  - In the case of highly non-normal error, may want to consider alternative (e.g., bootstrap)

Outline

(A4) Non-spherical errors

Assuming (A1)-(A3), the asymptotic distribution of $\hat{\beta}_{OLS}$ is:

$$\hat{\beta}_{OLS} \overset{a}{\sim} \mathcal{N}\left(\beta_0, (X'X)^{-1}X'\Sigma X(X'X)^{-1'}\right)$$

→ need a consistent estimate of the asymptotic vcov matrix in order to do sampling-based statistical inference → need Σ (the vcov matrix of the error term)

(A4) Non-spherical errors

Assuming (A1)-(A3), the asymptotic distribution of $\hat{\beta}_{OLS}$ is:

$$\hat{\beta}_{OLS} \overset{a}{\sim} \mathcal{N}\left(\beta_0, (X'X)^{-1}X'\Sigma X(X'X)^{-1'}\right)$$

→ need a consistent estimate of the asymptotic vcov matrix in order to do sampling-based statistical inference → need Σ (the vcov matrix of the error term)

- Spherical *e*: $\Sigma = \sigma^2 I$, so we can simply consistently estimate the population variance σ^2 by the unbiased sample variance
- Heteroskedastic *e*: compute White SEs
- Autocorrelated *e*: compute Newey-West/Conley SEs if correlated in time/space...
- Clustered *e*: compute block-diagonal matrix using residuals

Notes on sandwich estimators

- If errors are autocorrelated in any way, it means your model is not capturing some feature of the DGP
  - Adjust for it after fitting the model (e.g., cluster SEs if autocorrelated by group)
  - Or incorporate into the structure of the model (e.g., multilevel structure)
- Sandwich estimators rely on asymptotics
  - Don't cluster SEs if you have too few clusters!
- These SEs only make sense for the linear regression model

Limited outcome models

A linear regression isn't appropriate when there is a *limited* y

- Binary: $y \in 0, 1$
- Count: $y \in 0, 1, 2, 3, \ldots$
- Censored

Limited outcome models

A linear regression isn't appropriate when there is a *limited y*

- Binary: $y \in 0, 1 \rightarrow$ probit, logit, ...
- Count: $y \in 0, 1, 2, 3, ... \rightarrow$ Poisson, negative binomial,...
- Censored $\rightarrow$ censored regression models

Use **generalized** linear models (GLMs) - flexible generalization of OLS:

$$g(\mathbb{E}[y|X]) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k, \quad e \overset{\text{iid}}{\sim} \mathcal{F}(0, ...)$$

- Invertible link function $g()$, which relates:
- $\mathbb{E}[y|X]$ to linear predictor vector $X\beta$
- Assume some data distribution $\mathcal{F}()$

Outline

The Bottom-Line

In practice, causal inference in observational studies means, *both*

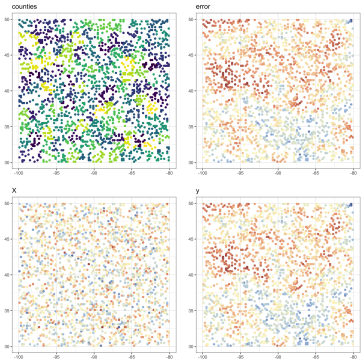
- A good design (identifying assumptions) → unbiasedness
- A model (modeling assumptions)

You shouldn't overlook the modeling assumptions!

Outline

Mini coding exercise

- https://bit.ly/causalIW
- Create fake data with spatial autocorrelation
- Test clustered and Conley SEs
- You can play around with:
  - Parameters of DGP
  - Strength and type of spatial correlation
  - Conley SEs distance parameter
  - Clustered SEs (number of clusters, etc.)
  - Omitted variable bias
  - R vs. Stata functions



Questions? Comments?

Thank you!

References

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

Appendix

More on limited outcome models from Claire's slides:

Limited outcome models

A *limited* y (categorical, or constrained to fall in a certain range) often arises. Linear regression isn't appropriate as it doesn't take into account the constraint on possible values of y.

Limited y	Appropriate regression models
binary: $y \in \{0, 1\}$	Probit, logit...
count: $y \in \{0, 1, 2, 3, ...\}$	Poisson, negative binomial...
censored	censored regression models

Use generalized linear models (GLMs):

$$g\big(\mathbb{E}[y|X]\big) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k, \qquad e \overset{\text{iid}}{\sim} F(0, ...)$$

An invertible link function $g()$ relates $\mathbb{E}[y|X]$ to the linear predictor vector $X\beta$, and we assume a data distribution $F(y|g^{-1}(X\beta))$.

- Ex: logistic reg model is a GLM with binomial data and logit link $\ln\left(\frac{\cdot}{1-\cdot}\right)$
- Ex: the Poisson reg model is a GLM with Poisson data and log link $\ln()$

Ex: Binary outcome models

$$y|X \sim Ber(\pi) = \begin{cases} 1 & \text{with probability } \pi \\ 0 & \text{with probability } 1-\pi \end{cases}$$

A regression model expresses the conditional probability $\pi \equiv \mathrm{P}[y{=}1|X]$ as a function of X and β:

$$y_i|X_i \sim Ber(\pi_i), \quad \pi_i = g^{-1}(X_i'\beta)$$

- Linear probability model $\pi_i = X_i'\beta + e_i$
  will not constrain the predicted values to be in [0,1]; in almost all circumstances, yields biased and *inconsistent* estimates (i.e., gives the wrong answer, with almost certainty, even with an infinitely large sample) [3]
- **Logistic regression model**

$$\pi_i = \text{logit}^{-1}(X_i'\beta) \equiv \frac{e^{X_i'\beta}}{1+e^{X_i'\beta}} \iff \text{logit}(\pi_i) = X_i'\beta$$

Ex: Count data models

$y_i \in \{0, 1, 2, ...\}$: number of occurrences of an event. *Ex: number of children in a household, number of doctor visits per year.*

- **Poisson regression model**

  Assume $y_i|X_i \sim Pois(\lambda_i)$. A single parameter $\lambda > 0$ (mean rate of occurrence of the event) to be explained by the predictors.

  $$\mathbb{E}[y_i|X_i] = \mathbb{V}[y_i|X_i] = \exp(X_i'\beta)$$

  Pb: implies equi-dispersion: $\mathrm{Var}[y_i|x_i] = \mathbb{E}[y_i|x_i]$, whereas we generally have overdispersion (ex: a few traders will do many trades, many traders will do a few).

- **Negative binomial model**

  The additional parameter $r > 0$ in the $NB(p, r)$ distribution captures overdispersion.