Unconfoundedness

Huan Deng

April 29, 2025

Outline

Introduction

- Definitions
- Simpson's Paradox

Unconfoundedness: Identification

- Identification
- Propensity Score

Unconfoundedness: Estimation

- Stratification
- Multivariate Regression
- Using Propensity Score: Stratification and IPW
- Doubly Robust Methods

Heterogeneous Treatment Effect

- T-learner
- A Semi-parametric Model

Introduction

Introduction

Definitions

Overview

Today, we will go beyond random assignment and extend unconditional independence to conditional independence.

1. Random assignment: $\{Y(0), Y(1)\} \perp D$
2. Unconfoundedness: $\{Y(0), Y(1)\} \perp D \mid X$
3. We will also see names such as Conditional Independence Assumption (CIA), Ignorability, and Selection on Observables
4. Identification-wise, there is not a lot to discuss.
5. But there are many estimators to use and recently, estimating heterogeneous treatment effects is a hot topic.

Overview

1. Unconfoundedness might sound a strong identification assumption, especially for pure observational studies
2. But for experiments, or situations where we can observe how a policymaker makes decisions, unconfoundedness might be a reasonable assumption.
3. More generally, conditioning per se might not be enough to clear all the confounding factors, but it can be complementary to other research designs.
4. For example, conditional parallel trend in difference-in-differences.
5. This lecture draws materials from Imbens and Rubin (2015), Wager's lecture notes (2024), Angrist and Pischeke (2009), and Mackenzie and Pearl (2018).

Introduction

Simpson’s Paradox

Simpson's Paradox

- Simpson's paradox might be the most striking example to show how misleading our conclusions can be when we mistakenly assume random assignment
- Dr. Simpson did RCTs for young and old people respectively to assess the efficacy of a drug.
- Dr. Simpson first looks at the data for the young and old respectively, then he pools the data together to check the results for the whole population.

Table: Fictitious data from Mackenzie and Pearl (2018)

	Control Group (No Drug)		**Treatment Group** (Took Drug)	
	Heart attack	*No heart attack*	*Heart attack*	*No heart attack*
Young	1	19	3	37
Old	12	28	8	12
Total	13	47	11	49

Simpson's Paradox

- We compute the difference-in-means estimates for the young, old, and the whole population. We get the below results:

$$\tau_y = \frac{3}{3+37} - \frac{1}{1+19} = 0.025 \tag{1}$$

$$\tau_o = \frac{8}{8+12} - \frac{12}{12+28} = 0.1 \tag{2}$$

$$\tau = \frac{11}{11+49} - \frac{13}{13+47} = -0.033 \tag{3}$$

- So, the drug increases the risk of heart attack for the young, for the old, but not for "people"?
- This can't be true. Since we know that for subpopulation, we correctly estimate the treatment effect, then the only explanation is: the difference-in-means estimate fails to identify the ATE for the whole population

Unconfoundedness: Identification

Unconfoundedness: Identification

Identification

Overview

- Under Unconfoundedness, we introduce another treatment effect, the Conditional Average Treatment Effect (CATE):

$$CATE = E[Y_i(1) - Y_i(0)|X = x] \tag{4}$$

- Under Unconfoundedness, CATE can be rewritten as:

$$\begin{aligned} CATE &= E[Y_i(1) - Y_i(0)|X_i = x] \\ &= E[Y_i(1)|X_i = x, D_i = 1] - E[Y_i(0)|X_i = x, D_i = 0] \\ &= E[Y_i|X_i = x, D_i = 1] - E[Y_i|X_i = x, D_i = 0] \end{aligned} \tag{5}$$

- The last two terms in the above equation can be estimated from data as long as there is overlapping support:

$$P(D_i = 1|X_i = x) \in (0, 1), \forall x \in X \tag{6}$$

- Then we can identify CATE

Overview

- With CATE identified, ATE can be identified as well, since ATE is a weighted average of CATE:

$$
\begin{aligned}
ATE &= \sum E[Y_i(1) - Y_i(0)|X_i = x] \times p(X_i = x) \\
&= \sum CATE(x) \times p(X_i = x) \qquad (7)
\end{aligned}
$$

- $p(X_i = x)$ is directly identified from the data
- Unconfoundedness is not testable because we can only observe one of $Y(0)$ and $Y(1)$
- But the overlapping support condition can be tested since it only involves observables.

Overview

- Now we revisit the Simpson's paradox rigorously by writing out the expression for the difference-in-means estimate:

$$
\begin{aligned}
E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = &\sum E[Y_i|D_i = 1, X_i = x] \times p(X_i = x|D_i = 1) - \\
&\sum E[Y_i|D_i = 0, X_i = x] \times p(X_i = x|D_i = 0) \\
= &\sum CATE(x) \times p(X_i = x|D_i = 1) + \\
&\sum E[Y_i|D_i = 0, X_i = x][p(X_i = x|D_i = 1) - p(X_i = x|D_i = 0)]
\end{aligned}
\tag{8}
$$

- $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$ is not a convex combination of CATEs
- Even if $CATE(x) = \tau, \forall x$, we don't necessarily have $E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \tau$
- So when random assignment doesn't hold but we mistakenly implement the difference-in-means estimator, we won't get a good estimate for our estimand.

Unconfoundedness: Identification

Propensity Score

Overview

- We naturally come up with an estimation strategy where we first estimate CATE and then aggregate all the CATEs into ATE
- This is good if the cardinality if X is a small number
- But what if X is high-dimensional or continuous?
- We need to find a way to reduce the dimension of our estimation problem.

Dimension Reduction

- To show that we can reduce the dimension of our problem, we need to first introduce an important concept: the **propensity score**:

$$e(x) = p[D_i = 1|X_i = x] \tag{9}$$

- Rosenbaum and Rubin (1983):

$$\{Y(0), Y(1)\} \perp D \mid X = \{Y(0), Y(1)\} \perp D \mid e(X) \tag{10}$$

- The propensity serves as a “sufficient statistics”: knowing the propensity score is enough to guarantee the validity of stratified estimators under unconfoundedness.

Unconfoundedness: Estimation

Overview

- Now we discuss several major estimators that exploit the Unconfoundedness assumption
- To estimate CATE, we need to either estimate the CEF function or the propensity score
- Different estimation methods can be understood as different ways to use CEF or propensity score

Unconfoundedness: Estimation

Stratification

Stratification

- As we discussed before, when the cardinality of X is small, then the most natural estimator is the stratified estimator:

$$\hat{\tau}_{STRAT} = \sum_{x \in \mathcal{X}} \frac{n_x}{n} \hat{\tau}(x), \tag{11}$$

$$\hat{\tau}(x) = \frac{1}{n_{x1}} \sum_{\{X_i = x, D_i = 1\}} Y_i - \frac{1}{n_{x0}} \sum_{\{X_i = x, D_i = 0\}} Y_i \tag{12}$$

  where $n_x = |\{i : X_i = x\}|$ and $n_{xd} = |\{i : X_i = x, D_i = d\}|$
- The stratified estimate can be shown to be consistent and asymptotically normal

Bias-Variance Trade-off

- In practice, we need to decide on how to stratify X.
- The finer our stratification, the more likely we are to satisfy the unconfoundedness assumption; thus, we will have a lower bias.
- However, then it is possible that each strata might only have a small number of treated or control units; our estimation will suffer from a high variance, i.e., we have a noisier estimation.

Unconfoundedness: Estimation

Multivariate Regression

Multivariate Regression

- We might have seen tons of papers doing multivariate regression that "controls for" X
- Multivariate regression also exploits the unconfoundedness assumption.
- Regression is just another method of aggregating CATE
- More generally, we can treat regression as an aggregation method that uses some weighting scheme to add up "building block" estimands into higher-level estimands
- This perspective will be particularly useful when we discuss the difference-in-differences method

Multivariate Regression

- Consider the below regression model:

$$Y_i = \alpha D_i + \beta X_i + \epsilon_i, \tag{13}$$

where X_i is a vector of dummy variables that saturate the regression model, which means the CEF function of Y on X is linear.

- By the FWL theorem, we can show that:

$$\begin{aligned}\hat{\alpha} &\xrightarrow{p} \frac{E[(D_i - E[D_i|X_i])^2 \delta_X]}{E[(D_i - E[D_i|X_i])^2]} \\ &= \frac{E\{E[(D_i - E[D_i|X_i])^2|X_i]\delta_X\}}{E\{E[(D_i - E[D_i|X_i])^2|X_i]\}} \\ &= \frac{E[\sigma_D^2(X_i)\delta_X]}{E[\sigma_D^2(X_i)]},\end{aligned} \tag{14}$$

where $\delta_X = CATE(X)$, $\sigma_D^2(X_i) \equiv E[(D_i - E[D_i|X_i])^2|X_i]$. Check Angirst and Pischeke (2009) P74-75 for detailed proof.

Unconfoundedness: Estimation

Using Propensity Score: Stratification and IPW

Stratification Based on Propensity Score

- Thanks to the findings from Rosenbaum and Rubin (1983), we know that if CIA holds for the covariates X, then it must also hold for $p(D = 1|X = x)$, which has only one dimension.
- Then we can apply the stratification estimation method to the propensity score.

We take the below steps:

- Obtain an estimate $\hat{e}(x)$ of the propensity score via non-parametric regression, and choose a number of strata J.
- Sort the observations according to their propensity scores, such that
$$\hat{e}(X_{i_1}) \leq \hat{e}(X_{i_2}) \leq \cdots \leq \hat{e}(X_{i_N}).$$
- Split the sample into J evenly sized strata using the sorted propensity score and, for each stratum, compute the simple difference-in-means estimate:
$$\hat{\tau}_j = \frac{\sum_{i=\lfloor (j-1)N/J \rfloor + 1}^{\lfloor jN/J \rfloor} D_i Y_i}{\sum_{i=\lfloor (j-1)N/J \rfloor + 1}^{\lfloor jN/J \rfloor} D_i} - \frac{\sum_{i=\lfloor (j-1)N/J \rfloor + 1}^{\lfloor jN/J \rfloor} (1 - D_i) Y_i}{\sum_{i=\lfloor (j-1)N/J \rfloor + 1}^{\lfloor jN/J \rfloor} (1 - D_i)}.$$
- Estimate the average treatment by averaging across strata:
$$\hat{\tau}_{PSTRAT} = \frac{1}{J} \sum_{j=1}^{J} \hat{\tau}_j.$$

Inverse Propensity Weighting (IPW)

- We can also exploit the propensity score in a weighting scheme.
- The Horvitz-Thompson estimator exploits the following two equalities:

$$E\left[\frac{D_i \cdot Y_i^{\text{obs}}}{e(X_i)}\right] = E[Y_i(1)] \tag{15}$$

$$E\left[\frac{(1 - D_i) \cdot Y_i^{\text{obs}}}{1 - e(X_i)}\right] = E[Y_i(0)] \tag{16}$$

- The proof exploits the Law of Iterated Expectation and the Unconfoundedness assumption.

Inverse Propensity Weighting (IPW)

- The two equalities suggest estimating $E[Y_i(1)]$ and $E[Y_i(0)]$ as

$$E[\widehat{Y_i(1)]} = \frac{1}{N}\sum_{i=1}^{N}\frac{D_i \cdot Y_i^{\text{obs}}}{e(X_i)} \tag{17}$$

$$E[\widehat{Y_i(0)]} = \frac{1}{N}\sum_{i=1}^{N}\frac{(1-D_i) \cdot Y_i^{\text{obs}}}{1-e(X_i)}, \tag{18}$$

- Then we can estimate the ATE using the below Horvitz-Thompson estimator:

$$\tilde{\tau}^{\text{ht}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i \cdot Y_i^{\text{obs}}}{e(X_i)} - \frac{(1-D_i) \cdot Y_i^{\text{obs}}}{1-e(X_i)}\right) \tag{19}$$

IPW

- We also call the above estimator "Oracle IPW" because we rarely know the propensity score in practice.
- Instead, we can weight using the estimated propensity score $\hat{e}(X_i)$, and use the estimator:

$$\hat{\tau}^{\text{ht}} = \frac{\sum_{i=1}^{N} \frac{D_i \cdot Y_i^{\text{obs}}}{\hat{e}(X_i)}}{\sum_{i=1}^{N} \frac{D_i}{\hat{e}(X_i)}} - \frac{\sum_{i=1}^{N} \frac{(1-D_i) \cdot Y_i^{\text{obs}}}{1-\hat{e}(X_i)}}{\sum_{i=1}^{N} \frac{1-D_i}{1-\hat{e}(X_i)}}. \tag{20}$$

- The estimation error in the propensity scores affects the accuracy of IPW, and how exactly the estimation error carries over is also complex. See Wager (2024) Chapter 2 for more discussion.

Unconfoundedness: Estimation

Doubly Robust Methods

Overview

- Recall that under unconfoundedness, we can expresse ATE using CATE:

$$\begin{aligned} ATE &= \sum CATE(x) \times p[X_i = x] \\ &= \sum \{\mathbf{E[Y_i|D_i = 1, X_i = x] - E[Y_i|D_i = 0, X_i = x]}\} \times p[X_i = x] \end{aligned} \tag{21}$$

- We denote $\mu_d(x) = E[Y_i|D_i = d, X_i = x]$ as the **Conditional Response Surface**
- We can also express ATE using the IPW weighting

$$\begin{aligned} ATE &= E[Y_i(1) - Y_i(0)] \\ &= \mathbf{E}\left[\frac{\mathbf{D_i \cdot Y_i^{obs}}}{\mathbf{e(X_i)}}\right] - \mathbf{E}\left[\frac{\mathbf{(1 - D_i) \cdot Y_i^{obs}}}{\mathbf{1 - e(X_i)}}\right] \end{aligned} \tag{22}$$

Augmented IPW

- So, to estimate ATE, we need to either estimate the propensity score and reweight, or to directly estimate the conditional response surface and then aggregate
- Combining both strategies, we have the augmented IPW (AIPW) estimator from Robins, Rotnitzky, and Zhao (1994):

$$\hat{\tau}_{AIPW} = \frac{1}{N}\sum[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + D_i\frac{Y_i - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - (1 - D_i)\frac{Y_i - \hat{\mu}_0(X_i)}{1 - \hat{e}(X_i)}] \tag{23}$$

AIPW is Doubly Robust

- An estimator is doubly robust if it is consistent if either the propensity score is consistent or the regression function is consistent.
- AIPW is doubly robust under unconfoundedness
- Wager (2024): “Qualitatively, AIPW can be seen as first making a best effort attempt at ATE by estimating the conditional response surfaces; then, it deals with any biases of the conditional response surfaces by applying IPW to the regression residuals.”

Why is AIPW Doubly Robust?

- If the conditional response surfaces are consistent, AIPW is consistent even if propensity score is inconsistent:

$$\hat{\tau}_{AIPW} = \frac{1}{N}\sum_{i=1}^{N} \underbrace{\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)\right)}_{\text{the regression estimator}}$$

$$+\frac{1}{N}\sum_{i=1}^{N} \underbrace{\left(\frac{D_i}{\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(1)}(X_i)\right) - \frac{1-D_i}{1-\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(0)}(X_i)\right)\right)}_{\approx\text{mean-zero noise}},$$

- If the propensity score is consistent, AIPW is consistent even if the conditional response surfaces are inconsistent:

$$\hat{\tau}_{AIPW} = \frac{1}{N}\sum_{i=1}^{N}\underbrace{\left(\frac{D_i Y_i}{\hat{e}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{e}(X_i)}\right)}_{\text{the IPW estimator}}$$

$$+\frac{1}{N}\sum_{i=1}^{N}\underbrace{\left(\hat{\mu}_{(1)}(X_i)\left(1-\frac{D_i}{\hat{e}(X_i)}\right) - \hat{\mu}_{(0)}(X_i)\left(1-\frac{1-D_i}{1-\hat{e}(X_i)}\right)\right)}_{\approx\text{mean-zero noise}}.$$

Heterogeneous Treatment Effect

Heterogeneous Treatment Effect

T-learner

Overview

- Sometimes, just knowing ATE is not enough to make decisions: a medicine that saves young people but kills the old might seem good "on average".
- So far, we use CATE as the building block for ATE, but CATE can be of direct interest to us too.
- Now we discuss how to estimate $CATE = \tau(x) = E[Y_i(1) - Y_i(0) | X_i = x]$

T-learner

- We have showed that under confoundedness, CATE can be expressed as the difference between two Conditional Response Surfaces:

$$\tau(x) = \mu_1(x) - \mu_0(x) \tag{24}$$

- T-learner: get consistent conditional response surfaces and then take the difference

$$\hat{\tau}_T(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x) \tag{25}$$

- T-leaner is consistent but suffers from the Regularization Bias in finite sample
- Why? We estimate the two conditional response surfaces separately, these two functions may end up being regularized in different ways from each other

Graphical Illustration of Regularization Bias from Wager (2024)

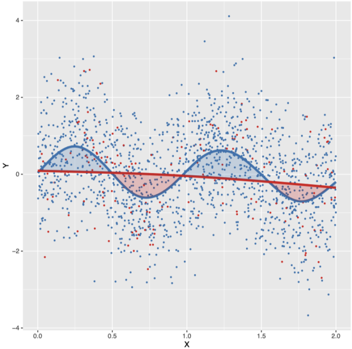

Figure 4.1: Illustration of regularization bias. Both control (blue) and treated (red) units are drawn from the same distribution. Data is generated from an RCT with $\pi = 0.1$, and so there are more controls than treated units. Spline regression learns a more oscillatory model for $\mu_{(0)}(x)$ and a flat one for $\mu_{(1)}(x)$. This results in an oscillatory CATE estimate, illustrated via shading, whereas the true CATE here is identically 0.

Heterogeneous Treatment Effect

A Semi-parametric Model

A Semiparametric Model

- Instead of estimating two conditional response surfaces separately, we can directly model the CATE function:

$$\tau(x) = \psi(x) \cdot \beta, \quad \psi : \mathcal{X} \to \mathbb{R}^d, \quad \beta \in \mathbb{R}^d. \tag{26}$$

- We put no functional form restrictions on the conditional response surfaces but instead put a parametric form on the CATE function
- We need the below CEF functions:

$$Y_i(1) = E[Y_i(1)|X_i = x] + \epsilon_i(1) \tag{27}$$

$$Y_i(0) = E[Y_i(0)|X_i = x] + \epsilon_i(0) \tag{28}$$

- Since $Y_i = D_i(Y_i(1) - Y_i(0)) + Y_i(0)$, we get:

$$Y_i = D_i\tau(x) + \mu_0(x) + D_i\epsilon_i(1) + (1 - D_i)\epsilon_i(0) \tag{29}$$

$$\begin{aligned} E[Y_i|X_i = x] &= m(x) \\ &= e(x)\tau(x) + \mu_0(x) + E[D_i\epsilon_i(1) + (1 - D_i)\epsilon_i(0)|X_i = x] \end{aligned} \tag{30}$$

Residual on Residual

- Take the difference with Y_i and $E[Y_i|X_i = x]$, we have the below model:

$$\begin{aligned} Y_i - m(X_i) &= (D_i - e(X_i))\tau(X_i) + \eta_i \\ &= (D_i - e(X_i))\psi(x) \cdot \beta + \eta_i, \end{aligned} \tag{31}$$

  where
  $\eta_i = D_i\epsilon_i(1) + (1 - D_i)\epsilon_i(0) - E[D_i\epsilon_i(1) + (1 - D_i)\epsilon_i(0)|X_i = x]$
- The above equation suggests a residual-on-residual estimator:
  1. Run non-parametric regressions $Y \sim X$ and $D \sim X$ using a flaxible method to get $\hat{m}(x)$ and $\hat{e}(x)$ respectively.
  2. Get the cross-fit residuals:

$$\tilde{Y}_i = Y_i - \hat{m}^{(-k(i))}(X_i), \quad \text{and} \tilde{Z}_i = \psi(X_i)(D_i - \hat{e}^{(-k(i))}(X_i)). \tag{32}$$

  3. Estimate $\hat{\beta}$ by running a linear regression $\tilde{Y}_i \sim \tilde{Z}_i$.

Further Reading

- Since we don't want to put strong functional forms on propensity score or the conditional response surfaces, machine learning techniques become particularly useful
- Susan Athey have discussed how to estimate HTE using machine learning methods and it is easy to find relevant resources
- For example, you can check Athey's 2018 AEA Continuing Education `https://bit.ly/2CGLfes`
- This strand of literature is making rapid progress, so make sure that you check the latest update when you need to implement these methods