Causal Inference

Tuo Wang

2020-10-18

Contents

Part I

Introduction

This is the notes of Causal Inference. Most of the materials come from STAT 992 in UW-Madison instructed by Prof. Kang and from several textbooks in causal inference.

Recommended textbook:

- 1. Paul Rosenbaum. “Design of OBservational Studies” (2010)
- 2. Miguel Hernan & James Robins. “Causal Inference” (2019)
- 3. Guido Imbens and Don Rubin. “Causal Inference for Statistics, Social, and Biomedical Sciences” (2015)

Chapter 1

Propensity score

Chapter 2

Matching

This lecture note will cover the following topics:

- Exact Matching
- Propensity Score Matching
- Multivariate Caliper Matching
- Matching with Multiple Controls
- Full Matching

2.1 What is matching?

In observational study, absent random assignment, treated and control individuals may differ in terms of covariates, so direct comparison of the outcomes of treated individuals and controls may compare individuals who are not comparable - that is, a direct comparison may be biased as an estimate of the effect caused by the treatment.

Pros:

- Simple and intuitive
- Blinding to outcome info
- Diagnostic for overlap is easy

Cons:

- Theory is difficult
- It requires a lot of practice.

2.2 Exact Matching

Idea: Pair people with identical Xs

2.3 Propensity Score Matching

Quick Review on Propensity Score: Propensity score is the conditional probability of exposure to treatment given the observed covariates, $e(\mathbf{x}) = \mathbb{Pr}(A = 1|\mathbf{x})$, where $\mathbf{x}$ is the covaraites and A is the intervention variable. The propensity score is unkown, but it can be estimated from the data. For example, we can use the logistic regression:

$$\log\{\frac{e(\mathbf{x_i})}{1 - e(\mathbf{x_i})}\} = \beta_0 + \mathbf{x_i}^T$$

Idea: Pair people with similar $e(X)$. We know $X \perp A|e(X)$.

Calculate the propensity score for all subjects. Define the distance between subject i in the treatment group and subject j in the control group as:

$$d(\mathbf{x_i}, \mathbf{x_j}) = \left|\text{logit}(e(\mathbf{x_i})) - \text{logit}(e(\mathbf{x_i}))\right|$$

2.4 Multivariate Caliper Matching

Why we need multivariate caliper matching?

In lecture 6, we learned matching with propensity score, which tends to balance all of the covariates used to build that score, but two individuals with the same propensity score may differ in important ways. Also, propensity score matching is a single covariate matching, which ignore the interaction between different covarites.

2.4.1 (a) Mahalanobis Distances

Let $\mathbf{x}$ be the covariates random vector. Let $\hat{\Sigma}$ be the sample covariance matrix of $\mathbf{x}$. Then the estimated Mahalanobis distance between subject i and j, who has covarates $\mathbf{x_i}$ and $\mathbf{x_j}$ repectively, is defined as:

$$d(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} - \mathbf{x_j})^T \hat{\Sigma}^{-1} (\mathbf{x_i} - \mathbf{x_j})$$

See the following comments on Mahalanobis distance from Rosenbaum's book, Design of Observational Studies:

> "The Mahalanobis distance was originally developed for use with multivariate Nomal data, and for data of that type it works fine. With data that are not Normal, the Mahalanobis distance can exhibit some rather odd behavior. If one covariate contains extreme outliers of has a long-tailed distribution, its standard deviation will be inflated, and the Mahalanobis distance will tend to ignore that covariate in matching."

A simple alternative to the Mahalanobis distance is the 'rank-based Mahalanobis distance' which replaces each of the covariates by its ranks. Check out Design of Observational Studies, Chapter 8.3 for details.

2.4.2 (b) Penalize large distance

The idea is that two individuals can be close on the propensity score to a degree, once this degree is achieved, covariates $\mathbf{x}$ may affect the distance. Define w as the caliper width. With w, if two individuals have propensity scores that differ more than w, we will add a penalty to the ahalanobis distance between subject i and j. Explicitly,

$$d_{new}(\mathbf{x_i}, \mathbf{x_j}) = \begin{cases} d(\mathbf{x_i}, \mathbf{x_j}) + p \times \left|\text{logit}(e(\mathbf{x_i})) - \text{logit}(e(\mathbf{x_j}))\right|, & \text{if } \left|\text{logit}(e(\mathbf{x_i})) - \text{logit}(e(\mathbf{x_j}))\right| \geq w \\ d(\mathbf{x_i}, \mathbf{x_j}), & \text{if } \left|\text{logit}(e(\mathbf{x_i})) - \text{logit}(e(\mathbf{x_j}))\right| < w \end{cases}$$

where $e(\mathbf{x})$ is the propensity score of covariates $\mathbf{x}$, $d(\mathbf{x_i}, \mathbf{x_j})$ is the Mahalanobis distance between individual i and j and p is the penalty term.

In practical, we may want to care about the following things:

What's the value of p?

Usually $p = 1000$. In Paul Rosenbaum's 2019 review on matching: he used $d_{new}(\mathbf{x_i}, \mathbf{x_j}) = +\infty$ when $\left|\text{logit}(e(\mathbf{x_i})) - \text{logit}(e(\mathbf{x_j}))\right| \geq w$.

What's the value of w?

People use $w = 0.5 \times SD(\text{logit}(e(\mathbf{x})))$ or $w = 0.2 \times SD(\text{logit}(e(\mathbf{x})))$ in practice.

Comments: Use of the Mahalanobis distance inside propensity score calipers will balance covariates and also pair similar individuals. Also, using both Mahalanobis distance and propensity score calipers adds a protection against the failure of a single matching technique.

2.5 Matching with Multiple controls

In previous matching algorithms, we only match one treatment with one control. In matching with multiple controls, each treatment is matched to at least one control. For example if we match 1 treatment with 2 controls, assume we have 10 subjects in treatment group, we will end up with 10 matched sets, which each contains 1 treatment and 2 controls. Note that the 20 controls included in matched sets need to be different.

How many controls?

Let m be the number of controls to be matched for one treatment. In Paul Rosenbaum's 2019 review on matching:

> "Under a simple, familiar, conventional Gaussian model for matched sets, the variance of the estimator is proportional to 1+1/m, where the omitted constant of proportionality does not depend on m, but depends on all sorts of other things: the sample size, the variance of errors, and so on (see, for instance, Rosenbaum 2010, section 8.7)"

For example, $m = 1$ means paired matching, then $1 + \frac{1}{m} = 2$. If $m = +\infty$, then $1+\frac{1}{m} = 1$. Based on Rosenbaum's comment, $1+\frac{1}{m}$ represents the variability. By adding controls from 1 to $+\infty$, the sampling uncertainty from controls reduces but the uncertainty from treatment is unchanged. For example when $m = 10$, $1 + \frac{1}{m} = 1.1$ so almost all the uncertainty come from treatment group. In practice, $m = 3 \sim 5$.

Pros and cons:

- Pros: Efficiency. More samples are included inside matched sets.
- Cons: Balance becomes terrible.

2.6 Full Matching

In full matching, we can accept one treated matched with multiple controls and many treated matched with one control.

Assessing Balance from Full Matching: Before-matching standardized difference calculation is the same as before. For after matching, we use a weighted harmonic mean. Specifically, let $s = 1, \cdots, S$ denote the S matched sets. For each covariate X, the standarized difference after matching is computed as:

$$\Delta_X = \frac{\frac{\sum_{s=1}^{s} w_s\left(\bar{X}_{s,A=1} - \tilde{X}_{s,A=0}\right)}{\sum_{s=1}^{s} w_s}}{\text{SE from before matching}}, \quad w_s = \frac{1}{\frac{1}{\text{Number of treated in } s} + \frac{1}{\text{Number of Controls in } s}}$$

A matched pair has weight $w_s = 1$, so this generalization makes sense.

Chapter 3

Sensitivity Analysis I

This chapter will cover the following topics:

Sensitivity analysis is designed to answer what does an unmeasurable variable do in terms of impacting our causal condlusions about A and Y.

3.1 Sensitivity Analysis for Matched Sets with Binary Outcomes

Matched set id	subject	covariate
Matched set 1	Bob-Scott	$X_{Bob} \approx X_{Scott}$
Matched set 2	Jenn-Kathleen	$X_{Jenn} \approx X_{Kathleen}$
$\vdots$	$\vdots$	$\vdots$
Matched set I		

By ignorability assumption, $Y_i(1), Y_i(0), X_i \perp A_i$, either one of the pair is equally likely to get treatment. Formally:

$$p_1 = \text{Prob that Bob Smokes and Scott doesn't}$$
$$1 - p_1 = \text{Prob that Scott Smokes and Bob doesn't}$$

Under assumption: $p_1 = \frac{1}{2}$. More generally, $p_s = \frac{1}{2}, \forall s = 1, \cdots, I$. Suppose Bob is more like to smoke than Scott because he carries gene and p_1 could range between certain values, say:

$$\frac{1}{1+\Gamma} \leq p_1 \leq \frac{\Gamma}{1+\Gamma}, \ \Gamma \geq 1$$

For example, $\Gamma = 1 \implies p_1 = \frac{1}{2}$, $\Gamma = 3 \implies 0.25 \leq p_1 \leq 0.75$.

Γ is the sensitivity parameters. $\Gamma = 1 \implies$ random experiment given X. $\Gamma > 1 \implies$ non-random experiment of A due to uncertain variable U.

Rewrite we can get,

$$\frac{1}{\Gamma} \leq \frac{p_1}{1-p_1} \leq \Gamma,$$

where $\frac{p_1}{1-p_1}$ is the odds ratio of Bob smoking versus Scott smoking.

Conside $H_0 : Y_i(1) = Y_i(0)\ \forall i$

$$\hat{ATE} = \frac{1}{I}\sum_{s=1}^{I} \delta_s,$$

where δ_s is the difference in outcome between treated and control in matched set s. We can use $\hat{ATE}$ to test H_0. For simplicity, $T = \frac{1}{I}\sum_{s=1}^{I} \delta_s$. Under H_0:

$$\begin{aligned} Y_{Bob}(1) &= Y_{Bob}(0) \\ Y_{Scott}(1) &= Y_{Scott}(0) \end{aligned}$$

Thus,

$$\begin{aligned} \delta_1 &= \begin{cases} Y_{Bob}(1) - Y_{Scott}(0), & p_1 \\ Y_{Scott}(1) - Y_{Bob}(0), & 1-p_1 \end{cases} \\ &= \begin{cases} Y_{Bob}(0) - Y_{Scott}(0), & p_1 \\ Y_{Scott}(0) - Y_{Bob}(0), & 1-p_1 \end{cases} \end{aligned}$$

δ_1 only changes in sign denpending on who gets treated. Thus,

$$\delta_1 = Q_1 \times |Y_{Bob}(0) - Y_{Scott}(0)|,$$

where $Q_1 = \begin{cases} 1, & p_1 \\ -1, & 1-p_1 \end{cases}$

Thus under H_0,

$$T = \sum_{s=1}^{I} \delta_s = \sum_{s=1}^{I} Q_s C_s = \sum_{s:C_s \neq 0} Q_s,$$

where $Q_s = \begin{cases} 1, & p_1 \\ -1, & 1-p_1 \end{cases}$ and $C_s = 1$ or 0 in binary outcome

We can add a term transform the term in the summation. That is $Q_s = -1$ or 1, we can do a easy linear transformation to let term equal 0 and 1.

Let $d = \sum_{s=1}^{I} \mathbb{I}(C_s \neq 0)$. Then,

$$\frac{T+d}{2} = \sum_{s:C_s \neq 0} \frac{Q_s+1}{2} \sim Bin(d, \frac{1}{2})$$

$\frac{T+d}{2}$ follows a binomial distribuiton with $n = d$ and $p = \frac{1}{2}$.Thus the p-value = $\mathbb{P}(Bin(d, \frac{1}{2}) > \frac{T_{obs}+d}{2})$(Note that d is a constant once given observations).

Now suupose $p_1 \neq \frac{1}{2}$ e.g. $p_1 > \frac{1}{2}$ (Bob is more likely to smoke than scott). Instead we can assume p_1 is in a certain range i.e. $\frac{1}{1+\Gamma} \leq p_1 \leq \frac{\Gamma}{1+\Gamma}$.

What is the new p-value if $\Gamma > 1$? We may not be able to find the exact p-value but we can find the range of the p-value, i.e. we can find the upper and lower bound of the p-value.

$$\mathbb{P}(Bin(d, \frac{1}{1+\Gamma}) > \frac{T_{obs}+d}{2}) \leq \mathbb{P}(\frac{T+d}{2} > \frac{T_{obs}+d}{2}) \leq \mathbb{P}(Bin(d, \frac{\Gamma}{1+\Gamma}) > \frac{T_{obs}+d}{2})$$

How to choose the value of Γ:

- 1. Compute the range of p-values for each Γ, e.g $\Gamma = 1, 2, \cdots$
- 2. Stop if the range of p-value contains $\alpha = 0.05$

Chapter 4

causal Inference with Models

Throughout this chapter, we have the following set up :

- Y_i, A_i, X_i i.i.d $\sim F$

and assumptions:

- 1. $Y_i(1)A_i + Y_i(0)(1 - A_i) = Y_i$
- 2. $Y_i(1), Y_i(0) \perp A_i | X_i$
- 3. $0 < \mathbb{P}(A_i = 1|X_i) < 1$

Goal: Estimate $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$

We introduce four methods:

- IPW: Inverse probability weighting
- Outcome Regression /G-formula
- Double Robust estimation
- Machine learning

Under assumption 1-3:

$$\tau = \mathbb{E}\left[\frac{Y_i A_i}{e(X_i)} - \frac{Y_i(1 - A_i)}{1 - e(X_i)}\right] \quad \text{(inverse probability weighting)}$$

$$\tau = \mathbb{E}_X\{\mathbb{E}[Y_i|A_i = 1, X_i] - \mathbb{E}[Y_i|A_i = 0, X_i]\} = \mathbb{E}_X[f_1(X_i) - f_0(X_i)] \quad \text{(Outcome regression)}$$

Prove: $\mathbb{E}[Y_i(1)] = \mathbb{E}[\frac{Y_i A_i}{e(X_i)}]$ and $\mathbb{E}[Y_i(0)] = \mathbb{E}[\frac{Y_i(1-A_i)}{1-e(X_i)}]$

proof :

By assumption 3, $\mathbb{E}[\frac{Y_i A_i}{e(X_i)}]$ is well defined. Thus,

$$\begin{aligned}
\mathbb{E}[\frac{Y_i A_i}{e(X_i)}] &= \mathbb{E}_X\{\mathbb{E}[\frac{Y_i A_i}{e(X_i)}|X_i]\} \\
&= \mathbb{E}_X\{\frac{1}{e(X_i)}\mathbb{E}[Y_i A_i|X_i]\} \\
&= \mathbb{E}_X\{\frac{1}{e(X_i)}\mathbb{E}[(Y_i(1)A_i + Y_i(0)(1-A_i))A_i|X_i]\} \\
&= \mathbb{E}_X\{\frac{1}{e(X_i)}\mathbb{E}[Y_i(1)A_i|X_i]\} \\
&= \mathbb{E}_X\{\frac{1}{e(X_i)}\mathbb{E}[Y_i(1)|X_i]\mathbb{E}[A_i|X_i]\} \\
&= \mathbb{E}_X\{\mathbb{E}[Y_i(1)|X_i]\} \\
&= \mathbb{E}[Y_i(1)]
\end{aligned}$$

We can prove $\mathbb{E}[Y_i(0)] = \mathbb{E}[\frac{Y_i(1-A_i)}{1-e(X_i)}]$ by using the same procedure.

Prove: $\mathbb{E}[Y_i(1)] = \mathbb{E}_X\{\mathbb{E}[Y_i|A_i = 1, X_i]\}$ and $\mathbb{E}[Y_i(0)] = \mathbb{E}_X\{\mathbb{E}[Y_i|A_i = 0, X_i]\}$

proof:

$$\begin{aligned}
\mathbb{E}_X\{\mathbb{E}[Y_i|A_i = 1, X_i]\} &= \mathbb{E}_X\{\mathbb{E}[Y_i(1)A_i + Y_i(0)(1-A_i)|A_i = 1, X_i]\} \\
&= \mathbb{E}_X\{\mathbb{E}[Y_i(1)|A_i = 1, X_i]\} \\
&= \mathbb{E}_X\{\mathbb{E}[Y_i(1)|X_i]\} \\
&= \mathbb{E}[Y_i(1)]
\end{aligned}$$

We can prove $\mathbb{E}[Y_i(0)] = \mathbb{E}_X\{\mathbb{E}[Y_i|A_i = 0, X_i]\}$ by using the same procedure.

4.1 IPW estimator

Idea: replace $\mathbb{E}[.]$ with sample means.

Define:

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i A_i}{e(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i(1-A_i)}{1-e(X_i)}$$

To use $\hat{\tau}_{IPW}$, you need to know $e(X_i) = \mathbb{P}(A_i = 1|X_i)$.

From **Law of Large Number**:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i A_i}{e(X_i)} \to \mathbb{E}[\frac{Y_i A_i}{e(X_i)}]$$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i(1-A_i)}{1-e(X_i)} \to \mathbb{E}[\frac{Y_i(1-A_i)}{1-e(X_i)}]$$

Thus,

$$\hat{\tau}_{IPW} \to \tau$$

Rewrite,

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_iA_i}{e(X_i)} - \frac{Y_i(1-A_i)}{1-e(X_i)}\right]$$

Further, from **Central Limit Theorem:** we have:

$$\sqrt{n}(\hat{\tau}_{IPW} - \tau) \to N(0, \text{var}(\frac{Y_iA_i}{e(X_i)} - \frac{Y_i(1-A_i)}{1-e(X_i)}))$$

4.2 Outcome Regression

Define:

$$\hat{\tau}_{OR} = \frac{1}{n}\sum_{i=1}^{n}[f_1(X_i) - f_0(X_i)],$$

where $f_1(X_i) = \mathbb{E}[Y_i|A_i = 1, X_i]$ and $f_0(X_i) = \mathbb{E}[Y_i|A_i = 0, X_i]$.

To use $\hat{\tau}_{OR}$, you need to know $f_1(X_i)$ and $f_0(X_i)$.

From **Law of Large Number:**

$$\hat{\tau}_{OR} = \frac{1}{n}\sum_{i=1}^{n}[f_1(X_i) - f_0(X_i)] \to \mathbb{E}[f_1(X_i) - f_0(X_i)] = \tau$$

And further from **Central Limit Theorem:**

$$\sqrt{n}(\hat{\tau}_{OR} - \tau) \to N(0, \text{var}(f_1(X_i) - f_0(X_i)))$$

4.3 Doubly robust estimator

Define:

$$\hat{\tau}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(Y_i - f_1(X_i))A_i}{e(X_i)} + f_1(X_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(Y_i - f_0(X_i))(1-A_i)}{1-e(X_i)} + f_0(X_i)\right]$$

Suppose $f_1(X_i)$, $f_0(X_i)$, $e(X_i)$ are correct, we can show $\hat{\tau}_{DR} \to \tau$

proof:

$$\begin{aligned}\mathbb{E}\left[\frac{(Y_i - f_1(X_i))A_i}{e(X_i)}\right] &= \mathbb{E}_X\left\{\mathbb{E}\left[\frac{(Y_i - f_1(X_i))A_i}{e(X_i)}|X_i\right]\right\} \\ &= \mathbb{E}_X\left\{\frac{1}{e(X_i)}(\mathbb{E}[Y_iA_i|X_i] - \mathbb{E}[f_1(X_i)A_i|X_i])\right\} \\ &= \mathbb{E}_X\left\{\frac{1}{e(X_i)}(\mathbb{E}[Y_i|A_1 = 1, X_i]\mathbb{P}(A_i = 1|X_i) - f_1(X_i)\mathbb{E}[A_i|X_i])\right\} \\ &= 0\end{aligned}$$

The same holds for the other part. Then,

$$\hat{\tau}_{DR} \to \mathbb{E}[f_1(X_i)] - \mathbb{E}[f_0(X_i)] \to \tau$$

From **Central Limit Theorem:**

$$\sqrt{n}(\hat{\tau}_{DR} - \tau) \to N(0, \text{var}(\frac{(Y_i - f_1(X_i))A_i}{e(X_i)} + f_1(X_i) - \frac{(Y_i - f_0(X_i))(1 - A_i)}{1 - e(X_i)} - f_0(X_i)))$$

Suppose the estimate propensity score is different from the true propensity score, i.e. $\hat{e}(X) \neq e(X)$. Assume f_1 and f_0 are known.

$$\hat{\tau}_{DR,\hat{e}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(Y_i - f_1(X_i))A_i}{\hat{e}(X_i)} + f_1(X_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(Y_i - f_0(X_i))(1 - A_i)}{1 - \hat{e}(X_i)} + f_0(X_i)\right]$$

We still have

$$\hat{\tau}_{DR,\hat{e}} \to \tau$$

Suppose $\hat{f}_1 \neq f_1$ and $\hat{f}_0 \neq f_0$ but e is correct.

$$\begin{aligned}\hat{\tau}_{DR,\hat{f}_1,\hat{f}_0} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(Y_i - \hat{f}_1(X_i))A_i}{e(X_i)} + \hat{f}_1(X_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(Y_i - \hat{f}_0(X_i))(1 - A_i)}{1 - e(X_i)} + \hat{f}_0(X_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_iA_i}{e(X_i)} - \frac{Y_i(1 - A_i)}{1 - e(X_i)}\right) + \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_1(X_i) - \frac{\hat{f}_1(X_i)A_i}{e(X_i)} - \hat{f}_0(X_i) - \frac{\hat{f}_0(X_i)A_i}{1 - e(X_i)}\right) \\ &= \hat{\tau}_{IPW} + \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_1(X_i) - \frac{\hat{f}_1(X_i)A_i}{e(X_i)} - \hat{f}_0(X_i) - \frac{\hat{f}_0(X_i)A_i}{1 - e(X_i)}\right)\end{aligned}$$

We know the first part $\hat{\tau}_{IPW} \to \tau$ and it is easy to show the latter part goes to zero. Thus,

$$\hat{\tau}_{DR,\hat{f}_1,\hat{f}_0} \to \tau$$

Suppose replace f_1,f_0,e with estimate function $\hat{f}_1$,$\hat{f}_0$,$\hat{e}$. If $\hat{f}_1$,$\hat{f}_0$,$\hat{e}$ are all l2 converge to f_1,f_0,e, then

$$\hat{\tau}_{DR,\hat{e},\hat{f}_1,\hat{f}_0} \to \tau$$

4.4 Asymptotic variance

Asymptotic variance of $\hat{\tau}_{IPW}$:

$$\begin{aligned}
\mathrm{Var}\left[\frac{Y_i A_i}{e(X_i)} - \frac{Y_i(1-A_i)}{1-e(X_i)}\right] &= \mathbb{E}\left[\frac{Y_i A_i}{e(X_i)} - \frac{Y_i(1-A_i)}{1-e(X_i)}\right]^2 - \left(\mathbb{E}\left[\frac{Y_i A_i}{e(X_i)} - \frac{Y_i(1-A_i)}{1-e(X_i)}\right]\right)^2 \\
&= \mathbb{E}\left[\frac{Y_i A_i}{e(X_i)} - \frac{Y_i(1-A_i)}{1-e(X_i)}\right]^2 - \tau^2 \\
&= \mathbb{E}\left[\frac{Y_i^2 A_i^2}{e(X_i)^2} - 2\frac{Y_i A_i Y_i(1-A_i)}{e(X_i)(1-e(X_i))} + \frac{Y_i^2(1-A_i)^2}{[1-e(X_i)]^2}\right] - \tau^2 \\
&= \mathbb{E}\left[\frac{Y_i^2 A_i}{e(X_i)^2} + \frac{Y_i^2(1-A_i)}{[1-e(X_i)]^2}\right] - \tau^2 \\
&= \mathbb{E}\Bigg[\frac{(Y_i(1)A_i + Y_i(0)(1-A_i))^2 A_i}{e(X_i)^2} \\
&\quad + \frac{(Y_i(1)A_i + Y_i(0)(1-A_i))^2(1-A_i)}{[1-e(X_i)]^2}\Bigg] - \tau^2 \\
&= \mathbb{E}\left[\frac{Y_i(1)^2 A_i}{e(X_i)^2} + \frac{Y_i(0)^2(1-A_i)}{[1-e(X_i)]^2}\right] - \tau^2 \\
&= \mathbb{E}\left\{\mathbb{E}\left[\frac{Y_i(1)^2 A_i}{e(X_i)^2} + \frac{Y_i(0)^2(1-A_i)}{[1-e(X_i)]^2}|X_i\right]\right\} - \tau^2 \\
&= \mathbb{E}\left[\frac{Y_i(1)^2}{e(X_i)} + \frac{Y_i(0)^2}{1-e(X_i)}\right] - \tau^2
\end{aligned}$$

Asymptotic variance of $\hat{\tau}_{OR}$:

$$\begin{aligned}
\mathrm{Var}(f_1(X_i) - f_0(X_i)) &= \mathbb{E}[f_1(X_i) - f_0(X_i)]^2 - [\mathbb{E}(f_1(X_i) - f_0(X_i))]^2 \\
&= \mathbb{E}[f_1(X_i) - f_0(X_i)]^2 - \tau^2
\end{aligned}$$

Asymptotic variance of $\hat{\tau}_{DR}$:

$$
\begin{aligned}
&\mathrm{Var}\left[\frac{(Y_i - f_1(X_i))A_i}{e(X_i)} + f_1(X_i) - \frac{(Y_i - f_0(X_i))(1-A_i)}{1-e(X_i)} - f_0(X_i)\right] \\
&= \mathbb{E}\left[\frac{(Y_i - f_1(X_i))A_i}{e(X_i)} + f_1(X_i) - \frac{(Y_i - f_0(X_i))(1-A_i)}{1-e(X_i)} - f_0(X_i) - \tau\right]^2 \\
&= \mathbb{E}\Bigg\{\left[\frac{(Y_i - f_1(X_i))A_i}{e(X_i)} - \frac{(Y_i - f_0(X_i))(1-A_i)}{1-e(X_i)}\right]^2 \\
&\quad + 2\left[\frac{(Y_i - f_1(X_i))A_i}{e(X_i)} - \frac{(Y_i - f_0(X_i))(1-A_i)}{1-e(X_i)}\right]\left[f_1(X_i) - f_0(X_i) - tau\right] \\
&\quad + \left[f_1(X_i) - f_0(X_i) - tau\right]^2\Bigg\} \\
&= \mathbb{E}\Bigg\{\left[\frac{(Y_i - f_1(X_i))A_i}{e(X_i)} - \frac{(Y_i - f_0(X_i))(1-A_i)}{1-e(X_i)}\right]^2 \\
&\quad + \left[f_1(X_i) - f_0(X_i) - tau\right]^2\Bigg\} \\
&= \mathbb{E}\left[\frac{\left(Y_i(1) - \mathbb{E}\left[Y_i(1)|X_i\right]\right)^2}{e\left(X_i\right)} + \frac{\left(Y_i(0) - \mathbb{E}\left[Y_i(0)|X_i\right]\right)^2}{1-e\left(X_i\right)} + \left(f_1\left(X_i\right) - f_0\left(X_i\right) - \tau\right)^2\right]
\end{aligned}
$$

Chapter 5

Sensitivity Analysis II

Reference:

- Peng and Tyler (2014), "Generalized Cornfield conditions for the risk difference"
- Peng and Tyler (2016), "Sensitivity Analysis Without Assumptions"
- Tyler and Peng (2017), "Sensitivity Analysis in Observational Research: Introducing the E-Value"

This chapter will cover the following topics:

- E-values
- Sensitivity analysis for IPW, OR and DR estimators.

A central question in causal inference with observational studies is the sensitivity of conclusions to unmeasured confounding.

First we introduce three variables to describe causal effect:

- Risk Difference =

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{P}(Y_i(1) = 1) - \mathbb{P}(Y_i(0) = 1)$$

- Relative Risk (RR) =

$$\frac{\mathbb{P}(Y_i(1) = 1)}{\mathbb{P}(Y_i(0) = 1)}$$

  e.g. RR=3 means treatment causal increases the probability of $Y = 1$ by a factor of 3
- Odds Ratio (OR) =

$$\frac{\mathbb{P}(Y_i(1) = 1)}{1 - \mathbb{P}(Y_i(1) = 1)} \Big/ \frac{\mathbb{P}(Y_i(0) = 1)}{1 - \mathbb{P}(Y_i(0) = 1)}$$

  e.g. OR=3 means treatment causal increases the odss of $Y = 1$ by 3 fold.

Goal: we want to assess what happens to estimators of tau and RR when an observed confounder U is present

Review the assumptions:

- SUTUA: $Y_i = A_i Y_i(1) + (1 - A_i)Y_i(0)$
- Conditional ignorability: $Y_i(1), Y_i(0) \perp A_i | X_i$
- $0 << P(A_i = 1 | X_i = x) << 1 \ \forall x$

When U is present, the assumptions becomes:

- Conditional ignorability: $Y_i(1), Y_i(0) \perp A_i | X_i, U_i$
- $0 << P(A_i = 1 | X_i = x, U_i = u) << 1 \ \forall x, u$

Graphically:

Let RR_{AY} be the relative risk of treatment on outcome. Let RR_{UA} be the relative risk of confounder on treatment. Let RR_{UY} be the relative risk of confounder on outcome. In the presence of U, our estimators of tau and RR are biased. We want to measure how much is this bias.

Define:

$$B = \frac{RR_{AU} RR_{UY}}{RR_{AU} + RR_{UY} - 1}$$

Theorem:

$$RR_{AY} \geq \frac{\hat{RR}}{B}$$

Note that:

- This bound is sharp.
- U can be any type (one dimension and multi-dimension).
- This formula can also be applied on the lower bound of the confidence interval.

Example: Study the causal effect of baby formula (A=1) on respiratory death (Y=1)

- $\hat{RR} = 3.9$, 95% CI:(1.8,8.7)
- They controlled for age, ... (Xs)

What if there is an U? How would this U change the result?

$\exists U$ s.t. $RR_{UA} = 2$, probability of taking infants formula increase 2 and $RR_{UY} = 4$ If $U = 1$, then baby is 4 times more likely to die than if $U = 0$. Then

$$RR_{AY} \geq \frac{3.9}{1.6} = 2.43$$

and

$$L_{AY} \geq \frac{\hat{L}_{AY}}{B} = \frac{1.8}{1.6} = 1.125$$

5.1 E-value

Suppose $RR_{UA} = RR_{UY}$, then $B = \frac{RR_{UA}^2}{2RR_{UA}-1}$. Then consider the point in the lower bound where $RR_{AY} = 1$,

$$1 = RR_{AY} \geq \frac{\hat{RR}}{B} = \frac{\hat{RR}}{\frac{RR_{AU}^2}{2RR_{AU}-1}}$$

Solve for RR_{UA}, we get the

$$\text{E-value} = \hat{RR} + \sqrt{\hat{RR}(\hat{RR}-1)}$$

Note that:

- Higher E-value means the study is more robust to unmeasured U.
- E-calue is the minimum strength of assoication on the risk ratio scale, that an unmeasured confounder would need to have with both the treatment and outcome, conditional on the measured covariates, to explain away a treatment-outcome association.
- E-value is a continuous measure of an association's robustness to potential uncontrolled condounders. The lowerst possible E-value is 1 (that is, no unmeasured confounding is needed to explain away the observed association). The higher the E-value, the stronger the confounder associations must be to explain away the effect.i

Risk difference τ: Let $\hat{\tau}$ be the estimator, Y binary, $p_1 = \mathbb{P}(Y_i = 1|A_i = 1)$, $p_0 = \mathbb{P}(Y_i = 1|A_i = 0)$ and $f = P(A_i = 1)$.

Theorem:

$$\tau_{AY} \geq (p_1 - Bp_0)\sqrt{f + \frac{1-f}{B}}$$

Chapter 6

Instrumental Variable

6.1 IV without covariates X

Goal: Estimate the causal effect of A on Y, especially when:

- A is "hopelessly" confounded
- A can never be randomized

Idea: Use "external" random variation in the data to "extract" randomness out of confounded A

Controversy: It's "impossible" to find such Z.

Set up:

- $Z \in 0, 1$, Instrumental Variable
- $A \in 0, 1$, Treatment
- $Y \in \mathbb{R}$, Outcome

Potential Outcome:

- $A_i^{(Z=1)}$: potential treatment when $Z = 1$
- $A_i^{(Z=0)}$: potential treatment when $Z = 0$
- $Y_i^{(Z=1,A=1)}$,$Y_i^{(Z=0,A=1)}$,$Y_i^{(Z=1,A=0)}$,$Y_i^{(Z=0,A=0)}$
- $Y_i^{(z,a)}$ is the potential outcome if unit i received z and a

Assumptions:

- *A1*: SUTVA

$$A_i = Z_i A_i(1) + (1 - Z_i)A_i(0)$$

$$Y_i = Z_i Y_i^{(1,A_i(1))} + (1 - Z_i)Y_i^{(0,A_i(0))}$$

- *A2*: $Z \perp Y^{(z,a)}, A^{(z)}$, I.V. is randomized
- *A3*: $0 < \mathbb{P}(Z_i = 1) < 1$

- *A4*: $\mathbb{E}[A(1) - A(0)] \neq 0$, I.V.(Z) has casual effect on A
- *A5*: $Y_i^{(Z=1,a)} = Y_i^{(Z=0,a)} = Y_i^{(a)}$
- *A6*: $A^{(1)} \geq A^{(0)}$, monotoncity

Notes on the assumptions:

- If Z was completely randomized like in a RCT, A1-A3 are plausible.
- A4 can be checked from data if Z was randomized
- A5 can never be fully verified with data!! We can observe $Y_i^{(1,a)}$,$Y_i^{(0,a)}$ at the same time.
- A6 means one-sided complicance exists in the experiment design. One-sided complicance: $A_i(0) = 0$, i.e $A_i(0) = 1$ will never happen.

Accoding to this paper, Angrist, Imbens and Rubin (1996) "Identification of Causal Effects Using Instrumental Variables", we can divide the subjects into four categories:

- *G1*: $A_i(1) = 1, A_i(0) = 1$: Always takers (doesn't exist for one-sided compliance)
- *G2*: $A_i(1) = 0, A_i(0) = 0$: Never takers
- *G3*: $A_i(1) = 1, A_i(0) = 0$: Compliers
- *G4*: $A_i(1) = 0, A_i(0) = 1$: Defiers

G1-4 partitions the population in a non-overlapping manner. If monotonicity holds, then defiers cannot exists.

Theorem: If A1-A6 held, then

$$\begin{aligned}\mathbb{E}[Y_i(1) - Y_i(0)|\, i \text{ is complier }] &= \mathbb{E}[Y_i(1) - Y_i(0)|A_i(1) = 1, A_i(0) = 1] \\ &= \frac{\mathbb{E}(Y_i|Z_i = 1) - \mathbb{E}(Y_i|Z_i = 0)}{\mathbb{E}(A_i|Z_i = 1) - \mathbb{E}(A_i|Z_i = 0)}\end{aligned}$$

Note that the estimand identifies the ATE, but among a group of people (compliers) and we don't know who they are!

proof:

First, let us look at the numerator.

$$
\begin{aligned}
& E(Y_i|Z_i=1) - E(Y_i|Z_i=0) \\
=^{A_1}\ & E(Z_i Y_i^{(1,A_i(1))} + (1-Z_i)Y_i^{(0,A_i(0))}|Z_i=1) - \\
& E(Z_i Y_i^{(1,A_i(1))} + (1-Z_i)Y_i^{(0,A_i(0))}|Z_i=0) \\
=\ & E(Y_i^{(1,A_i(1))}|Z_i=1) - E(Y_i^{(0,A_i(0))}|Z_i=0) \\
=^{A_2}\ & E(Y_i^{(1,A_i(1))}) - E(Y_i^{(0,A_i(0))}) \\
=\ & E[(Y_i^{(1,A_i(1))}) - E(Y_i^{(0,A_i(0))})I(A_i(1)=1, A_i(0)=0) + \\
& (Y_i^{(1,A_i(1))}) - E(Y_i^{(0,A_i(0))})I(A_i(1)=1, A_i(0)=1) + \\
& (Y_i^{(1,A_i(1))}) - E(Y_i^{(0,A_i(0))})I(A_i(1)=0, A_i(0)=0)] \\
=\ & E[(Y_i^{(1,A_i(1))}) - E(Y_i^{(0,A_i(0))})I(A_i(1)=1, A_i(0)=0)] \\
=\ & E[(Y_i^{(1)}) - E(Y_i^{(0)})I(A_i(1)=1, A_i(0)=0)]
\end{aligned}
$$

Then, we look at the denominator.

$$
\begin{aligned}
& E[A_i|Z_i=1] - E[A_i|Z_i=0] \\
=\ & E[Z_i A_i(1) + (1-Z_i)A_i(0)|Z_i=1] - \\
& E[Z_i A_i(1) + (1-Z_i)A_i(0)|Z_i=0] \\
=\ & E[A_i(1)|Z_i=1] - E[A_i(0)|Z_i=0] \\
=\ & E[A_i(1)] - E[A_i(0)] \\
=\ & E[A_i(1) - A_i(0)]
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& E[Y_i(1) - Y_i(0)|A_i(1)=1, A_i(0)=1] \\
=\ & \frac{E[(Y_i(1) - Y_i(0))I(A_i(1)=1, A_i(0)=1)]}{E[I(A_i(1)=1, A_i(0)=1)]} \\
=\ & \frac{E[(Y_i(1) - Y_i(0))I(A_i(1)=1, A_i(0)=1)]}{E[A_i(1) - A_i(0)]}
\end{aligned}
$$

How to estmate?

First method: We can use the sample mean to estimate the terms in the theorem, i.e.

$$\hat{E}(Y_i|Z_i=1) = \frac{\sum_{i=1}^n Y_i Z_i}{\sum_{i=1}^n Z_i}$$

Theorem: $\hat{E}(Y_i|Z_i=1) \xrightarrow{p} E(Y_i|Z_i=1)$

proof:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i Z_i \xrightarrow{p} E[Y_i Z_i]$$

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{p} E[Z_i]$$

Thus, by using Slutsky's theorem:

$$\frac{\sum_{i=1}^{n} Y_i Z_i}{\sum_{i=1}^{n} Z_i} \xrightarrow{p} \frac{E[Y_i Z_i]}{E[Z_i]} = E[Y_i | Z_i = 1]$$

Theorem:

$$\hat{\tau}_w = \frac{\hat{E}_{YZ} - \hat{E}_{Y(1-Z)}}{\hat{E}_{AZ} - \hat{E}_{A(1-Z)}} \xrightarrow{p} \tau = \frac{E_{YZ} - E_{Y(1-Z)}}{E_{AZ} - E_{A(1-Z)}}$$

and

$$\sqrt{\hat{\tau}_w - \tau} \to N(0, E\left[\frac{\frac{Y_i(1)^2}{p} + \frac{Y_i(0)^2}{1-p}}{\tau_A^2}\right] - \tau^2),$$

where $p = \mathbb{P}(Z_i = 1)$ and $\tau_A = E[A_i | Z_i = 1] - E[A_i | Z_i = 0]$

Second method: Two Stage Least Square

Theorem: If Z is binary, then $\hat{\tau}_{TSLS} = \hat{\tau}_w$

6.2 IV with covariates X

Assumptions:

- *A1*: SUTVA

$$A_i = Z_i A_i(1) + (1 - Z_i) A_i(0)$$
$$Y_i = Z_i Y_i^{(1, A_i(1))} + (1 - Z_i) Y_i^{(0, A_i(0))}$$

- *A2*: $Z_i \perp A_i(1), A_i(0), Y_i^{(1, A_i(1))}, Y_i^{(0, A_i(0))} | X_i$
- *A3*: $0 < \mathbb{P}(Z_i = 1 | X_i = x) < 1 \ \forall x$
- *A4*: $\mathbb{E}[A(1) - A(0) | X_i = x] \neq 0 \ \forall x$
- *A5*: $Y_i^{(Z=1,a)} = Y_i^{(Z=0,a)} = Y_i^{(a)}$
- *A6*: $A^{(1)} \geq A^{(0)}$, monotoncity

Theorem: Under assumptions *A1-A6*, The ATE among compliers:

$$E[Y_i(1) - Y_i(0) | A_i(1) = 1, A_i(0) = 1] = \frac{E_X\{E[Y_i | Z_i = 1, X_i] - E[Y_i | Z_i = 0, X_i]\}}{E_X\{E[A_i | Z_i = 1, X_i] - E[A_i | Z_i = 0, X_i]\}}$$

Strategy of estimating with Xs:

1:Two stage least square

Property of $\hat{\tau}_{TSLS}$

- $\hat{\tau}_{TSLS} \to \tau$
- $\sqrt{n}(\hat{\tau}_{TSLS} - \tau) \to N(0, V_{TSLS})$

2:Matching (non-parametric)

Take the ratio of the coef estimate between $Y \sim Z + .$ and $A \sim Z + .$

Property of $\hat{\tau}_{match}$

- $\hat{\tau}_{match} \to \tau$
- $\sqrt{n}(\hat{\tau}_{match} - \tau) \to N(0, V_{match})$
- $V_{match} \geq V_{TSLS}$

3:IPW OR DR

6.3 New topics in I.V.

- Multiple IVs
- Privacy and IV

Chapter 7

notes on M-estimator

References:

- 1. Stefanski and Boos's paper on M-estimation.
- 2. Chapter 5 of Asymptotic Statistics by van der Vaart.
- 3. Pages 29 to 32 in Semiparametric Theory and Missing Data by Tsiatis.