Causal Inference
CS 477-677

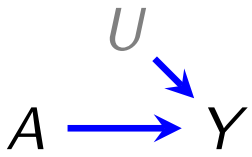
Instrumental Variables

Ilya Shpitser

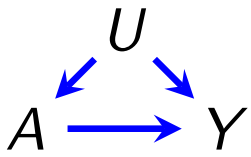
Outline

Last Time

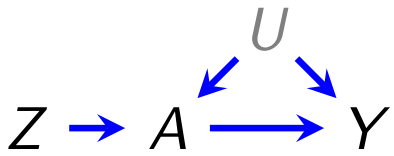
- Randomize (Daniel 1-15, Lind, Pierce, Neyman, Fisher):



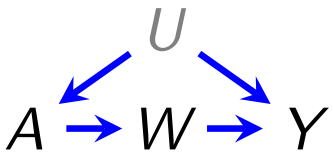
- Observe confounders/stratify, last time:



- Instrumental variable + assumptions (P. Wright, 1928), today:

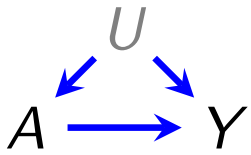


- Find a strong independent mediator (Pearl), later:



Why Do We Need A Trick?

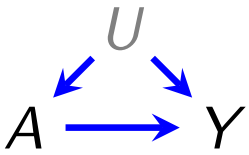
- Wait a second... this is a latent variable model:



$$p(Y, A) = \sum_u p(Y \mid A, U) p(A \mid U) p(U)$$

Why Do We Need A Trick?

- Wait a second... this is a latent variable model:

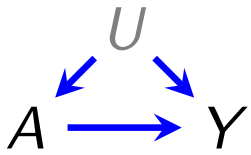


$$p(Y, A) = \sum_u p(Y \mid A, U) p(A \mid U) p(U)$$

- Why can we not do this:
  - Posit a model $p(Y, A, U; \alpha)$.

Why Do We Need A Trick?

- Wait a second... this is a latent variable model:

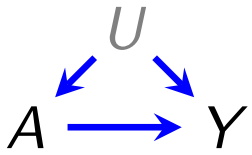


$$p(Y, A) = \sum_u p(Y \mid A, U) p(A \mid U) p(U)$$

- Why can we not do this:
  - Posit a model $p(Y, A, U; \alpha)$.
  - Use a latent variable fitting algorithm (expectation maximization) to find $\hat{\alpha}$.

Why Do We Need A Trick?

- Wait a second... this is a latent variable model:



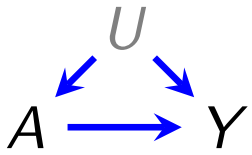
$$p(Y, A) = \sum_u p(Y \mid A, U) p(A \mid U) p(U)$$

- Why can we not do this:
  - Posit a model $p(Y, A, U; \alpha)$.
  - Use a latent variable fitting algorithm (expectation maximization) to find $\hat{\alpha}$.
  - Use adjustment formula with resulting parameters:

$$p(Y(a)) = \sum_U p(Y \mid A, U; \hat{\alpha}) p(U; \hat{\alpha}).$$

Why Do We Need A Trick?

- Wait a second... this is a latent variable model:



$$p(Y, A) = \sum_u p(Y \mid A, U) p(A \mid U) p(U)$$

- Why can we not do this:
  - Posit a model $p(Y, A, U; \alpha)$.
  - Use a latent variable fitting algorithm (expectation maximization) to find $\hat{\alpha}$.
  - Use adjustment formula with resulting parameters:

$$p(Y(a)) = \sum_U p(Y \mid A, U; \hat{\alpha}) p(U; \hat{\alpha}).$$

- Unfortunately, this doesn't work.

The Identification Problem

- Observed data on $\vec{X}$, observed data model $p(\vec{X}; \alpha)$.
- "Larger model" of interest: $p(\vec{X}, \vec{W}; \beta)$.
- Could be: latent variable model, causal model, etc.
- Larger model induces observed data model:

$$p(\vec{X}; \alpha) = f(p(\vec{X}, \vec{W}; \beta)).$$

- Interested in parameter β_i in β.
- β_i is said to be identified from observed data if for any $p(\vec{X}, \vec{W}; \beta)$ in the model, β_i is a function of $p(\vec{X}; \alpha)$.
- In other words, we want to "invert" f:

$$\beta_i = g(p(\vec{X}; \alpha)).$$

- Abstract, so let's do an example.

Identification Example: Ignorable Causal Model

- Observed data on A, Y, observed data model $p(Y, A; \alpha)$.
- Larger model: causal model on $p(Y(1), Y(0), A; \beta)$.
- Remember, we don't get to see all of the larger model, but have assumptions to help us:

$$Y(A) = Y$$
$$\{Y(1), Y(0)\} \perp\!\!\!\perp A$$

- Parameter of interest $\beta_i = E[Y(1)] - E[Y(0)]$.

Identification Example: Ignorable Causal Model

- Observed data on A, Y, observed data model $p(Y, A; \alpha)$.
- Larger model: causal model on $p(Y(1), Y(0), A; \beta)$.
- Remember, we don't get to see all of the larger model, but have assumptions to help us:

$$Y(A) = Y$$
$$\{Y(1), Y(0)\} \perp\!\!\!\perp A$$

- Parameter of interest $\beta_i = E[Y(1)] - E[Y(0)]$.
- Parameter identified:

$$\beta_i = E[Y \mid A = 1] - E[Y \mid A = 0].$$

Identification Example: Ignorable Causal Model

- Observed data on A, Y, observed data model $p(Y, A; \alpha)$.
- Larger model: causal model on $p(Y(1), Y(0), A; \beta)$.
- Remember, we don't get to see all of the larger model, but have assumptions to help us:

$$Y(A) = Y$$
$$\{Y(1), Y(0)\} \perp\!\!\!\perp A$$

- Parameter of interest $\beta_i = E[Y(1)] - E[Y(0)]$.
- Parameter identified:

$$\beta_i = E[Y \mid A = 1] - E[Y \mid A = 0].$$

- If observed data model is a linear regression: $Y = w_0 + w_1 A + \epsilon$,

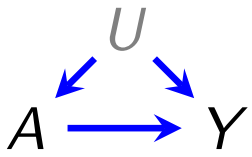
$$\beta_i = w_1.$$

Identification And Estimation

- Formally, have to make sure parameter is identified before estimating.
- Otherwise, parameter is not a unique function of observed data, so estimation problem is not well-posed.
- In classical statistics, parameters are often identified, so the issue is not explicitly discussed.
- Sometimes parameters are not identified.
- Classical example in machine learning: labels in mixture models.
- That type of non-identification is benign (no real effect on decision making).
- In causal inference, failure of identification is common and very important!

Example of Unidentified Parameter

- Observed data on A, Y, observed data model $p(Y, A; \alpha)$.
- Larger model: causal model on $p(Y(1), Y(0), A, U; \beta)$.
- As before, we don't get to see all of the larger model



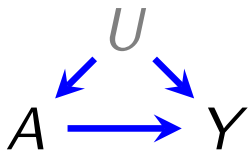
- Consistency holds, ignorability does not:

$$Y(A) = Y$$
$$\{Y(1), Y(0)\} \not\perp\!\!\!\perp A$$
$$\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid U$$

- Parameter of interest (as before): $\beta_i = E[Y(1)] - E[Y(0)]$.
- β_i is not identified in this model. Why?

Non-identification (Counterexample)

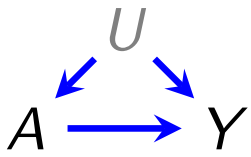
- Will construct two models that agree on $p(A, Y)$, disagree on $p(Y(a))$, for:



- Both models: U drawn from a fair coin, $A = U$.
- Model 1: $Y = A$ xor $U + \mathcal{N}(0, 1)$. Model 2: $Y = \mathcal{N}(0, 1)$.
- Observed data (both models): A and Y are independent, A is a fair coin, Y is a standard normal $\mathcal{N}(0, 1)$.
- Model 1: $Y(1) \sim \mathcal{N}(0, 1) +$ fair coin flip,
  Model 2: $Y(1) \sim \mathcal{N}(0, 1)$.

Non-identification (Counterexample)

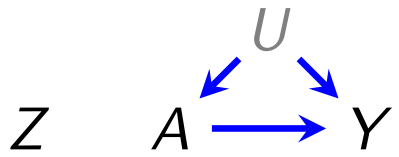
- Will construct two models that agree on $p(A, Y)$, disagree on $p(Y(a))$, for:



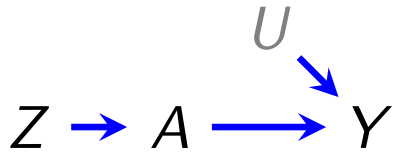
- Both models: U drawn from a fair coin, $A = U$.
- Model 1: $Y = A$ xor $U + \mathcal{N}(0, 1)$. Model 2: $Y = \mathcal{N}(0, 1)$.
- Observed data (both models): A and Y are independent, A is a fair coin, Y is a standard normal $\mathcal{N}(0, 1)$.
- Model 1: $Y(1) \sim \mathcal{N}(0, 1) +$ fair coin flip,
  Model 2: $Y(1) \sim \mathcal{N}(0, 1)$.
- Lesson: no latent variable modeling approach will get you $Y(a)$.
- $Y(a)$ is not a function of $p(Y, A)$, and thus a model $\sum_u p(Y, A, U = u; \alpha)$ will not help.

Why Do Instrumental Variables Help?

- If $Z \perp\!\!\!\perp A$, we are back to previous problem:



- Imagine Z was a really strong predictor of A:



- Then confounding goes away, and $p(Y(a)) = p(Y \mid A = a)$.
- What if we are somewhere in the middle?

Instrumental Variable Conditions

- Already saw picture of the IV model.
- Let's state assumptions on potential outcomes.
  - Relevance: Z is associated with A: $Z \not\perp\!\!\!\perp A$.
  - Marginal ignorability: $Y(a, z) \perp\!\!\!\perp Z$.
  - Exclusion restriction: Z does not cause Y directly $Y(a, z) = Y(a, z')$ for all a, z, z'.
- Want to randomize A, but cannot. Can randomize Z related to A in the appropriate way.
- Examples: genes (Mendelian randomization), intent vs implementation, regression discontinuity.

IV In Linear Models

- Assume linear regressions: for A and Y:

$$A = v_0 + v_1 \cdot Z + \epsilon_1$$
$$Y = u_0 + u_1 \cdot A + \epsilon_2$$
$$Y = w_0 + w_1 \cdot Z + \epsilon_3$$

- Claim:

$$E[Y(1)] - E[Y(0)] = \frac{E[Y(z=1)] - E[Y(z=0)]}{E[A(z=1)] - E[A(z=0)]} = w_1/v_1.$$

IV In Linear Models

- Assume linear regressions: for A and Y:

$$A = v_0 + v_1 \cdot Z + \epsilon_1$$
$$Y = u_0 + u_1 \cdot A + \epsilon_2$$
$$Y = w_0 + w_1 \cdot Z + \epsilon_3$$

- Claim:

$$E[Y(1)] - E[Y(0)] = \frac{E[Y(z=1)] - E[Y(z=0)]}{E[A(z=1)] - E[A(z=0)]} = w_1/v_1.$$

- Intuition: path analysis (Sewall Wright).

On the board

Path Analysis

- Imagine linear models for a chain:

$$A = v_0 + v_1 \cdot Z + \epsilon_1$$
$$Y = u_0 + u_1 \cdot A + \epsilon_2$$
$$Y = w_0 + w_1 \cdot Z + \epsilon_3$$

$$Z \longrightarrow A \longrightarrow Y$$

Path Analysis

- Imagine linear models for a chain:

$$A = v_0 + v_1 \cdot Z + \epsilon_1$$
$$Y = u_0 + u_1 \cdot A + \epsilon_2$$
$$Y = w_0 + w_1 \cdot Z + \epsilon_3$$

$$Z \longrightarrow A \longrightarrow Y$$

- $E[Y(z=1)] - E[Y(z=0)] = u_0 + u_1 \cdot (v_0 + v_1) - u_0 - u_1 \cdot (v_0) = u_1 \cdot v_1.$

Path Analysis

- Imagine linear models for a chain:

$$A = v_0 + v_1 \cdot Z + \epsilon_1$$
$$Y = u_0 + u_1 \cdot A + \epsilon_2$$
$$Y = w_0 + w_1 \cdot Z + \epsilon_3$$

$$Z \longrightarrow A \longrightarrow Y$$

- $E[Y(z=1)] - E[Y(z=0)] = u_0 + u_1 \cdot (v_0 + v_1) - u_0 - u_1 \cdot (v_0) = u_1 \cdot v_1$.
- $E[Y(a=1)] - E[Y(a=0)] = u_0 + u_1 - u_0 - u_1 \cdot 0 = u_1$.

Path Analysis

- Imagine linear models for a chain:

$$A = v_0 + v_1 \cdot Z + \epsilon_1$$
$$Y = u_0 + u_1 \cdot A + \epsilon_2$$
$$Y = w_0 + w_1 \cdot Z + \epsilon_3$$

$$Z \longrightarrow A \longrightarrow Y$$

- $E[Y(z=1)] - E[Y(z=0)] = u_0 + u_1 \cdot (v_0 + v_1) - u_0 - u_1 \cdot (v_0) = u_1 \cdot v_1.$
- $E[Y(a=1)] - E[Y(a=0)] = u_0 + u_1 - u_0 - u_1 \cdot 0 = u_1.$
- $E[A(z=1)] - E[A(z=0)] = v_0 + v_1 - v_0 - v_1 \cdot 0 = v_1.$

Path Analysis

- Imagine linear models for a chain:

$$A = v_0 + v_1 \cdot Z + \epsilon_1$$
$$Y = u_0 + u_1 \cdot A + \epsilon_2$$
$$Y = w_0 + w_1 \cdot Z + \epsilon_3$$

$$Z \longrightarrow A \longrightarrow Y$$

- $E[Y(z=1)] - E[Y(z=0)] = u_0 + u_1 \cdot (v_0 + v_1) - u_0 - u_1 \cdot (v_0) = u_1 \cdot v_1$.
- $E[Y(a=1)] - E[Y(a=0)] = u_0 + u_1 - u_0 - u_1 \cdot 0 = u_1$.
- $E[A(z=1)] - E[A(z=0)] = v_0 + v_1 - v_0 - v_1 \cdot 0 = v_1$.
- Since $u_1 = (u_1 \cdot v_1)/v_1$, we have:

$$E[Y(a=1)] - E[Y(a=0)] = \frac{E[Y(z=1)] - E[Y(z=0)]}{E[A(z=1)] - E[A(z=0)]}.$$

Path Analysis

- Imagine linear models for a chain:

$$A = v_0 + v_1 \cdot Z + \epsilon_1$$
$$Y = u_0 + u_1 \cdot A + \epsilon_2$$
$$Y = w_0 + w_1 \cdot Z + \epsilon_3$$

$$Z \longrightarrow A \longrightarrow Y$$

- $E[Y(z=1)] - E[Y(z=0)] = u_0 + u_1 \cdot (v_0 + v_1) - u_0 - u_1 \cdot (v_0) = u_1 \cdot v_1$.
- $E[Y(a=1)] - E[Y(a=0)] = u_0 + u_1 - u_0 - u_1 \cdot 0 = u_1$.
- $E[A(z=1)] - E[A(z=0)] = v_0 + v_1 - v_0 - v_1 \cdot 0 = v_1$.
- Since $u_1 = (u_1 \cdot v_1)/v_1$, we have:

$$E[Y(a=1)] - E[Y(a=0)] = \frac{E[Y(z=1)] - E[Y(z=0)]}{E[A(z=1)] - E[A(z=0)]}.$$

- What if $cov(\epsilon_1, \epsilon_2) \neq 0$? Same derivation.

Weakening Assumptions

- If the model is entirely linear, can deal with unobserved U!
- Very strong assumption.
- Can we weaken this?
- Yes, assume effect independent of Z given A.
- Effect: $E[Y(1) - Y(0)]$, so for $a \in \{0, 1\}$,

$$E[Y(1) - Y(0)|Z = 1, A = a] = E[Y(1) - Y(0)|Z = 0, A = a].$$

- Assumption implies usual identification:

$$E[Y(a = 1)] - E[Y(a = 0)] = \frac{E[Y(z = 1)] - E[Y(z = 0)]}{E[A(z = 1)] - E[A(z = 0)]}.$$

- If interested, can read proof in technical point 16.3 in HR.

IV Models and Non-Compliance

- Non-compliance means units don't do what they are told.
- Common issue in certain randomized trials.
- Arm assignment is randomized, but actual drugs aren't always taken in active arm.
- Can view this as an IV model.
- Confounding by "type of person":
  - "Compliers": $A(z = 1) = 1, A(z = 0) = 0$.
  - "Always takers": $A(z = 1) = A(z = 0) = 1$.
  - "Never takers": $A(z = 1) = A(z = 0) = 0$.
  - "Defiers": $A(z = 1) = 0$, $A(z = 0) = 1$.
- Often assume lack of defiers, a type of monotonicity assumption.

Complier-Specific Causal Effect

- If no defiers, can obtain
  $E[Y(1) - Y(0) \mid A(z=1) = 1, A(z=0) = 0]$ via usual estimator

$$\frac{E[Y(z=1)] - E[Y(z=0)]}{E[A(z=1)] - E[A(z=0)]}.$$

- Proof due to Imbens and Angrist (1994), see technical point 16.5 in HR.

Two Stage Least Squares

- Fit $E[A \mid Z; \alpha] = \alpha_0 + \alpha_1 \cdot Z$ by MLE to obtain $\hat{\alpha}$.
- Fit $E[Y \mid Z; \beta] = \beta_0 + \beta_1 E[\hat{A} \mid Z; \hat{\alpha}]$ by MLE to obtain $\hat{\beta}$. Then

$$\hat{\beta}_1 = \frac{E[Y(z=1)] - E[Y(z=0)]}{E[A(z=1)] - E[A(z=0)]},$$

  interpreted as before.
- Same idea with a set of baseline variables $\vec{X}$, just include them in both models.

When Identification Fails

- Two big problems with IV:
  - Reliance on parametric assumptions (even with infinite data).
  - Output very sensitive to violations.
- What can we do if assumptions are false?
- Then ACE is not identified (not a function of observed data).
- Then we are dead...

When Identification Fails

- Two big problems with IV:
  - Reliance on parametric assumptions (even with infinite data).
  - Output very sensitive to violations.
- What can we do if assumptions are false?
- Then ACE is not identified (not a function of observed data).
- Then we are dead... but maybe only mostly dead!

When Identification Fails

- Two big problems with IV:
  - Reliance on parametric assumptions (even with infinite data).
  - Output very sensitive to violations.
- What can we do if assumptions are false?
- Then ACE is not identified (not a function of observed data).
- Then we are dead... but maybe only mostly dead!
- IV model is not an arbitrary model (does not contain all distributions $p(A, Y, Z)$).
- That means we may be able to restrict ACE to a subset of possible values.
- In other words, try to find bounds for the ACE.

IV Bounds (Balke And Pearl)

- Let $p_{00.0} = p(Y = 0, A = 0 \mid Z = 0)$ (same for other values).
- Can show using computer algebra ACE lies within the following polytope:

$$\left\{ \begin{array}{c} p_{00.0} + p_{11.1} - 1 \\ p_{00.1} + p_{11.1} - 1 \\ p_{11.0} + p_{00.1} - 1 \\ p_{00.0} + p_{11.0} - 1 \\ 2p_{00.0} + p_{11.0} + p_{10.1} + p_{11.1} - 2 \\ p_{00.0} + p_{11.0} + p_{00.1} + p_{01.1} - 2 \\ p_{10.0} + p_{11.0} + p_{00.1} + p_{11.1} - 2 \\ p_{00.0} + p_{01.0} + p_{00.1} + p_{11.1} - 2 \end{array} \right\} \leq \text{ACE} \leq \left\{ \begin{array}{c} 1 - p_{10.0} - p_{01.1} \\ 1 - p_{01.0} - p_{10.1} \\ 1 - p_{01.0} - p_{10.0} \\ 1 - p_{01.1} - p_{10.1} \\ 2 - 2p_{01.0} - p_{10.0} - p_{10.1} - p_{11.1} \\ 2 - 2p_{01.0} - p_{10.0} - p_{00.1} - p_{01.1} \\ 2 - 2p_{10.0} - p_{11.0} - p_{01.1} - p_{10.1} \\ 2 - 2p_{00.0} - p_{01.0} - p_{01.1} - p_{10.1} \end{array} \right\}$$

- Sometimes these are informative (e.g. bound ACE from 0).
- Often very wide, however.
- This is the best we can do non-parametrically.

IV Bounds (Balke And Pearl)

- Let $p_{00.0} = p(Y = 0, A = 0 \mid Z = 0)$ (same for other values).
- Can show using computer algebra ACE lies within the following polytope:

$$\left\{ \begin{array}{c} p_{00.0} + p_{11.1} - 1 \\ p_{00.1} + p_{11.1} - 1 \\ p_{11.0} + p_{00.1} - 1 \\ p_{00.0} + p_{11.0} - 1 \\ 2p_{00.0} + p_{11.0} + p_{10.1} + p_{11.1} - 2 \\ p_{00.0} + p_{11.0} + p_{00.1} + p_{01.1} - 2 \\ p_{10.0} + p_{11.0} + p_{00.1} + p_{11.1} - 2 \\ p_{00.0} + p_{01.0} + p_{00.1} + p_{11.1} - 2 \end{array} \right\} \leq \text{ACE} \leq \left\{ \begin{array}{c} 1 - p_{10.0} - p_{01.1} \\ 1 - p_{01.0} - p_{10.1} \\ 1 - p_{01.0} - p_{10.0} \\ 1 - p_{01.1} - p_{10.1} \\ 2 - 2p_{01.0} - p_{10.0} - p_{10.1} - p_{11.1} \\ 2 - 2p_{01.0} - p_{10.0} - p_{00.1} - p_{01.1} \\ 2 - 2p_{10.0} - p_{11.0} - p_{01.1} - p_{10.1} \\ 2 - 2p_{00.0} - p_{01.0} - p_{01.1} - p_{10.1} \end{array} \right\}$$

- Sometimes these are informative (e.g. bound ACE from 0).
- Often very wide, however.
- This is the best we can do non-parametrically.
- Aside: this is related to the structure of latent variable models. More later.

Next time: Decomposing Causal Effects Into Direct and Indirect Effects.