CSC2541: Introduction to Causality

Lecture 5 - Estimation (cont.) and Instrumental Variables

Instructor: Rahul G. Krishnan

TA: Vahid Balazadeh-Meresht

October 17, 2022

Recap - Lecture 4

- Identification
  - Backdoor criteria: Identical to adjustment via the G-formula,
  - Frontdoor criteria: Using mediators to identify causal effect on outcomes.
- Do-Calculus: Three rules to identify causal effects:
  1. Insertion or deletion of observations : Generalization of d-separation,
  2. Interchanging actions with observations : Generalization of the backdoor criteria,
  3. Insertion or deletion of actions
- Parametric Estimation:
  - Conditional outcome models
  - Grouped conditional outcome models
  - TAR-Net

e

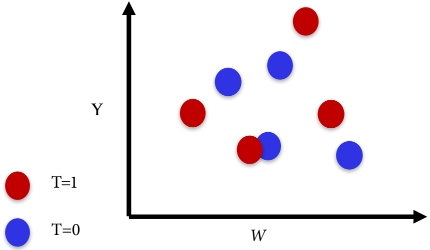
Matching

1. For each observation in the treatment group, find "statistical twins" in the control group with similar covariates X (and vice versa), where X is a valid adjustment set
2. Use the Y values of the matched observations as the counterfactual outcomes for one at hand
3. Estimate average treatment effect as the difference between observed and imputed counterfactual values

Slide credit to David Sontag

Matching

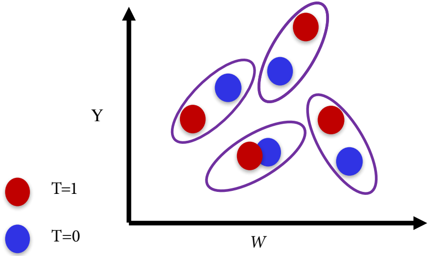
1. For each observation in the treatment group, find "statistical twins" in the control group with similar covariates X (and vice versa), where X is a valid adjustment set
2. Use the Y values of the matched observations as the counterfactual outcomes for one at hand
3. Estimate average treatment effect as the difference between observed and imputed counterfactual values



Slide credit to David Sontag

Matching

1. For each observation in the treatment group, find "statistical twins" in the control group with similar covariates X (and vice versa), where X is a valid adjustment set
2. Use the Y values of the matched observations as the counterfactual outcomes for one at hand
3. Estimate average treatment effect as the difference between observed and imputed counterfactual values



Slide credit to David Sontag

Matching - Formal definition

Let the data $\mathcal{D} = \{(T^i, X^i, Y^i)\}_{i=1}^N$. To estimate the counterfactual Y_0^i for a sample i in the treatment group, we use (similar) samples from the control group $(T = 0)$:

$$\hat{Y}_0^i = \sum_{j \text{ s.t. } T^j=0} w_{ij} Y^j$$

Similarly, to estimate the counterfactual Y_1^i for a sample i in the control group, we use samples from the treatment group:

$$\hat{Y}_1^i = \sum_{j \text{ s.t. } T^j=1} w_{ij} Y^j$$

An estimation of ATE will be

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_i Y_1^i - Y_0^i = \frac{1}{N} \left[\sum_{i;T^i=1} \left(Y^i - \hat{Y}_0^i \right) + \sum_{i;T^i=0} \left(\hat{Y}_1^i - Y^i \right) \right]$$

Different matching algorithms use different definitions of w_{ij}

Types of matching

- **Exact matching**: $w_{ij} = \begin{cases} \frac{1}{k_i} & \text{if } X^i = X^j \\ 0 & o.w. \end{cases}$ with k_i as the number of samples j with $X^i = X^j$
  - Problem: For high-dimensional X, it will be less likely to find an exact match
- **Multivariate distance matching (MDM)**: Use (Euclidean) distance metric to find "close" observations as potential matches
  - We can use KNN algorithm to find the k closes observations in the control (treatment) group for each treated (controlled) sample, i.e.,

$$w_{ij} = \begin{cases} \frac{1}{k} & \text{if } X^j \in \text{KNN}(X^i) \\ 0 & o.w. \end{cases}$$

Matching - Pros and Cons

+ Interpretable, especially in small samples
+ Non-parametric
- KNN-matching can be biased since $X^i \approx X^j \implies Y_0^i \approx Y_0^j, Y_1^i \approx Y_1^j$ (See Abadie and Imbens, 2011 for bias-correction for matching estimators)
- Curse of dimensionality - it gets harder to find good matches as dimension grows

Source: `https://mobile.twitter.com/HallaMartin/status/1569311697717927937`

Propensity scores

- Matching can suffer from curse of dimensionality of X
- Let's look at probability of treatment assignment given X

$$e(X) := P(T = 1|X)$$

Propensity scores

- Matching can suffer from curse of dimensionality of X
- Let's look at probability of treatment assignment given X

$$e(X) := P(T = 1|X)$$

- $e(X)$ summarizes high-dimensional variables X into one dimension!

Theorem - Propensity Score

Assume X satisfies the backdoor criterion (conditional ignorability) w.r.t. T, Y. Given positivity, $e(X)$ will also satisfy conditional ignorability, i.e.,

$$Y_0, Y_1 \perp\!\!\!\perp T|e(X)$$

- Helpful for matching!

Propensity score theorem - Intuition

$$Y_0, Y_1 \perp\!\!\!\perp T | X \implies Y_0, Y_1 \perp\!\!\!\perp T | e(X)$$

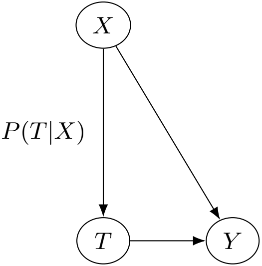
$$\begin{tikzcd} & X \arrow[dl] \arrow[d] \\ T \arrow[r] & Y \end{tikzcd}$$

For the formal proof, see Rosenbaum and Rubin, 1983.

Slide credit to Brady Neal

Propensity score theorem - Intuition

$$Y_0, Y_1 \perp\!\!\!\perp T | X \implies Y_0, Y_1 \perp\!\!\!\perp T | e(X)$$

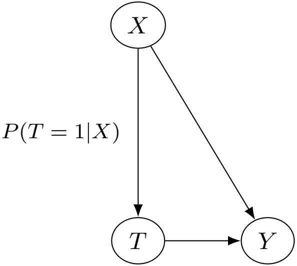


For the formal proof, see Rosenbaum and Rubin, 1983.

Slide credit to Brady Neal

Propensity score theorem - Intuition

$$Y_0, Y_1 \perp\!\!\!\perp T | X \implies Y_0, Y_1 \perp\!\!\!\perp T | e(X)$$

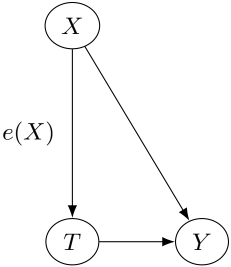


For the formal proof, see Rosenbaum and Rubin, 1983.

Slide credit to Brady Neal

Propensity score theorem - Intuition

$$Y_0, Y_1 \perp\!\!\!\perp T | X \implies Y_0, Y_1 \perp\!\!\!\perp T | e(X)$$

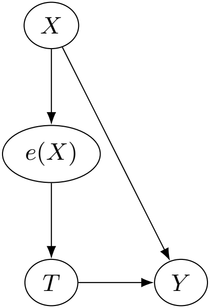


For the formal proof, see Rosenbaum and Rubin, 1983.

Slide credit to Brady Neal

Propensity score theorem - Intuition

$$Y_0, Y_1 \perp\!\!\!\perp T | X \implies Y_0, Y_1 \perp\!\!\!\perp T | e(X)$$

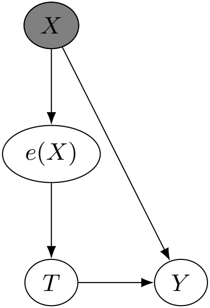


For the formal proof, see Rosenbaum and Rubin, 1983.

Slide credit to Brady Neal

Propensity score theorem - Intuition

$$Y_0, Y_1 \perp\!\!\!\perp T | X \implies Y_0, Y_1 \perp\!\!\!\perp T | e(X)$$

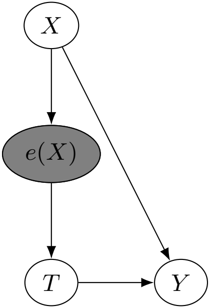


For the formal proof, see Rosenbaum and Rubin, 1983.

Slide credit to Brady Neal

Propensity score theorem - Intuition

$$Y_0, Y_1 \perp\!\!\!\perp T | X \implies Y_0, Y_1 \perp\!\!\!\perp T | e(X)$$



For the formal proof, see Rosenbaum and Rubin, 1983.

Slide credit to Brady Neal

Propensity score matching

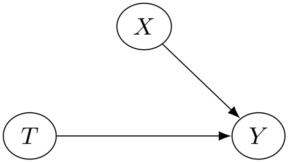
- Instead of computing multivariate distances, we can match the one-dimensional propensity score:
- Step 1: Estimate $e(X)$ using a **parametric** method
- Step 2: Apply a matching algorithm (KNN) with distance $|e(X_i) - e(X_j)|$

Propensity score matching

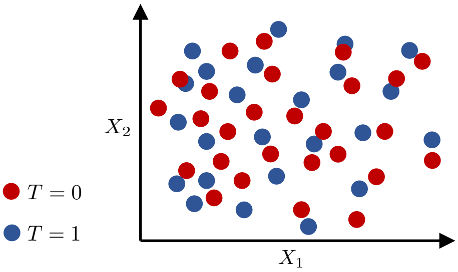
- Instead of computing multivariate distances, we can match the one-dimensional propensity score:
- Step 1: Estimate $e(X)$ using a **parametric** method
- Step 2: Apply a matching algorithm (KNN) with distance $|e(X_i) - e(X_j)|$
- This is not a magic, we still need to estimate $P(T = 1|X)$!
- A perfect predictor of T is not always good - we can include more variables as X to get better treatment assignment predictions
  - Can increase variance,
  - See "Why Propensity Scores Should Not Be Used for Matching" by King and Nielsen, 2019.

Inverse probability weighting (IPW)

- Causal estimation in RCTs is easier (control and treatment groups are similar)

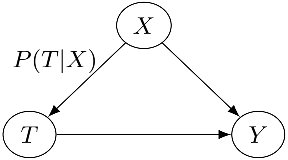


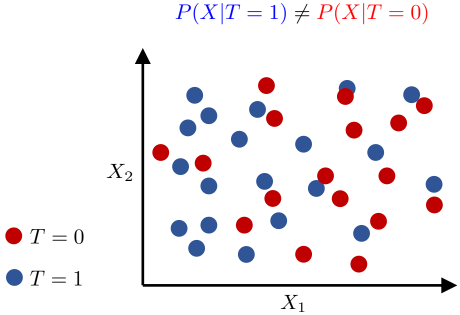
$P(X|T=1) = P(X|T=0)$



Inverse probability weighting (IPW)

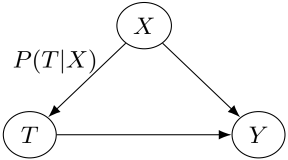
- In observational studies, however, the treatment and control groups are not comparable.

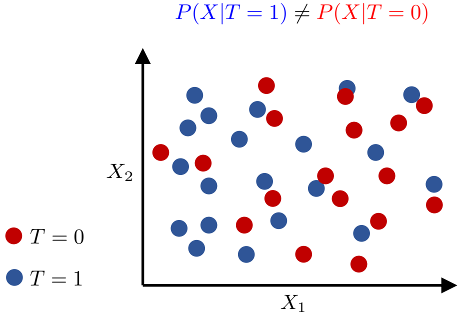




Inverse probability weighting (IPW)

- In observational studies, however, the treatment and control groups are not comparable. Can we make a pseudo-RCT by re-weighting samples?



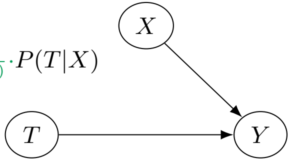


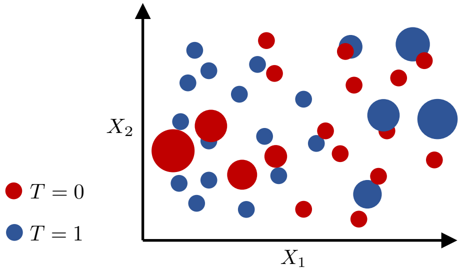
Inverse probability weighting (IPW)

- In observational studies, however, the treatment and control groups are not comparable. Can we make a pseudo-RCT by re-weighting samples?

$$w_1(X) \cdot P(X|T=1) \approx w_0(X) \cdot P(X|T=0)$$

$\frac{1}{P(T|X)} \cdot P(T|X)$





Samples re-weighted by the inverse propensity score of the treatment they received

Inverse probability weighting (IPW) - Formal

$$\mathbb{E}[Y_t] = \mathbb{E}_X\left[\mathbb{E}[Y|X, T=t]\right] \qquad \text{(conditional ignorability)}$$

Inverse probability weighting (IPW) - Formal

$$\begin{aligned}
\mathbb{E}[Y_t] &= \mathbb{E}_X\left[\mathbb{E}[Y|X, T=t]\right] && \text{(conditional ignorability)} \\
&= \sum_x \mathbb{E}[Y|X=x, T=t] P(X=x) \\
&= \sum_x \sum_y y P(y|x,t) P(x)
\end{aligned}$$

Inverse probability weighting (IPW) - Formal

$$
\begin{aligned}
\mathbb{E}[Y_t] &= \mathbb{E}_X\left[\mathbb{E}[Y|X,T=t]\right] && \text{(conditional ignorability)} \\
&= \sum_x \mathbb{E}[Y|X=x,T=t]P(X=x) \\
&= \sum_x \sum_y yP(y|x,t)P(x) \\
&= \sum_x \sum_y yP(y|x,t)P(x)\frac{P(t|x)}{P(t|x)} \\
&= \sum_{x,y} \frac{1}{P(t|x)} yP(x,y,t) && (P(y|x,t)P(x)P(t|x) = P(x,y,t))
\end{aligned}
$$

Inverse probability weighting (IPW) - Formal

$$
\begin{aligned}
\mathbb{E}[Y_t] &= \mathbb{E}_X\left[\mathbb{E}[Y|X, T=t]\right] && \text{(conditional ignorability)} \\
&= \sum_x \mathbb{E}[Y|X=x, T=t]P(X=x) \\
&= \sum_x \sum_y yP(y|x,t)P(x) \\
&= \sum_x \sum_y yP(y|x,t)P(x)\frac{P(t|x)}{P(t|x)} \\
&= \sum_{x,y} \frac{1}{P(t|x)} yP(x,y,t) && (P(y|x,t)P(x)P(t|x) = P(x,y,t)) \\
&= \sum_{x,y,t'} \underbrace{\frac{\mathbb{I}(t'=t)}{P(t|x)}y}_{f(x,y,t')} P(x,y,t') && \text{(sum over } T\text{)} \\
&= \sum_{x,y,t'} f(x,y,t')P(x,y,t')
\end{aligned}
$$

Inverse probability weighting (IPW) - Formal

$$
\begin{aligned}
\mathbb{E}[Y_t] &= \mathbb{E}_X\left[\mathbb{E}[Y|X, T=t]\right] && \text{(conditional ignorability)} \\
&= \sum_x \mathbb{E}[Y|X=x, T=t] P(X=x) \\
&= \sum_x \sum_y y P(y|x,t) P(x) \\
&= \sum_x \sum_y y P(y|x,t) P(x) \frac{P(t|x)}{P(t|x)} \\
&= \sum_{x,y} \frac{1}{P(t|x)} y P(x,y,t) && (P(y|x,t)P(x)P(t|x) = P(x,y,t)) \\
&= \sum_{x,y,t'} \underbrace{\frac{\mathbb{I}(t'=t)}{P(t|x)} y}_{f(x,y,t')} P(x,y,t') && \text{(sum over } T\text{)} \\
&= \sum_{x,y,t'} f(x,y,t') P(x,y,t') \\
&= \mathbb{E}\left[f(X,Y,T)\right] = \mathbb{E}\left[\frac{\mathbb{I}(T=t)Y}{P(t|X)}\right]
\end{aligned}
$$

Inverse probability weighting (IPW) - Formal

- Hence,

$$\begin{aligned}\text{ATE} = \mathbb{E}\left[Y_1 - Y_0\right] &= \mathbb{E}\left[\frac{\mathbb{I}(T=1)Y}{P(T=1|X)}\right] - \mathbb{E}\left[\frac{\mathbb{I}(T=0)Y}{P(T=0|X)}\right] \\ &= \mathbb{E}\left[\frac{\mathbb{I}(T=1)Y}{e(X)}\right] - \mathbb{E}\left[\frac{\mathbb{I}(T=0)Y}{1-e(X)}\right]\end{aligned}$$

Inverse probability weighting (IPW) - Formal

- Hence,

$$\begin{aligned}\text{ATE} = \mathbb{E}\left[Y_1 - Y_0\right] &= \mathbb{E}\left[\frac{\mathbb{I}(T=1)Y}{P(T=1|X)}\right] - \mathbb{E}\left[\frac{\mathbb{I}(T=0)Y}{P(T=0|X)}\right] \\ &= \mathbb{E}\left[\frac{\mathbb{I}(T=1)Y}{e(X)}\right] - \mathbb{E}\left[\frac{\mathbb{I}(T=0)Y}{1-e(X)}\right]\end{aligned}$$

- For a given dataset $\mathcal{D} = \{(x^i, t^i, y^i)\}_{i=1}^N$, an estimate of ATE will be

$$\widehat{\text{ATE}} = \frac{1}{N_1}\sum_{i;t^i=1}\frac{y^i}{\hat{e}(x^i)} - \frac{1}{N_0}\sum_{i;t^i=0}\frac{y^i}{1-\hat{e}(x^i)}$$

for $N_1 = |\{i; t^i = 1\}|, N_0 = N - N_1$.

Inverse probability weighting (IPW) - Formal

- Hence,

$$\begin{aligned} \text{ATE} = \mathbb{E}\left[Y_1 - Y_0\right] &= \mathbb{E}\left[\frac{\mathbb{I}(T=1)Y}{P(T=1|X)}\right] - \mathbb{E}\left[\frac{\mathbb{I}(T=0)Y}{P(T=0|X)}\right] \\ &= \mathbb{E}\left[\frac{\mathbb{I}(T=1)Y}{e(X)}\right] - \mathbb{E}\left[\frac{\mathbb{I}(T=0)Y}{1-e(X)}\right] \end{aligned}$$

- For a given dataset $\mathcal{D} = \{(x^i, t^i, y^i)\}_{i=1}^N$, an estimate of ATE will be

$$\widehat{\text{ATE}} = \frac{1}{N_1} \sum_{i;t^i=1} \frac{y^i}{\hat{e}(x^i)} - \frac{1}{N_0} \sum_{i;t^i=0} \frac{y^i}{1-\hat{e}(x^i)}$$

for $N_1 = |\{i; t^i = 1\}|, N_0 = N - N_1$.

- Still we need to estimate $e(X)$. If positivity is violated, propensity scores become non-informative and miscalibrated
- Small propensity scores can create large variance/errors

Questions?

Question

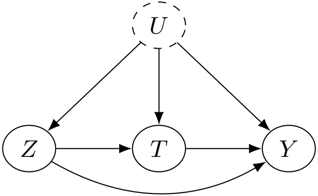
Any questions on weighting based estimators?

Instrumental Variables

- Unobserved confounding (variables that we know exist, but do not observe) is a real concern when attempting to identify causal effects in practical scenarios,
- In such scenarios, we might be able to rely on the use of *instruments* to help us,
- Instruments can be thought of as random variables in a causal Bayesian network that:
  - Are independent of unobserved confounding and,
  - Are related to the outcome *only through* the treatment,

Motivation via causal Bayesian networks

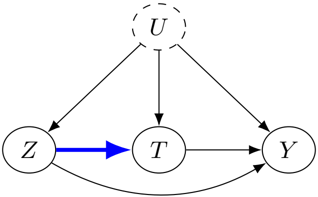
- Consider the following complete graph with unobserved U and observed Z (which as we'll see is the instrument variable),
- We care about estimating the causal effect of T on Y,
- The causal effect of T on Y is non-identifiable (why?). We'll make *assumptions* to make causal inference feasible:



Presentation credit to Brady Neal

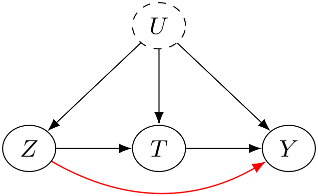
Assumption 1: Relevance

- First, we'll need to assume that there exists an edge from Z to T,
- This is saying the instrument has an effect on treatment.



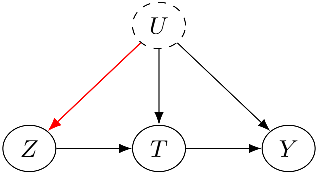
Assumption 2: Exclusion Restriction

- Next, we'll need to assume that there is no edge from Z to Y,
- This is equivalent to saying that the only effect that Z can have on Y is through T.



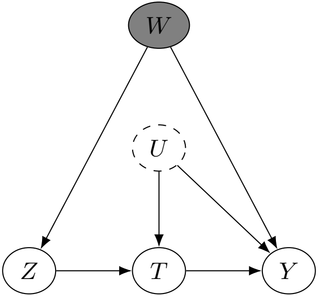
Assumption 3: Instrumental Unconfoundedness

- Finally, we'll need to assume that there is no edge from U to Z,
- This is equivalent to saying that the instrument is independent of the confounder.



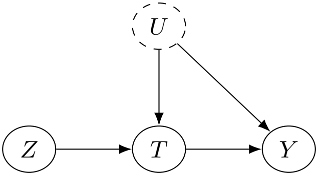
Assumption 3: (Conditional) Instrumental Unconfoundedness

- If there exists a W that couples Z and Y, we can still obtain a valid instrument by conditioning on W.



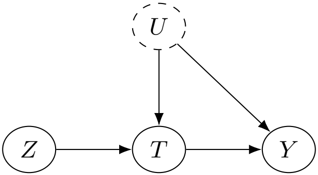
Instrumental variables - Intuition

- How to estimate ATE (or CATE) with unobserved U?
- *Intuition:*
  - Changes in the instrument Z lead to changes in the treatment T, and consequently the outcome Y,
  - If we modify Z, then T, Y will co-vary based on the relationship induced by U,
  - If we can modify Z in different ways, we can see how T, Y co-vary and subtract off the influence of U.



Intuition - Partial derivatives and differences in conditional expectations

- Note that $\frac{\partial y}{\partial z}$ represents the effect on the outcome by perturbation of the instrument,
- In the (implicit) SCM for the figure below, what we really want is to assess $\frac{\partial y}{\partial t}$,
- We have $\frac{\partial y}{\partial t} = \frac{\partial y}{\partial z}\frac{\partial z}{\partial t} = \frac{\frac{\partial y}{\partial z}}{\frac{\partial t}{\partial z}}$,
- In the binary setting, $\frac{\partial y}{\partial z}$ can be seen as $\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]$, and $\frac{\partial y}{\partial t}$ as $\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$.



Binary Linear Model

Assume $Y = \delta T + \alpha U + \epsilon$:

$$\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]$$
$$= \mathbb{E}[\delta T + \alpha U + \epsilon|Z=1] - \mathbb{E}[\delta T + \alpha U + \epsilon|Z=0]$$
$$= \mathbb{E}[\delta T + \alpha U|Z=1] - \mathbb{E}[\delta T + \alpha U|Z=0] + \cancel{\mathbb{E}[\epsilon|Z=1] - \mathbb{E}[\epsilon|Z=0]}$$

Binary Linear Model

Assume $Y = \delta T + \alpha U + \epsilon$:

$$
\begin{aligned}
&\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] \\
&= \mathbb{E}[\delta T + \alpha U + \epsilon|Z=1] - \mathbb{E}[\delta T + \alpha U + \epsilon|Z=0] \\
&= \mathbb{E}[\delta T + \alpha U|Z=1] - \mathbb{E}[\delta T + \alpha U|Z=0] + \cancel{\mathbb{E}[\epsilon|Z=1] - \mathbb{E}[\epsilon|Z=0]} \\
&= \delta(\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0]) + \alpha\underbrace{(\mathbb{E}[U|Z=1] - \mathbb{E}[U|Z=0])}_{U \perp\!\!\!\perp Z}
\end{aligned}
$$

Binary Linear Model

Assume $Y = \delta T + \alpha U + \epsilon$:

$$\begin{aligned}
&\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] \\
&= \mathbb{E}[\delta T + \alpha U + \epsilon|Z=1] - \mathbb{E}[\delta T + \alpha U + \epsilon|Z=0] \\
&= \mathbb{E}[\delta T + \alpha U|Z=1] - \mathbb{E}[\delta T + \alpha U|Z=0] + \cancel{\mathbb{E}[\epsilon|Z=1] - \mathbb{E}[\epsilon|Z=0]} \\
&= \delta(\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0]) + \alpha \underbrace{(\mathbb{E}[U|Z=1] - \mathbb{E}[U|Z=0])}_{U \perp\!\!\!\perp Z} \\
&= \delta(\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0]) + \alpha \cancel{(\mathbb{E}[U] - \mathbb{E}[U])} \\
&= \delta(\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0])
\end{aligned}$$

Binary Linear Model

Assume $Y = \delta T + \alpha U + \epsilon$:

$$
\begin{aligned}
&\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] \\
&= \mathbb{E}[\delta T + \alpha U + \epsilon|Z=1] - \mathbb{E}[\delta T + \alpha U + \epsilon|Z=0] \\
&= \mathbb{E}[\delta T + \alpha U|Z=1] - \mathbb{E}[\delta T + \alpha U|Z=0] + \cancel{\mathbb{E}[\epsilon|Z=1] - \mathbb{E}[\epsilon|Z=0]} \\
&= \delta(\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0]) + \alpha\underbrace{(\mathbb{E}[U|Z=1] - \mathbb{E}[U|Z=0])}_{U \perp\!\!\!\perp Z} \\
&= \delta(\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0]) + \alpha\cancel{(\mathbb{E}[U] - \mathbb{E}[U])} \\
&= \delta(\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0])
\end{aligned}
$$

Simplifying gives us the Wald Estimand:

$$\delta = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0]}$$

Binary Linear Model

Assume $Y = \delta T + \alpha U + \epsilon$:

$$
\begin{aligned}
&\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] \\
&= \mathbb{E}[\delta T + \alpha U + \epsilon|Z=1] - \mathbb{E}[\delta T + \alpha U + \epsilon|Z=0] \\
&= \mathbb{E}[\delta T + \alpha U|Z=1] - \mathbb{E}[\delta T + \alpha U|Z=0] + \cancel{\mathbb{E}[\epsilon|Z=1] - \mathbb{E}[\epsilon|Z=0]} \\
&= \delta(\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0]) + \alpha \underbrace{(\mathbb{E}[U|Z=1] - \mathbb{E}[U|Z=0])}_{U \perp\!\!\!\perp Z} \\
&= \delta(\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0]) + \alpha \cancel{(\mathbb{E}[U] - \mathbb{E}[U])} \\
&= \delta(\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0])
\end{aligned}
$$

Simplifying gives us the Wald Estimand:

$$\delta = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0]}$$

We can estimate this from data via the Wald Estimator:

$$\hat{\delta} = \frac{\frac{1}{n_1}\sum_{i:z_i=1} Y_i - \frac{1}{n_1}\sum_{i:z_i=0} Y_i}{\frac{1}{n_1}\sum_{i:z_i=1} T_i - \frac{1}{n_1}\sum_{i:z_i=0} T_i}$$

Continuous Linear Model

In the continuous case, we use a similar intuition but instead of differences in conditional expectations, we look at $\mathrm{Cov}(Y, Z)$.

$$
\begin{aligned}
&\mathrm{Cov}(Y, Z) \\
&= \mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z] \\
&= \mathbb{E}[(\delta T + \alpha U + \epsilon)Z] - \mathbb{E}[(\delta T + \alpha U + \epsilon)]\mathbb{E}[Z]
\end{aligned}
$$

Continuous Linear Model

In the continuous case, we use a similar intuition but instead of differences in conditional expectations, we look at $\mathrm{Cov}(Y, Z)$.

$$
\begin{aligned}
&\mathrm{Cov}(Y, Z) \\
&= \mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z] \\
&= \mathbb{E}[(\delta T + \alpha U + \epsilon)Z] - \mathbb{E}[(\delta T + \alpha U + \epsilon)]\mathbb{E}[Z] \\
&= \delta\mathbb{E}[TZ] + \alpha\mathbb{E}[UZ] - \delta\mathbb{E}[T]\mathbb{E}[Z] - \alpha\mathbb{E}[U]\mathbb{E}[Z]
\end{aligned}
$$

Continuous Linear Model

In the continuous case, we use a similar intuition but instead of differences in conditional expectations, we look at $\mathrm{Cov}(Y, Z)$.

$$\begin{aligned}
&\mathrm{Cov}(Y, Z) \\
&= \mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z] \\
&= \mathbb{E}[(\delta T + \alpha U + \epsilon)Z] - \mathbb{E}[(\delta T + \alpha U + \epsilon)]\mathbb{E}[Z] \\
&= \delta\mathbb{E}[TZ] + \alpha\mathbb{E}[UZ] - \delta\mathbb{E}[T]\mathbb{E}[Z] - \alpha\mathbb{E}[U]\mathbb{E}[Z] \\
&= \delta(\mathbb{E}[TZ] - \mathbb{E}[T]\mathbb{E}[Z]) + \alpha\underbrace{(E[UZ] - \mathbb{E}[U]\mathbb{E}[Z])}_{\mathrm{Cov}(U,Z)=0 \quad U \perp\!\!\!\perp Z} \\
&= \delta\mathrm{Cov}(T, Z)
\end{aligned}$$

Continuous Linear Model

In the continuous case, we use a similar intuition but instead of differences in conditional expectations, we look at $\text{Cov}(Y, Z)$.

$$\begin{aligned}
&\text{Cov}(Y, Z) \\
&= \mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z] \\
&= \mathbb{E}[(\delta T + \alpha U + \epsilon)Z] - \mathbb{E}[(\delta T + \alpha U + \epsilon)]\mathbb{E}[Z] \\
&= \delta\mathbb{E}[TZ] + \alpha\mathbb{E}[UZ] - \delta\mathbb{E}[T]\mathbb{E}[Z] - \alpha\mathbb{E}[U]\mathbb{E}[Z] \\
&= \delta(\mathbb{E}[TZ] - \mathbb{E}[T]\mathbb{E}[Z]) + \alpha\underbrace{(E[UZ] - \mathbb{E}[U]\mathbb{E}[Z])}_{\text{Cov}(U,Z)=0 \quad U \perp\!\!\!\perp Z} \\
&= \delta\text{Cov}(T, Z)
\end{aligned}$$

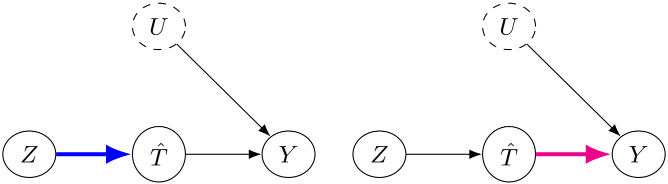
Simplifying gives us

$$\delta = \frac{\text{Cov}(Y, Z)}{\text{Cov}(T, Z)}$$

We can estimate this from data via the empirical covariances.

Another approach - Two-stage estimator

1. Estimate (via linear regression) $\mathbb{E}[T|Z]$. The model then gives us $\hat{T}$,
2. Estimate (via linear regression) $\mathbb{E}[Y|\hat{T}]$. The coefficient in front of $\hat{T}$ is our estimate $\hat{\delta}$.



For one-dimensional variables, this method matches the previous one:

$$\hat{T} = \frac{\text{Cov}(T, Z)}{\text{Var}(Z)} Z$$

$$\hat{\delta} = \frac{\text{Cov}(\hat{T}, Y)}{\text{Var}(\hat{T})} = \frac{\frac{\text{Cov}(T,Z)}{\text{Var}(Z)} \text{Cov}(Z, Y)}{\left(\frac{\text{Cov}(T,Z)}{\text{Var}(Z)}\right)^2 \text{Var}(Z)} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(T, Z)}$$

Questions?

Question

Any questions on IV estimators?

Heterogeneity in treatment effects

- Let's say we run the data science division of an app in use right now.
- We want to assess the causal effect of a push notification on purchases by the user.[1]
- Collect 10K users and randomly assign a push notification.
- But not everyone gets the notification! Furthermore, people do not behave in a homogenous manner.
- Older vs newer phones, people who turn off all notifications.

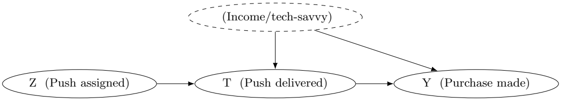


Figure: Causal graph of purchases in an app

[1] https://matheusfacure.github.io/python-causality-handbook/09-Non-Compliance-and-LATE.html

Effect of pushing notifications

- Push is randomly assigned so there is no bias.
- Let's start with
  $\text{ATE} = \mathbb{E}[Y|Z \text{ (push assigned)} = 1] - \mathbb{E}[Y|Z \text{ (push assigned)} = 0]$,
- Is this what we want?

Effect of pushing notifications

- Push is randomly assigned so there is no bias.
- Let's start with
  $\text{ATE} = \mathbb{E}[Y|Z \text{ (push assigned)} = 1] - \mathbb{E}[Y|Z \text{ (push assigned)} = 0]$,
- No, the above equation measures the effect of treatment assignment!

Effect of pushing notifications

- Push is randomly assigned so there is no bias.
- Let's start with
  ATE $= \mathbb{E}[Y|Z \text{ (push assigned)} = 1] - \mathbb{E}[Y|Z \text{ (push assigned)} = 0]$,
- Can we translate the above effect into the effect of treatment?

Effect of pushing notifications

- Push is randomly assigned so there is no bias.
- Let's start with
  ATE $= \mathbb{E}[Y|Z \text{ (push assigned)} = 1] - \mathbb{E}[Y|Z \text{ (push assigned)} = 0]$,
- Not quite – there is heterogeneity in how the population responds to treatment assignment.

Categorizations of treatment effect

We can split up the population into four groups based on how they respond to treatment assignment.

- Define $T_{Z_i=k}$ as the potential outcome of treatment T given the assignment $Z = k$.
- Compliers are those for whom $T_{Z_i=0} = 0$, $T_{Z_i=1} = 1$
- Defiers are those for whom $T_{Z_i=0} = 1$, $T_{Z_i=1} = 0$
- Always Takers are those for whom $T_{Z_i=0} = 1$, $T_{Z_i=1} = 1$
- Never Takers are those are those for whom $T_{Z_i=0} = 0$, $T_{Z_i=1} = 0$
- Can we estimate treatment effects when we have heterogeneity?

Categorizations of treatment effect

We can split up the population into four groups based on how they respond to treatment assignment.

- Define $T_{Z_i=k}$ as the potential outcome of treatment T given the assignment $Z = k$.
- Compliers are those for whom $T_{Z_i=0} = 0$, $T_{Z_i=1} = 1$
- Defiers are those for whom $T_{Z_i=0} = 1$, $T_{Z_i=1} = 0$
- Always Takers are those for whom $T_{Z_i=0} = 1$, $T_{Z_i=1} = 1$
- Never Takers are those are those for whom $T_{Z_i=0} = 0$, $T_{Z_i=1} = 0$
- Yes, with the monotonicity assumption $T_{Z_i=1} \geq T_{Z_i=0}$

Deriving treatment effects

Let's follow along the derivation of using Z as the instrument [1]

$$\begin{aligned}
&\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] \\
&\quad = \mathbb{E}[Y_{Z=1} - Y_{Z=0}|T_{Z_i=0}=0, T_{Z_i=1}=1]P(T_{Z_i=0}=0, T_{Z_i=1}=1) \\
&\quad + \mathbb{E}[Y_{Z=1} - Y_{Z=0}|T_{Z_i=0}=1, T_{Z_i=1}=0]P(T_{Z_i=0}=1, T_{Z_i=1}=0) \\
&\quad + \mathbb{E}[Y_{Z=1} - Y_{Z=0}|T_{Z_i=0}=1, T_{Z_i=1}=1]P(T_{Z_i=0}=1, T_{Z_i=1}=1) \\
&\quad + \mathbb{E}[Y_{Z=1} - Y_{Z=0}|T_{Z_i=0}=0, T_{Z_i=1}=0]P(T_{Z_i=0}=0, T_{Z_i=1}=0)
\end{aligned}$$

[1] Adapted from Brady Neal's course notes

Deriving treatment effects

Let's follow along the derivation of using Z as the instrument [1]

$$\begin{aligned}
&\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] \\
&\quad = \mathbb{E}[Y_{Z=1} - Y_{Z=0} | T_{Z_i=0} = 0, T_{Z_i=1} = 1] P(T_{Z_i=0} = 0, T_{Z_i=1} = 1) \\
&\quad + 0 \text{ (Monotonicity)} \\
&\quad + 0 \text{ (Invalidity of the instrument)} \\
&\quad + 0 \text{ (Invalidity of the instrument)}
\end{aligned}$$

[1] Adapted from Brady Neal's course notes

Deriving treatment effects

Let's follow along the derivation of using Z as the instrument [1]

$$\begin{aligned}
&\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0] \\
&\qquad\qquad \Longrightarrow \mathbb{E}[Y_{Z=1}-Y_{Z=0}|T_{Z_i=0}=0, T_{Z_i=1}=1] \\
&\qquad\qquad = \frac{\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]}{P(T_{Z_i=0}=0, T_{Z_i=1}=1)}
\end{aligned}$$

Simplifying the denominator as follows, we get:

$$\begin{aligned}
P(T_{Z_i=0}=0, T_{Z_i=1}=1) &= 1 - P(T=0|Z=1) - P(T=1|Z=0) \\
&= 1 - (1 - P(T=1|Z=1)) - P(T=1|Z=0) \\
&= P(T=1|Z=1) - P(T=1|Z=0) \\
&= \mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0]
\end{aligned}$$

[1] Adapted from Brady Neal's course notes

Local Average Treatment Effect

$$\mathbb{E}[Y_{Z=1} - Y_{Z=0} | T_{Z_i=0} = 0, T_{Z_i=1} = 1] = \underbrace{\frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0]}}_{\text{Wald Estimator}}$$

- When we have heterogeneity in the treatment effect, the instrumental variable only recovers the local average treatment effect.
- This is different from the Average Treatment Effect over the entire population!
- Required us to use monotonicity (which is not always satisfied).

Recap

1. Matching based estimators (propensity score, inverse propensity weighting)
2. Instrumental variables and identification of effects
3. What do IV estimators yield when we have heterogeneity?

Abadie, Alberto and Guido W Imbens (2011). “Bias-corrected matching estimators for average treatment effects”. In: *Journal of Business & Economic Statistics* 29.1, pp. 1–11.

Rosenbaum, Paul R and Donald B Rubin (1983). “The central role of the propensity score in observational studies for causal effects”. In: *Biometrika* 70.1, pp. 41–55.

King, Gary and Richard Nielsen (2019). “Why propensity scores should not be used for matching”. In: *Political Analysis* 27.4, pp. 435–454.