Introduction to Causal Inference

Caleb Leedy

September 5, 2023

Frameworks for Causal Inference

- Potential Outcomes
- Directed Acyclic Graphs (DAGs)

Background and Notation

Potential Outcomes

- Outcome variable Y
- Treatment assignment $A \in \{0, 1\}$
- Covariates X
- The *potential outcomes* of Y for individual i are

$$Y_i(0) \text{ and } Y_i(1).$$

Potential Outcomes

- Often the goal is to understand: $Y(1) - Y(0)$.
- For each individual i, we only observe *either* $Y_i(0)$ *or* $Y_i(1)$.

Assumptions

- Stable unit treatment value assumption (SUTVA):

$$Y_i = Y_i(1)A_i + Y_i(0)(1 - A_i).$$

- No Unmeasured Confounders (NUC):

$$Y(1), Y(0) \perp A \mid X$$

- Positivity: For X with $p(X) > 0$,

$$0 < \Pr(A = 1 \mid X) < 1.$$

Goals:

- Average treatment effect (ATE):

$$E[Y(1) - Y(0)]$$

- Average treatment effect of the treated (ATT):

$$E[Y(1) - Y(0) \mid A = 1]$$

- Conditional average treatment effect (CATE):

$$E[Y(1) - Y(0) \mid X]$$

- Individual treatment effect (ITE):

$$Y_i(1) - Y_i(0)$$

Average Treatment Effect

- Consider the ATE denoted by

$$\tau = E[Y(1) - Y(0)].$$

Outcome Model

- From our assumptions, $E[Y(a)] = E[E[Y \mid X, A = a]]$.
- We can construct a model, $m_a(X) = E[Y \mid X, A = a]$ and estimate τ using
$$\hat{\tau}_{\text{OR}} = \hat{m}_1(X) - \hat{m}_0(X).$$
- If the model for $\hat{m}_a(X)$ is correctly specified, $\hat{\tau}_{\text{OR}}$ is consistent.

Propensity Score

- Instead of modeling the outcome $E[Y \mid X]$ we can model the response probability $\pi(X) = \Pr(A = 1 \mid X)$.

Inverse Propensity Score Weighting

- If $\pi(X)$ is known then we can estimate τ with $\hat{\tau}_{\text{IPW}} = \hat{\mu}_1 - \hat{\mu}_0$.
- We estimate $E[Y(1)]$ with

$$\hat{\mu}_1 = n^{-1} \sum_{i=1}^{n} \frac{A_i Y_i}{\pi(X_i)}.$$

- We can estimate $E[Y(0)]$ with

$$\hat{\mu}_0 = n^{-1} \sum_{i=1}^{n} \frac{(1 - A_i) Y_i}{1 - \pi(X_i)}.$$

- This is similar to a Horvitz-Thompson estimator.

Result

- The IPW estimator is consistent.

$$
\begin{aligned}
E\left[\frac{AY}{\pi(X)}\right] &= E\left[\frac{AY(1)}{\pi(X)}\right] \\
&= E\left[E\left[\frac{AY(1)}{\pi(X)} \mid Y(1), X\right]\right] \\
&= E\left[Y(1)E\left[\frac{A}{\pi(X)} \mid X\right]\right] \\
&= E[Y(1)].
\end{aligned}
$$

Double Robust Estimation

- We can combine our outcome model and response model together to get a doubly robust estimator: $\hat{\tau}_{\text{DR}} = \hat{\mu}_{1,\text{DR}} - \hat{\mu}_{0,\text{DR}}$ where

$$\hat{\mu}_{1,\text{DR}} = n^{-1} \sum_{i=1}^{n} \left(m_1(x_i) + \frac{A_i}{\hat{\pi}(X_i)} (Y_i - m_1(x_i)) \right)$$

and

$$\hat{\mu}_{0,\text{DR}} = n^{-1} \sum_{i=1}^{n} \left(m_0(x_i) + \frac{1 - A_i}{1 - \hat{\pi}(X_i)} (Y_i - m_0(x_i)) \right).$$

Result

- $\hat{\tau}_{\mathrm{DR}}$ is consistent if either the outcome or the response model is true.
- If the outcome model is correctly specified ($m_1(x) = E[Y(1) \mid X]$), then

$$
\begin{aligned}
& E[\hat{\mu}_{1,\mathrm{DR}}] \\
& = n^{-1} \sum_{i=1}^{n} \left(E[m_1(x_i)] + E\left[E\left[\left(\frac{A_i}{\hat{\pi}(X_i)} \right) (Y_i(1) - m_1(x_i)) \mid X \right] \right] \right) \\
& = n^{-1} \sum_{i=1}^{n} \left(E[m_1(x_i)] + E\left[E\left[\left(\frac{A_i}{\hat{\pi}(X_i)} \right) \mid X \right] E\left[(Y_i(1) - m_1(x_i)) \mid X \right] \right] \right) \\
& = n^{-1} \sum_{i=1}^{n} E[m_1(x_i)] \\
& = E[Y(1)].
\end{aligned}
$$

Result

- If the response model is correctly specified ($\hat{\pi}(X) = \pi(X)$), then

$$\begin{aligned}
&E[\hat{\mu}_{1,\mathsf{DR}}] \\
&= n^{-1} \sum_{i=1}^{n} E\left[\left(1 - \frac{A_i}{\pi(X_i)}\right) m_1(x_i) + \frac{A_i}{\pi(X_i)} Y_i\right] \\
&= n^{-1} \sum_{i=1}^{n} E\left[E\left[\left(1 - \frac{A_i}{\pi(X_i)}\right) m_1(x_i) + \frac{A_i}{\pi(X_i)} Y_i \mid X, Y\right]\right] \\
&= n^{-1} \sum_{i=1}^{n} E\left[\left(1 - \frac{\pi(X_i)}{\pi(X_i)}\right) m_1(x_i) + \frac{\pi(X_i)}{\pi(X_i)} Y_i(1)\right] \\
&= E[Y(1)].
\end{aligned}$$

Instrumental Variables (IVs)

- Overview of Instrumental Variables
- Instrumental Variables in Causal Inference

Ordinary Least Squares

- In OLS, we consider the model,

$$Y = X\beta + \varepsilon \tag{1}$$

  where $E[\varepsilon \mid X] = 0$ and $X \perp \varepsilon$.
- However, what if $\mathrm{Cov}(X, \varepsilon) \neq 0$?
- We say a variable X_k is **endogenous** if $\mathrm{Cov}(X_k, \varepsilon) \neq 0$.
- A variable X_k is **exogenous** if $X_k \perp \varepsilon$.

Modifying Previous Assumptions

- We previous discussed the assumptions of the potential outcomes framework. One of them was: No Unmeasured Confounders (NUC),

$$Y(1), Y(0) \perp A \mid X.$$

- If a variable X_k is endogenous, then the model does *not* satisfy the NUC condition.

Parametric Models

- Consider the following linear model:[1]

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

  with $x_1, x_2 \perp \varepsilon$ but $x_3 \not\perp \varepsilon$

- To estimate β_3 we need an instrumental variable.

[1] Example taken from (Wooldridge 2010).

Instrumental Variables

- A variable z_1 is an **instrumental variable** (IV) if it satisfies:

$$\mathrm{Cov}(z_1, \varepsilon) = 0 \tag{2}$$

$$\mathrm{Cov}(z_1, x_3) \neq 0 \tag{3}$$

- This makes sense because we want it to be exogenous with respect to Equation 1, yet we need it to influence x_3 if we are going to measure β_3.
- Note, that Equation 2 *cannot* be tested but Equation 3 can and should be tested.

Reduced Form Equations

- When we have an instrument z_1, we can estimate:

$$\hat{x}_3 = \hat{\gamma}_0 + \hat{\gamma}_1 x_1 + \hat{\gamma}_2 x_2 + \hat{\theta} z_1$$
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 \hat{x}_3$$

- This framework is called **two-stage least squares** (2SLS).
- This can be generalized to have K exogenous x_i variables and L instruments z_j.

Identification

- Then the IV solves the identification problem.
- Let $z = (x_1, x_2, z_1)$.
- Equation 2 implies that $E[z'\varepsilon] = 0$.
- The normal equations for the IV estimator are:

$$E[z'x]\beta = E[z'y].$$

- This has a unique solution if $E[z'x]$ has full rank, which happens if Equation 3 is satisfied.

Results for 2SLS

Under regularity conditions, 2SLS is

- consistent,
- asymptotically normal, and
- asymptotically efficient.

See (Wooldridge 2010), Chapter 5 for these proofs.

Problems with IVs

- Bias
- Weak instruments

Instrumental Variables

- Overview of Instrumental Variables
- Instrumental Variables in Causal Inference

Causal Models with No Unconfoundedness

- Suppose that we have the model,[2]

$$Y_i(a) = Y_i(0) + \tau A_i.$$

We can also express this as

$$Y_i = \alpha + A_i\tau + \varepsilon_i$$

- We do *not* use the NUC. So

$$Y(1), Y(0) \not\perp A \mid X.$$

- Notice that OLS does not work because

$$\tau_{OLS} = \frac{\mathrm{Cov}(Y_i, A_i)}{\mathrm{Var}\, A_i} = \frac{\mathrm{Cov}(\tau A_i + \varepsilon, A_i)}{\mathrm{Var}\, A_i} = \tau + \frac{\mathrm{Cov}(\varepsilon, A_i)}{\mathrm{Var}\, A_i}$$

[2]The rest of the slides were based off of Stefan Wager's S361 Causal Inference Notes (Wager 2020).

Causal Models with IVs

- We can add an instrument and have something similar to 2SLS,

$$\begin{aligned} Y_i &= \alpha + A_i\tau + \varepsilon_i \\ A_i &= Z_i\gamma + \eta_i \qquad \varepsilon_i \perp Z_i. \end{aligned}$$

- Then

$$\operatorname{Cov}(Y_i, Z_i) = \operatorname{Cov}(A_i\tau + \varepsilon_i, Z_i) = \tau \operatorname{Cov}(A_i, Z_i).$$

- Hence,

$$\tau = \frac{\operatorname{Cov}\left(Y_i, Z_i\right)}{\operatorname{Cov}(A_i, Z_i)}.$$

Optimal Instruments

- If Z is a d-dimensional vector then we have

$$\tau = \frac{\mathrm{Cov}\left(Y_i, w(Z_i)\right)}{\mathrm{Cov}(A_i, w(Z_i))}$$

  where $w : \mathbb{R}^d \to \mathbb{R}$.
- The optimal choice of $w(\cdot)$ that minimizes the variance of τ, is

$$w^*(Z) \propto E[A \mid Z].$$

Estimation

The previous slide suggests the following estimation strategy:

1. Estimate $\hat{w}(\cdot) = E[A \mid Z]$ nonparametrically, and then
2. Estimate the covariances using $\hat{w}$,

$$\hat{\tau} = \frac{\hat{\mathrm{Cov}}(Y_i, \hat{w}(Z_i))}{\hat{\mathrm{Cov}}(A_i, \hat{w}(Z_i))}$$

However, this can fail from overfitting with weak instruments.

Cross Fitting

A better strategy is to use cross-fitting, and solve

$$\hat{\tau} = \frac{\hat{\mathrm{Cov}}(Y_i, \hat{w}^{k(-i)}(Z_i))}{\hat{\mathrm{Cov}}(A_i, \hat{w}^{k(-i)}(Z_i))}$$

where $\hat{w}^{k(-i)}$ is the estimation of $\hat{w}$ on the k-th fold in which element i is missing.

Previously

- Continuous Treatment Effects
- Covariate Balancing
- Multiple Causes
- Dynamic Treatment Regimes
- Spatial Confounding
- Semiparametric IVs

References I

Wager, Stefan (2020). *Stats 361: Causal inference*.

Wooldridge, Jeffrey M (2010). *Econometric analysis of cross section and panel data*. MIT press.