

Notes of Causal Inference

Observational studies

Yue Wu

1 Introduction

[Why not express the results in words and highlight them!]

[A question: can we estimate CATE and impute the x_i' s in and get the estimation of ATE?]
(anything to do with balance?)

[remember to check $Y(0), Y(1) / Y(1), Y(0)$]

In previous chapters, we talk about , and following that, This note mainly focuses on causal inference with covariates. This is worth considering for the situation when we observe $\{X_i, Z_i, Y_i\}$ pairs.

Many contents comes with the coming information of covariates:

- power to estimate treatment effect under non-random-treatment-assignment situation
- assumption of unconfoundedness and overlap
- possibility to estimate CATE, and to estimate ATE by CATE
- propensity score
- covariate balance
- etc

The difference of OB and RCT situation.

- fundamentally, limited/super population
- the way and why we use X

2 how we understand the i.i.d sample

In observational point of view, we view the observed units as observed samples from an unknown population, which we make assumptions and inference of. This is more like the traditional statistical analysis, with sample is IID and stochastic.

$$\{X_i, Z_i, Y_i(1), Y_i(0)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \{X, Z, Y(1), Y(0)\}. \quad (1)$$

Since we can factorize the joint distribution into a product of conditional distributions, there are many ways to understand the sample data $\{X_i, Z_i, Y_i\}_{i=1}^n$ from the **data-generating point of view**:

- $\Pr(X, Z, Y(1), Y(0)) = \Pr(X) \times \Pr(Z, Y(1), Y(0) | X)$

We first sample $\{X_i\}_{i=1}^n$ from $\{X\}$, then for each of the n -large dataset, we sample $Z_i, Y_i(1), Y_i(0)$ from n same and independent distributions $\{Z_i, Y_i(1), Y_i(0)\} | X_i$ (we can write them as $\{Z, Y(1), Y(0) | X\}$).

Therefore, we can view the sample $\{X_i\} + \{Z_i, Y_i(1), Y_i(0)\} = \{X_i, Y_i(1), Y_i(0), Z_i\}$ as i.i.d sample of $\{X\} \times \{Z, Y(1), Y(0) | X\} = \{X, Z, Y(1), Y(0)\}$

- $\Pr(X, Z, Y(1), Y(0)) = \Pr(X) \times \Pr(Y(1), Y(0) | X) \times \Pr(Z | X, Y(1), Y(0))$

We first sample $\{X_i\}_{i=1}^n$ from $\{X\}$, then for each of the n -large dataset, we sample $Y_i(1), Y_i(0)$ from n same and independent distributions $\{Y_i(1), Y_i(0) | X_i\}$ (we can write them as $\{Y(1), Y(0) | X\}$). Then for each of the n -large $X_i, Y_i(1), Y_i(0)$ dataset, we sample Z_i from n same and independent distributions $\{Z_i | X_i, Y_i(1), Y_i(0)\}$ (we can write them as $\{Z | X, Y(1), Y(0)\}$).

Therefore, we can view the sample $\{X_i\} + \{Y_i(1), Y_i(0)\} + \{Z_i\} = \{X_i, Y_i(1), Y_i(0), Z_i\}$ as i.i.d sample of $\{X\} \times \{Y(1), Y(0) | X\} \times \{Z | X, Y(1), Y(0)\} = \{X, Z, Y(1), Y(0)\}$

- $\Pr(X, Z, Y(1), Y(0)) = \Pr(X) \times \Pr(Z | X) \times \Pr(Y(1), Y(0) | X, Z)$

We first sample $\{X_i\}_{i=1}^n$ from $\{X\}$, then for each of the n -large dataset, we sample Z_i from n same and independent distributions $\{Z_i | X_i\}$ (we can write them as $\{Z | X\}$). Then for each of the n -large X_i, Z_i dataset, we sample $Y_i(1), Y_i(0)$ from n same and independent distributions $\{Y_i(1), Y_i(0) | X_i, Z_i\}$ (we can write them as $\{Y(1), Y(0) | X, Z\}$).

Therefore, we can view the sample $\{X_i\} + \{Z_i\} + \{Y_i(1), Y_i(0)\} = \{X_i, Z_i, Y_i(1), Y_i(0)\}$ as i.i.d sample of $\{X\} \times \{Z | X\} \times \{Y(1), Y(0) | X, Z\} = \{X, Z, Y(1), Y(0)\}$

The above distribution $\{X, Z, Y(1), Y(0)\}$ is what we want to make inference of, but we can not observe its sample realization due to the fundamental missing problem. And what we can observe fully, Y , can be seemed as a nonparametric function of $\{Z, Y(1), Y(0)\}$, which is still a random variable.

$$Y = ZY(1) + (1 - Z)Y(0), \quad (2)$$

$$\{X_i, Z_i, Y_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \{X, Z, ZY(1) + (1 - Z)Y(0)\}. \quad (3)$$

3 Unconfoundedness

Assumption 1 (Unconfoundedness).

$$Y(z) \perp\!\!\!\perp Z | X \text{ for } z = 0, 1$$

Since Z is a binary random variable, we can also express this assumption as:

$$\Pr(Y(z) | Z = 1, X) = \Pr(Y(z) | Z = 0, X) \text{ for } z = 0, 1 \quad (4)$$

Under assumption 1, for $z = 0, 1$, we have

$$\begin{aligned} \Pr(Y | X, Z = z) &= \Pr(ZY(1) + (1 - Z)Y(0) | X, Z = z) \\ &= \Pr(zY(1) + (1 - z)Y(0) | X, Z = z) \\ &= \Pr(Y(z) | X, Z = z) \\ &= \Pr(Y(z) | X) \text{ (given } Y(z) \perp\!\!\!\perp Z | X) \end{aligned} \quad (5)$$

Which means the observable distribution of Y in each treatment arm (condition on) $Z = z$ equals the distribution of the potential outcome $Y(z)$, given covariates X . Unconfoundedness allows us to reach 'random assignment' within subpopulations defined by values of observed covariates $X = x$.

3.1 two forms of identification with X

Hence, we can express the mean of outcomes in both arms with the observed distribution, which means that we successfully reach identifiability of $E(Y(z) | X)$ by **writing the estimands as functions of distribution we can observe**. With the help of (conditional on) covariates X , two ways of identifications are

$$E(Y(1) | X) = E(Y | X, Z = 1), E(Y(0) | X) = E(Y | X, Z = 0) \quad (6)$$

$$E(Y(1) | X) = E\left(\frac{ZY}{e(X)}\right), \quad E(Y(0) | X) = E\left(\frac{(1-Z)Y}{1-e(X)}\right) \quad (7)$$

Therefore, for the comparison/function of the outcome means, CATE, we have the following two forms of identifications:

$$\tau(X) = E(Y | X, Z = 1) - E(Y | X, Z = 0) \quad (8)$$

$$\tau(X) = E\left(\frac{ZY}{e(X)}\right) - E\left(\frac{(1-Z)Y}{1-e(X)}\right) \quad (9)$$

3.2 reduce the causal question to two separate conditional distributions

In general, what we do in causal inference We can factorize the joint distribution as

$$\Pr(X, Z, Y(1), Y(0)) = \Pr(X) \times \Pr(Z, Y(1), Y(0) | X) \times \Pr(Z | X, Y(1), Y(0)) \quad (10)$$

where

(1) $\Pr(X)$ is the covariate distribution, which we often do not model because they are background information happening before the treatment and outcome,

(2) $\Pr(Y(1), Y(0) | X)$ is the outcome distribution conditional on the covariates X , which we often model to get the estimation [...], [we care about this!!!!]

(3) $\Pr(Z | X, Y(1), Y(0))$ is the treatment distribution conditional on the covariates X , also known as the treatment assignment mechanism. Assumptions(Unconfoundedness, Overlap) are made for this distribution to ensure that we can make estimation of causal estimands.

What we do in causal inference: make assumption of $\Pr(Z | X, Y(1), Y(0))$ and infer the characteristics of $\Pr(Y(1), Y(0) | X)$

Power of unconfoundedness Under unconfoundedness,

$$\begin{aligned} \Pr(Z, Y(1), Y(0) | X) &= \Pr(Y(1), Y(0) | X) \times \Pr(Z | \cancel{Y(1)}, \cancel{Y(0)}, X) \\ &= \Pr(Z | X) \times \Pr(Y(1), Y(0) | \cancel{Z}, X) \end{aligned} \quad (11)$$

By making conditional independence assumption, we actually **simplify the problem of modeling Z and $Y(1), Y(0)$ and X together -> just modeling $Z | X$ and $Y(1), Y(0) | X$**

X , two separate independent distributions.

Under this assumption, the 'relation of Z and X ' and 'relation of $Y(1), Y(0)$ and X ' has nothing to do with each other.

So the question is reduced to two separate distributions: (1) **the outcome model** $\Pr(Y(1), Y(0) | X)$ and (2) **the treatment assignment/propensity model** $\Pr(Z | X)$.

4 Overlap

Assumption 2 (Overlap).

$$0 < \Pr(Z = 1 | X) < 1$$

[...]

5 CATE

5.1 two estimands

With the information of covariates X , we can estimate the Conditional average treatment effect(CATE). We mainly talk about the following two estimands in observational studies.

- Average treatment effect(ATE)

$$\tau = E[Y(1) - Y(0)] = E(Y(1)) - E(Y(0)) \quad (12)$$

Average treatment effect is important when we wish to know the average of the causal effect for the population of interest. And when the estimand is ATE, we do not really care if the effect differs with different value of X , or to say that we care more about $\{Y(1), Y(0)\}$ instead of $\{Y(1), Y(0) | X\}$.

- Conditional average treatment effect(CATE)

$$\tau(X) = E(Y(1) - Y(0) | X) = E(Y(1) | X) - E(Y(0) | X)$$

Conditional average treatment effect is important when we wish to know the average of the causal effect conditional on the value of X for the population of interest. This value maybe related to X , i.e. there is heterogeneity for the causal effect. In other words, we care about $\{Y(1), Y(0) | X\}$.

Note that the ATE is a scalar value, so we do not necessarily need a model assumption to estimate this, and itself is the estimand. However, for CATE under heterogeneous effect, as it is a function of X , so we (usually?) need to make a model assumption, and the estimands are the parameters of the model.

By introducing the information of X and making conditional assumptions on X , we actually go into a world of parametricity.

5.2 relationship between ATE and CATE

5.2.1 from (single) outcome model view

We know that, for a certain population $\{X, Y(1), Y(0)\} \sim P$, ATE is related to CATE as

$$\tau(X) = E_P(Y(1) - Y(0) | X)$$

$$\tau = E_P(Y(1) - Y(0)) = E_P(E_P(Y(1) - Y(0) | X)) = E_P(\tau(X))$$

and we assume X_i are i.i.d samples of $\{X\}$, we can also estimate ATE by first estimating the CATE function $\hat{\tau}(x)$ and averaging its value with plug-in X_i .[??]

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \hat{\tau}(X_i), \quad i = 1, \dots, n$$

(Note that in this step, there is nothing to do with the assignment model [!!!] $\tau = E(\tau(X))$ is correct, but $Pr(Y(1)) \neq Pr(Y | Z = 1)$ since

$$E_{population}(Y(z) | X) = E(Y | X, Z = z)$$

5.2.2 response surfaces - from outcome model & treatment assignment model view

Suppose we have $\{X, Z, Y\} \sim \{X, Z, ZY(1) + (1 - Z)Y(0)\}$ which can be separated into $\{X, Y\} | Z = 1 \sim P_1$ and $\{X, Y\} | Z = 0 \sim P_0$. In P_z the conditional expectation of the variable Y given a particular value of X is called the response surface for Y in P_z , and at $X = x$ is denoted $R_z(x)$.

$$R_z(x) = E(Y | X = x, Z = z) = E(Y(z) | X = x, Z = z)$$

The difference in response surfaces at $X = x$, $R_1(x) - R_0(x)$, is the effect of the treatment variable at $X = x$.

$$\tau_{PF}(X) = R_1(X) - R_0(X) = E(Y(1) | X, Z = 1) - E(Y(0) | X, Z = 0)$$

Note that this expectation is not CATE $\tau(X) = E(Y(1) - Y(0) | X)$. But,

When we assume unconfoundedness we have $E(Y(1) | X, Z = 1) = E(Y(1) | X, Z = 0) = E(Y(1) | X)$, $E(Y(0) | X, Z = 1) = E(Y(0) | X, Z = 0) = E(Y(0) | X)$, and the difference between response surfaces equals CATE and the average difference of that is ATE.

6 Inverse propensity score weighting(IPW)

Recall that if $Y(z)$ and Z are independent, we have the equation $E(\tau) = E\left(\frac{ZY}{Pr(Z=1)} - \frac{(1-Z)Y}{1-Pr(Z=1)}\right)$, which motivates the nonparametric(no model assumption, without X) IPW estimator, $\hat{\tau}^{ipw} = \frac{1}{n} \sum_{i=1}^n \frac{Z_i Y_i}{Pr(Z=1)} - \frac{1}{n} \sum_{i=1}^n \frac{(1-Z_i) Y_i}{1-Pr(Z=1)}$. This is unbiased if $Y(z)$ and Z are independent. But now the assumption is conditional independence, and we can think of a similar estimator with X inside, which is a parametric way of estimation.

Equation(9) motivates the following nonparametric(moment-based) estimator of ATE, which is the difference in the mean of the weighted outcomes between group:

$$\hat{\tau}^{ipw} = \frac{1}{n} \sum_{i=1}^n \frac{Z_i Y_i}{e(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - Z_i) Y_i}{1 - e(X_i)}, \quad (13)$$

It is HT estimator in survey sampling [3]. and further in [4].

It is called IPW because we can see this in the way of $\hat{\tau}^{ipw} = \frac{1}{n} \sum_{i=1}^n Z_i Y_i w_1(X_i) - \frac{1}{n} \sum_{i=1}^n (1 - Z_i) Y_i w_0(X_i)$ where the weights are defined by the inverse of the probability of the unit being sampled($e(X_i)$)[?]

$$\begin{cases} w_1(X_i) = \frac{1}{e(X_i)}, & \text{for } Z_i = 1 \\ w_0(X_i) = \frac{1}{1-e(X_i)}, & \text{for } Z_i = 0. \end{cases} \quad (14)$$

Typically, we standardize the weights and get ipw estimator as:

$$\hat{\tau}^{ipw} = \frac{\sum_{i=1}^n Z_i Y_i / e(X_i)}{\sum_{i=1}^n Z_i / e(X_i)} - \frac{\sum_{i=1}^n (1 - Z_i) Y_i / (1 - e(X_i))}{\sum_{i=1}^n (1 - Z_i) / (1 - e(X_i))}, \quad (15)$$

[why standardize?]

6.1 properties

Since the propensity score $e(X)$ is unknown(what if we know the true $e(X)$?[2]), we use the sample to estimate this function(usually a binary logistic model), and plug-in the estimated function to the estimator.

$$\hat{\tau}^{ipw} = \frac{1}{n} \sum_{i=1}^n \frac{Z_i Y_i}{\hat{e}(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - Z_i) Y_i}{1 - \hat{e}(X_i)}, \quad (16)$$

If the postulated PS model is true, then $\hat{\tau}^{ipw}$ is consistent.

7 Outcome Regression

[what is exact(calculation results) of 'dengjia'? what is 'dengjia' in statistical world]

[impute the missing/impute both]

Motivated by equation(8), we can define $m_z(X) = E(Y(z) | X)$, which are estimable with sample dataset, with certain model assumptions. If we specify a model with parameter α , then we get the estimation of $\hat{m}_z(X) = m_z(X; \hat{\alpha})$.

So the estimator of ATE is

$$n^{-1} \sum_{i=1}^n \{\hat{m}_1(X_i) - \hat{m}_0(X_i)\}$$

, or

$$\hat{\tau}_{adj} = n^{-1} \sum_{i=1}^n \{Z_i(Y_i - \hat{m}_0(X_i)) + (1 - Z_i)(\hat{m}_1(X_i) - Y_i)\}$$

[property If the specified regression model is true, then $\hat{\tau}$ is consistent and efficient(?), but not otherwise]

A canonical choice of model assumption is to specify a regression relationship, and there is multiple ways of defining that.

7.1 Linear, No heterogeneity model assumption

If we assume a linear relation between the potential outcomes $Y(1), Y(0)$ conditional on X as

$$E(Y(z) | X) = \beta_0 + \beta_Z z + \beta_X^\top X, \quad z \in \{0, 1\} \quad (17)$$

which, with assumption 1, is equivalent with a linear relation of observed outcome Y conditional on X in group $Z = z$

$$\begin{aligned} E(Y(z) | X) &= \mu_z(X) = E(Y | X, Z = z) \\ &= \beta_0 + \beta_Z z + \beta_X^\top X, \quad z \in \{0, 1\} \end{aligned} \quad (18)$$

then we have

$$\begin{aligned} \tau(X) &= E(Y | X, Z = 1) - E(Y | X, Z = 0) \\ &= (\beta_0 + \beta_Z + \beta_X^\top X) - (\beta_0 + \beta_X^\top X) \\ &= \beta_Z \end{aligned} \quad (19)$$

By assuming this, we actually assume the treatment effect is homogeneous with no change according to the change of X . So CATE equals ATE for all value of X , and the estimands of both of them is a scalar. In this case, the information of X is for the unbiasedness estimation of ATE, we do not use it for getting a various value of CATE.

7.1.1 estimation of CATE(=ATE)

Equation(17) tells us that, in both treatment arms $Z = z$, the expectation of observed Y is a linear function of covariate X . Given Z is a binary variable and linearity, this also means that the observed Y is a linear function of both Z and X , as

$$E(Y | Z, X) = \beta_0 + \beta_Z Z + \beta_X^\top X$$

Now the estimation question reduces to a classic linear regression question.

How can we estimate the parameter of interest, β_Z ? An intuitive way is doing the regression using all data.

Regress $Y \sim Z + X$ and get the coefficient of Z as $\hat{\beta}_Z$

$$(\hat{\beta}_0, \hat{\beta}_Z, \hat{\beta}_X) = \arg \min_{\beta_0, \beta_Z, \beta_X} \frac{1}{n} \sum_{i=1}^n (Y_i - \beta_0 - \beta_Z Z_i - \beta_X^\top X_i)^2$$

$$\hat{\tau}(X) = \hat{\beta}_Z$$

Which is an OLS estimator, and it is unbiased and consistent.

$$E(\hat{\tau}) = E(\hat{\beta}_Z) = \beta_Z = \tau = \tau(X) \quad (20)$$

Here we regard X_i fixed and Y as a function of random variable Z (depending on X , sampled from $P(Z | X)$) and fixed X therefore a random variable depending on X .

$$\begin{aligned} P(Y | Z, X) &= P(Y, Z, X) / P(Z, X) \\ &= P(Y, Z | X) P(X) / P(Z | X) P(X) \\ &= P(Y, Z | X) / P(Z | X) \\ &= \end{aligned} \quad (21)$$

7.1.2 do we need balance? in this estimator

Let's think about another ways: In each treatment group $Z = z$, regress $Y \sim X$ and get the estimation of $\hat{\beta}_{0,z}$, then diff

$$(\hat{\beta}_{0,z}, \hat{\beta}_{X,z}) = \arg \min_{\beta_{0,z}, \beta_{X,z}} \frac{1}{n_z} \sum_{i:Z_i=z} (Y_i - \beta_{0,z} - \beta_{X,z}^\top X_i)^2, \quad z \in \{0, 1\}.$$

$$\hat{\tau}(X) = \hat{\beta}_{0,1} - \hat{\beta}_{0,0}$$

It is under the same assumption within this section, but what happens to our estimation using this way? It is still unbiased.

This estimator, which use separately each treatment group data for estimating .

Two separate models for each treatment group vs. one unified model with treatment indicator?

7.2 Linear, Heterogeneity model assumption

If we assume

$$E(Y(z) | X) = \beta_0 + \beta_Z z + \beta_X^\top X + \beta_{ZX}^\top X z, \quad z \in \{0, 1\} \quad (22)$$

then we have

$$\begin{aligned} \tau(X) &= E(Y | X, Z = 1) - E(Y | X, Z = 0) \\ &= (\beta_0 + \beta_Z + \beta_X^\top X + \beta_{ZX}^\top X) - (\beta_0 + \beta_X^\top X) \\ &= \beta_Z + \beta_{ZX}^\top X \end{aligned} \quad (23)$$

which assume that the treatment effect heterogeneity exists. This, coupled with ignorability[?], implies that

$$\begin{aligned} \tau &= E(\tau(X)) \\ &= E(\beta_Z + \beta_{ZX}^\top X) \\ &= \beta_Z + \beta_{ZX}^\top E(X) \end{aligned} \quad (24)$$

7.2.1 estimation of CATE & ATE

Equation(22) tells us that, the expectation of observed Y is a linear function of covariate Z and X with interaction term, as

$$E(Y | Z, X) = \beta_0 + \beta_Z Z + \beta_X^\top X + \beta_{ZX}^\top X Z$$

So how can we design estimator of $\tau(X)$ and τ under this setting? The parameters of the model can be classically estimated by OLS as

$$(\hat{\beta}_0, \hat{\beta}_Z, \hat{\beta}_X, \hat{\beta}_{ZX}) = \arg \min_{\beta_0, \beta_Z, \beta_X, \beta_{ZX}} \frac{1}{n} \sum_{i=1}^n (Y_i - \beta_0 - \beta_Z Z_i - \beta_X^\top X_i - \beta_{ZX}^\top (Z_i \cdot X_i))^2.$$

Hence, the estimated CATE is

$$\hat{\tau}(X) = \hat{\beta}_Z + \hat{\beta}_{ZX}^\top X. \quad (25)$$

What about ATE? As $\tau = E(\tau) = E(E(\tau(X) | X)) = \beta_Z + \beta_{ZX}^\top E(X)$, so if we know the true population mean of X , we can directly get the estimation of ATE as

$$\hat{\tau} = \hat{\beta}_Z + \hat{\beta}_{ZX}^\top E(X).$$

But since we do not know about the distribution of X , we can use the mean of it.[Does this harm unbiasedness?]

$$\hat{\tau} = \hat{\beta}_Z + \hat{\beta}_{ZX}^\top \frac{1}{n} \sum_{i=1}^n X_i.$$

7.2.2 properties of estimator

[prove that ATE is unbiased]

do we need balance? in this estimator

Firstly, is β_Z and β_{ZX} unbiased and consistent?

We can prove that ATE is unbiased as:

$$\begin{aligned}\hat{\tau} &= E(\hat{\beta}_Z + \hat{\beta}_{ZX}^\top \frac{1}{n} \sum_{i=1}^n X_i) \\ &= E(\hat{\beta}_Z) + E(\hat{\beta}_{ZX}^\top \frac{1}{n} \sum_{i=1}^n X_i) \\ &= \beta_Z + E(E(|))) \\ &= \beta_Z + \beta_{ZX} E(X) \\ &= \tau\end{aligned}\tag{26}$$

7.3 General(not necessarily linear), complex model assumption

7.4 Discussion

It is plausible to conditional on X and regard it as fixed, but we can never see Z as fixed.

A mixture of Sample OLS and Population OLS.

What if the linear model is wrong?

Unlike randomized experiments, the estimator is not consistent if the linear model is misspecified.

[[??]]If the imbalance of the covariates between the two groups is large, the model-based results heavily relies on extrapolation in the region with little overlap, which is sensitive to the model specification assumption.

8 DR estimator

9 About high-dimensional case

9.1 example of high-dimensional causal cases

IT company have many covariates.

Genetic analysis.

9.2 the two assumptions

see [1]

10 Some other discussions

10.1 Estimands with X

- individual treatment effect(ITE)

$$\tau_i = Y_i(1) - Y_i(0)$$

- conditional average causal effect(CATE)

$$\tau(X) = E(Y(1) - Y(0) | X)$$

The individual treatment effect(ITE) refers to the difference of potential outcomes for unit i , which is a difference and has nothing to do with distribution since the $Y_i(1)$ and $Y_i(0)$ are fixed values.

The conditional average causal effect(CATE) is the average of the individual treatment effect of all units with the covariate value x.

10.2 why we need X in randomized

Already unbiased.

1. X for covariate adjustment
2. X for designing a better assignment

10.3 what we assume, and what we observe

- definitly of randomized study, because assumption is fact

- overlap and covariate balance(the relation of plausible(?) assumption and plausible(?) sample, intertextuality) [search 'balancing score' in DPbook]

10.4 Designing a Randomized Experiment

can we better design a randomized experiment with some information known?(when we know X and we can do stratified sampling, when we have results from OB data we collected, Bayesian view for example we know the mean of some parameter(FanLi's work))

10.5 multi-treatment levels

11 summary

A proofs

The unbiasedness of $\hat{\tau}$ follows from the representation

$$\hat{\tau}^{DIM} = n_1^{-1} \sum_{i=1}^n Z_i Y_i - n_0^{-1} \sum_{i=1}^n (1 - Z_i) Y_i = n_1^{-1} \sum_{i=1}^n Z_i Y_i(1) - n_0^{-1} \sum_{i=1}^n (1 - Z_i) Y_i(0)$$

and the linearity of the expectation:

$$\begin{aligned} E(\hat{\tau}^{DIM}) &= E\left\{ n_1^{-1} \sum_{i=1}^n Z_i Y_i(1) - n_0^{-1} \sum_{i=1}^n (1 - Z_i) Y_i(0) \right\} \\ &= n_1^{-1} \sum_{i=1}^n E(Z_i) Y_i(1) - n_0^{-1} \sum_{i=1}^n E(1 - Z_i) Y_i(0) \\ &= n_1^{-1} \sum_{i=1}^n \frac{n_1}{n} Y_i(1) - n_0^{-1} \sum_{i=1}^n \frac{n_0}{n} Y_i(0) \\ &= n^{-1} \sum_{i=1}^n Y_i(1) - n^{-1} \sum_{i=1}^n Y_i(0) \\ &= \tau. \end{aligned}$$

The variance of $\hat{\tau}$ follows from Lemma C.2 of simple random sampling:

$$\begin{aligned} \text{var}(\hat{\tau}^{DIM}) &= \frac{n_1 n_0}{n(n-1)} \sum_{i=1}^n \left\{ \frac{Y_i(1)}{n_1} + \frac{Y_i(0)}{n_0} - \frac{\bar{Y}(1)}{n_1} - \frac{\bar{Y}(0)}{n_0} \right\}^2 \\ &= \frac{n_1 n_0}{n(n-1)} \left[\frac{1}{n_1^2} \sum_{i=1}^n \{Y_i(1) - \bar{Y}(1)\}^2 + \frac{1}{n_0^2} \sum_{i=1}^n \{Y_i(0) - \bar{Y}(0)\}^2 \right. \\ &\quad \left. + \frac{2}{n_1 n_0} \sum_{i=1}^n \{Y_i(1) - \bar{Y}(1)\} \{Y_i(0) - \bar{Y}(0)\} \right] \\ &= \frac{n_0}{n_1 n} S^2(1) + \frac{n_1}{n_0 n} S^2(0) + \frac{2}{n} S(1, 0). \end{aligned}$$

From Lemma 4.1, we can also write the variance as

$$\begin{aligned} \text{var}(\hat{\tau}^{DIM}) &= \frac{n_0}{n_1 n} S^2(1) + \frac{n_1}{n_0 n} S^2(0) + \frac{1}{n} \{S^2(1) + S^2(0) - S^2(\tau)\} \\ &= \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n}. \end{aligned}$$

References

- [1] Alexander D'Amour et al. "Overlap in observational studies with high-dimensional covariates". In: *Journal of Econometrics* 221.2 (2021), pp. 644–654.
- [2] Keisuke Hirano, Guido W Imbens, and Geert Ridder. "Efficient estimation of average treatment effects using the estimated propensity score". In: *Econometrica* 71.4 (2003), pp. 1161–1189.
- [3] Daniel G Horvitz and Donovan J Thompson. "A generalization of sampling without replacement from a finite universe". In: *Journal of the American statistical Association* 47.260 (1952), pp. 663–685.
- [4] Paul R Rosenbaum. "Model-based direct adjustment". In: *Journal of the American statistical Association* 82.398 (1987), pp. 387–394.