Perelman SCHOOL OF MEDICINE
UNIVERSITY of PENNSYLVANIA

BIOSTATISTICS FOR EPIDEMIOLOGIC METHODS II (EPID5270) PART II

Introduction to Causal Inference

Luke Stewart

2/13/2026

Roadmap

01 CAUSAL THINKING & POTENTIAL OUTCOMES

02 COMMON ASSUMPTIONS FOR CAUSAL INFERENCE

03 REVIEW OF REGRESSION-BASED METHODS

04 ALTERNATIVES TO REGRESSION-BASED METHODS

Disclaimer

This presentation will be somewhat dogmatic in that it presents a very specific view of what it means to ask and answer causal questions.

Although not the definitive perspective, the setup discussed here predominates modern causal inference research across statistics, medicine, epidemiology, public health, and the social sciences.

Other methods of research can provide information about causes and effects, but I do not touch on that here.

Disclaimer (part 2)

What I'll talk about today merely brushes on decades of research in the study of causal effects.

The field is much wider and much deeper than the content here. I encourage you to explore it more in electives and your own research!

Causal thinking

Which questions are causal (and which are not)?

- Example: Smoking and blood pressure

- Tempting question to ask when presented with a dataset of blood pressure measurements and many covariates: “What causes high blood pressure?”
  - Sounds causal, but isn’t (in the technical sense)

- Actual causal question: “What is the effect of smoking on blood pressure?”
  - As we’ll soon see, this question is also fraught with issues

- Causal inference asks, “What is the effect of exposure X on outcome Y?”
  - Causal inference does not (directly) ask, “What are all the causes Y?” or “What all explains the occurrence of Y?”

Roadmap to answer causal questions

1. Identify the **question of interest**

2. Write the question in terms of a specific **causal estimand**

3. Connect the causal estimand to observable data via **identifying assumptions**
   - This gives you a **statistical model**, the **parameters** of which you want to estimate

4. **Estimate** and conduct **inference** on the parameters of your statistical model (and thus, your causal estimand)

Potential outcomes

- For this lecture, we focus on a binary treatment A and an outcome Y that can be binary, categorical, count, or continuous
  - The terms “exposure” and “treatment” refer to the same concept, but are best suited to different studies
  - Motivating example: effect of getting the seasonal flu vaccine on flu incidence in the months following

- Let $Y_i(\mathrm{a})$ be the potential outcome for patient i when they experience treatment a
  - $Y_i(1)$ is an indicator of flu incidence if the patient received the flu vaccine
  - $Y_i(0)$ is an indicator of flu incidence if that same patient did not receive the vaccine

- For untreated patients ($\mathrm{A} = 0$), we observe $Y_i(0)$ in our data, but we also define and work with $Y_i(1)$, which we do not observe (and vice versa)

- Fundamental problem of causal inference (Holland 1986): Ideally, we would compare $Y_i(1)$ and $Y_i(0)$ for everyone, but we cannot observe both simultaneously

Causal estimands

- Instead of individual comparisons, we try to identify population aggregations of comparisons between potential outcomes:
  - Average treatment effect: $E[Y_i(1) - Y_i(0)]$
    - Equal to a risk difference in our setting with binary outcome: $P(Y_i(1) = 1) - P(Y_i(0) = 1)$
  - Average treatment effect on the treated: $E[Y_i(1) - Y_i(0) | \mathrm{A_i} = 1]$
  - Causal risk ratio: $P(Y_i(1) = 1)/P(Y_i(0) = 1)$
  - Causal odds ratio: $\frac{P(\mathrm{Y_i}(1)=1)/(1-\mathrm{P}(\mathrm{Y_i}(1)=1))}{\mathrm{P}(\mathrm{Y_i}(0)=1)/(1-\mathrm{P}(\mathrm{Y_i}(0)=1))}$

- Under specific assumptions (which vary by setting and by estimand), we can relate these quantities which involve unobservable values to our observable data

Causal assumptions

Three main assumptions

- The rest of lecture mostly focuses on the average treatment effect (ATE): $E[Y_i(1) - Y_i(0)]$

- Assume we have independent, identically distributed data on outcomes Y_i (flu incidence), exposure A_i (flu vaccine), and covariates X_i (e.g. demographics), for patients $\mathrm{i} = 1, \ldots, \mathrm{n}$

- In this setting, three assumptions together allow us to write the ATE in terms of observable data:
  1. Consistency: $Y_i(a) = Y_i$ for those with exposure a
  2. Positivity: $0 < P(A_i = 1|X_i) < 1$
  3. Exchangeability: $\mathrm{Y_i(a)} \perp \mathrm{A_i}|\mathrm{X_i}$

Consistency: $Y_i(a) = Y_i$

- Consistency directly relates a potential outcome to an observable outcome

- It has two main implications:
  1. One person's treatment does not affect another person's outcome
     - Does this hold for the flu vaccine example?
  2. There is only one "version" of treatment; treatment is well defined
     - Potentially ill-defined exposures include BMI and smoking status

- In some fields, consistency (or at least, some version of it) is known as the SUTVA (stable unit treatment value assumption)

Positivity (aka overlap): $0 < P(A_i = 1|X_i) < 1$

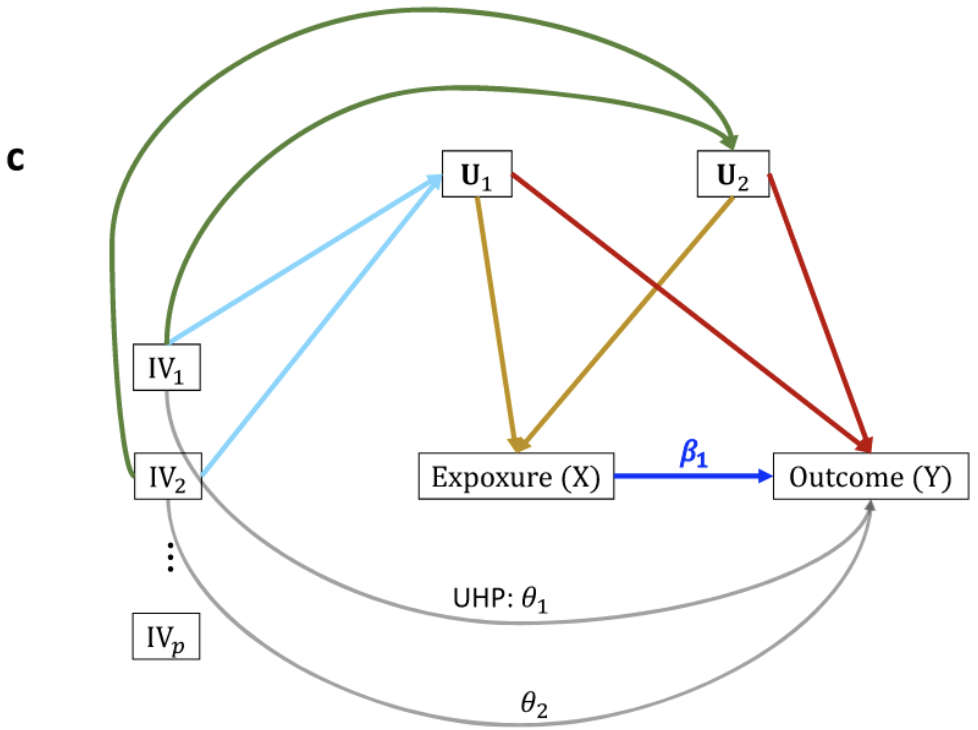
- Necessary to use probability properties in identification proofs
- Sometimes thought of as a more technical assumption, but has intuitive merit
- Intuition: for a patient in the treated group, if there is no one similar in the control group with which to compare, we cannot infer anything about the treatment effect for them without extrapolation
- Unlike the other assumptions, this is somewhat verifiable from the data

Exchangeability: $Y_i(a) \perp A_i|X_i$

- Within strata of X_i, the **potential outcomes** are independent of the treatment actually received
- Also referred to as “ignorability”; often synonymous with “no unobserved confounding”
- Can be a very strong assumption in some settings, but is very reasonable in others (i.e. randomized trials where randomization is stratified on X_i)
  - Would it hold in the vaccine example when X_i only includes demographics like race, sex, and age?

Quick aside: DAGs (Directed Acyclic Graphs)

- You've previously covered DAGs in the context of mediation and effect modification
- Algorithms have been developed to use a DAG to generate a set of confounders to include in your causal model (and eliminate bad controls)
- As clinicians, your ability to use domain knowledge to inform the DAG for a study is very valuable!



(Cheng et al. 2022)

Review: Regression-based methods

Using regression-based methods to estimate causal effects

- When we include covariates in a regression model of our outcome on treatment, what are we trying to do?

- Assume the following model for the outcome:

$$Y_i = \alpha + \tau \cdot A_i + \beta X_i + \epsilon_i$$

- Regressing Y_i on A_i and X_i conditions on X_i, so using our causal assumptions, τ is exactly equal to the average treatment effect

- Proof on next slide

Linear regression can estimate the ATE

$$
\begin{aligned}
E[Y_i(1) - Y_i(0)] &= E[E[Y_i(1) - Y_i(0)|X_i]] && \text{by law of iterated expectations} \\
&= E[E[Y_i(1)|X_i] - E[Y_i(0)|X_i]] && \text{by linearity of expectation} \\
&= E[E[Y_i(1)|X_i, A_i = 1] - E[Y_i(0)|X_i, A_i = 0]] && \text{by exchangeability and positivity} \\
&= E[E[Y_i|X_i, A_i = 1] - E[Y_i|X_i, A_i = 0]] && \text{by consistency} \\
&= E[\alpha + \tau + \beta X_i - (\alpha + \beta X_i)] && \text{plug in our model for observed } Y_i \\
&= E[\tau] = \tau
\end{aligned}
$$

When does covariate adjustment get the right answer?

- We used the law of iterated expectations and linearity of expectation (which always hold), in addition to our three causal assumptions (consistency, positivity, and exchangeability) to show that we estimate the ATE with linear regression.

- But we also assumed that our model for the outcome was correct!

- In general, if our model is wrong, the coefficient τ is not equal to the ATE.
  - For example, if the correct model included $log(X_i)$ instead of X_i, we would not be estimating the ATE

- **Bottom line: it's very important to get the model right by including all the covariates and choosing the right functional form for them (including covariate-by-covariate interactions)**

Flexible functional forms to the rescue?

- One way to have a better chance at getting the form of the model right is to use flexible models, like splines or machine learning models
- Upside: Assuming you include all the relevant confounders, these flexible methods are far more capable at finding the right form
- Downside: Getting confidence intervals (and performing hypothesis testing) can be far more difficult
  - For the most part, the theory that gives you confidence intervals for these methods is based on large-sample approximations
  - In small samples, the ATE estimates from these methods could be severely biased
- These methods are harder to use, and some audiences may not prefer their use

Alternatives* to modelling the outcome

Inverse probability of treatment weighting (IPTW)

- Under the same set of assumptions as before, the average treatment effect can also be identified as:

$$E[Y_i(1) - Y_i(0)] = E\left[\frac{\mathrm{A_i Y_i}}{P(\mathrm{A_i} = 1|\mathrm{X_i})} - \frac{(1 - \mathrm{A_i})\mathrm{Y_i}}{1 - P(\mathrm{A_i} = 1|\mathrm{X_i})}\right]$$

- $P(\mathrm{A_i} = 1|X_i)$ is known as the propensity score
- The quantity on the right-hand side can be estimated by the inverse probability of treatment weighting (IPTW) estimator

$$\hat{\theta}_{\mathrm{IPTW}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left[\frac{\mathrm{A_i Y_i}}{\widehat{\mathrm{P}}(\mathrm{A_i} = 1|\mathrm{X_i})} - \frac{(1 - \mathrm{A_i})\mathrm{Y_i}}{1 - \widehat{\mathrm{P}}(\mathrm{A_i} = 1|\mathrm{X_i})}\right]$$

- $\widehat{\mathrm{P}}(\mathrm{A_i} = 1|\mathrm{X_i})$ is an estimated propensity score from a model (e.g. logistic regression)
- Problem: the propensity score model must be correctly specified! Similar issues and solutions as to the outcome model approach
- Modern “doubly robust” approaches allow either the outcome model or propensity score model to be correctly specified to have consistent estimates

Matching

- Goal: Restrict the analysis to populations (and therefore samples) where treatment assignment is exchangeable

- Matching methods group treated individuals together with untreated individuals that are otherwise similar
  - More formally, matching algorithms minimize a “distance” between treated and untreated individuals that is calculated based on their covariates

- Four steps to matching (Stuart 2010):
  1. Defining “distance”
  2. Implementing a matching algorithm with the chosen distance
  3. Assessing the quality of the match (and adjusting by redoing steps 1 and 2)
  4. Analysis of the outcome

- Only step 4 involves looking at the outcome! We’re not p-hacking by repeating steps 1, 2, and 3.

Matching

- Typically, matching is used when we have more untreated units than treated units
- The “extra” untreated units that aren’t similar to any treated units will generally not be used in the match
- The result is a “pseudopopulation” that looks more like the exposed population than the unexposed population
- Sounds like this may solve our modelling problem, right? Not quite…

Connections between matching and IPTW

- Exact matching on more than a few covariates tends to be impossible
  - With continuous covariates, even more so
- One popular measure of “distance” in matching is the propensity score
- Both matching and IPTW create pseudopopulations
  - IPTW weights untreated and treated units in such a way to make the two groups comparable
  - Matching restricts the pseudopopulation to untreated units that are similar to treated units
- Matching estimators can be more stable (i.e. lower variance) than weighting estimators in some situations
- **Both methods can be used in conjunction with a regression model of the outcome, and both reduce the dependence of your result on correctly specifying the outcome model**

What if I have unobserved confounding?

Whirlwind tour of causal inference with unobserved confounding:

- Sensitivity analyses
  - Analysis of how sensitive the results of the previous methods are to hypothetical unobserved confounding
- Instrumental variables
  - Uses random variation in some other variable (the instrument) to inform how an exposure affects an outcome
  - Useful in trials where non-compliance is a problem (related to principal stratification)
- Discontinuity designs
  - Used when treatment is assigned by a cutoff in a running variable (like a diagnostic decision rule)
  - Compares individuals immediately on both sides of the cutoff to estimate a local treatment effect
- Longitudinal, quasi-experimental designs
  - Frequently used to analyze the effect of a policy implementation when units (most frequently geographies like counties or states) are observed over time
  - E.g. difference-in-differences, synthetic controls, interrupted time series
- And more!

To sum it all up…

- Ask causal questions, and carefully translate them into **causal estimands**
  - In particular, identify the **potential outcomes** you want to compare
- Make explicit, formal **identifying assumptions** to connect your question to observable data
  - Use your domain knowledge to construct a DAG to inform your study design
- Then (and only then), choose a statistical model/estimation method that makes sense for your data
- Critically evaluate your assumptions (before analyzing your outcome, if possible)

References and more reading

IPTW overview: Austin, Peter C., and Elizabeth A. Stuart. 2015. “Moving towards Best Practice When Using Inverse Probability of Treatment Weighting (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies.” *Statistics in Medicine* 34 (28): 3661–79. https://doi.org/10.1002/sim.6607.

Matching overview: Stuart, Elizabeth A. 2010. “Matching Methods for Causal Inference: A Review and a Look Forward.” *Statistical Science* 25 (1). https://doi.org/10.1214/09-STS313.

Instrumental variables: Baiocchi, Michael, Jing Cheng, and Dylan S. Small. 2014. “Instrumental Variable Methods for Causal Inference.” *Statistics in Medicine* 33 (13): 2297–340. https://doi.org/10.1002/sim.6128.

Difference-in-differences: Rothbard, Sarah, James C. Etheridge, and Eleanor J. Murray. 2024. “A Tutorial on Applying the Difference-in-Differences Method to Health Data.” *Current Epidemiology Reports* 11 (2): 85–95. https://doi.org/10.1007/s40471-023-00327-x.

Source for complicated DAG: Cheng, Qing, Xiao Zhang, Lin S. Chen, and Jin Liu. 2022. “Mendelian Randomization Accounting for Complex Correlated Horizontal Pleiotropy While Elucidating Shared Genetic Etiology.” *Nature Communications* 13 (1): 6490. https://doi.org/10.1038/s41467-022-34164-1.

- Side note: I don’t endorse this particular paper, but I’m happy to chat with anyone interested in genetic epidemiology about Mendelian randomization

Fundamental problem of causal inference source: Holland, Paul W. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81 (396): 945–60. https://doi.org/10.1080/01621459.1986.10478354.