Jai-Hua Kevin Yen
kevinyen@buffalo.edu

Casual Inference: What If Notes

2025-12-24
Version 1.0

Chapter 1: A Definition of Casual Effect

Casual Effect: Compare action A (two options for this action: taken, withheld)
For example, for a dichotomous treatment A, 1 to be treated and 0 to be not treated.

Potential Outcome: Let Y (1: death, 0: survival) be the outcome of interest. $Y^{a=1}$, $Y^{a=0}$ be the potential outcomes or called counterfactual outcome. Note that these are random variables.

Consistency: counterfactual outcome = observed outcome
Looks "obvious" but sometimes this assumption would be violated.
Question: Why is this assumption needed since we definitely see the observed outcome?

Average Causal Effect: Since identify individual causal effect is not possible, turn into retreiving "aggregated" causal effect. The ingredients needed here are:

- identify outcome of interest
- $a = 0$ vs. $a = 1$ to be compared
- $Y^{a=1}$ vs. $Y^{a=0}$ to be compared

This exists if: $P(Y^{a=1} = 1) \neq P(Y^{a=0} = 1)$ or in general, $E(Y^{a=1}) \neq E(Y^{a=0})$.

Note: Sharp causal null hypothesis is a test for no causal effect for any individual.

In this whole book, we have the following assumptions:

- Assumption of no interference: an individual's outcome is independent of othersone's treatment
- Assumption of no multiple versions of treatment: definition of a counterfactual outcome under treatment value a also implicitly assumes that there is only one version of treatment value $A = a$.

These assumptions are all one of the stable unit treatment value assumptions (SUTVA).

Meausre of Causal Effect: To represent causal null,

1. $P(Y^{a=1} = 1) - P(Y^{a=0} = 1) = 0$ causal risk difference (average individual causal effect)
2. $\frac{P(Y^{a=1}=1)}{P(Y^{a=0}=1)} = 1$ causal risk ratio
3. $\frac{P(Y^{a=1}=1)/P(Y^{a=1}=0)}{P(Y^{a=0}=1)/P(Y^{a=0}=0)} = 1$ causal odds ratio

These are called the effect measures.

Number Need to Treat (NNT): On average, NNT measures how many people we need to treat in order to save "1" life. i.e.

$$NNT = \frac{-1}{P(Y^{a=1} = 1) - P(Y^{a=0} = 1)}$$

Random Variability: The procedure to compute effect measures is implausible, and is from two variability:

1. sampling variability
2. non-deterministic counterfactuals (stochastic): we then aim to measure the average counterfactual outcome of population

$$\begin{aligned} E(Y^a) &= E[E(Y^a|\Theta_{Y^a}(\cdot))] \\ &= \int y dE[\Theta_{Y^a}(\theta)] \end{aligned}$$

where $\Theta_{Y^a}(\cdot)$ is the individual-specific statistical distribution.

Note that until Chapter 10, we ignore those two random variabilities.

Causation versus association: Due to the fact that we can only observe one of the potential outcomes, here we define the observed outcome Y from the treatment level A. Then the independent is defined in terms of Y and A. Some equivalent definitions of independence are:

1. $P[Y = 1|A = 1] - P[Y = 1|A = 0] = 0$ associational risk difference
2. $\frac{P[Y=1|A=1]}{P[Y=1|A=0]} = 1$ associational risk ratio
3. $\frac{P[Y=1|A=1]/P[Y=0|A=1]}{P[Y=1|A=0]/P[Y=0|A=0]} = 1$ associational odds ratio

Thus, when treatment A and outcome Y are dependent or associated, we can see $P[Y = 1|A = 1] \neq P[Y = 1|A = 0]$. These are called the association measures. In general, we quantify the association using expectation $E[Y = 1|A = 1] \neq E[Y = 1|A = 0]$.

Most Important Idea of Chapter 1: Inferences about causation are concerned with what if questions in counterfactual worlds, such as "what would be the risk if everybody had been treated?" and "what would be the risk if everybody had been untreated?", whereas inferences about association are concerned with questions in the actual world, such as "what is the risk in the treated?" and "what is the risk in the untreated?"

Chapter 2: Randomized Experiments

The reality of our "actual data": data are missing for the counterfactual outcome.
$\Rightarrow$ Randomization ensures the missing values occur by chance and somehow controled.

Ideal Randomized Experiment:

- no loss to follow up
- full adherence to the assigned treatment
- single version of treatment
- double blind assignment

Exchangeability: The risk would be the same if the treatment and control group switched. i.e.,

$$P(Y^a = 1|A = 1) = P(Y^a = 1|A = 0) = P(Y^a = 1) \qquad (*)$$

or $Y^a \perp\!\!\!\perp A$ for all a.
It is also called **exogeneity** in causal inference.
$\Rightarrow$ In ideal randomized experiments, association is causation.

Note:

- $Y^a \perp\!\!\!\perp A$ does not have the same meaning of $Y \perp\!\!\!\perp A$. i.e., if treatment has causal effect on the outcome. $\Rightarrow Y \not\perp\!\!\!\perp A$ since treatment is associated with observed outcome.
- $E[Y^a|A = 1] = E[Y^a|A = 0]$: mean exchangeability. For continuous, exchangeability $\Rightarrow$ mean exchangeability. But the other direction does not always true. The reason is that other distributional parameters other than mean (variance) may not be independent of the treatment.
- A study is a randomized experiment even if exchangeability does not hold.

Marginally Randomized Experiments: use several randomized probabilities that depend on the value of a variable (e.g. prognostic factor L)

Conditionally Randomized Experiments: use a single unconditional randomized probability common to all individuals Note: Conditionally randomized experiments will not generally result in exchangeability. If this experiment is simply combinations of marginally randomized experiments, then it is a marginally randomized experiments, it is exchangeable. i.e.,

$$P(Y^a = 1|A = 1, L = 1) = P(Y^a = 1|A = 0, L = 1)$$

or $Y^a \perp\!\!\!\perp A|L = 1$ for all a. For $L = 0$, it also satisfies the formula above.

When $Y^a \perp\!\!\!\perp A|L = l$ for all values of l, we say $Y^a \perp\!\!\!\perp A|L$.

Note:

- conditional randomization $\rightarrow$ conditional exchangeability

- marginally randomization → marginal exchangeability, conditional exchangeability

Under marginal exchangeability:

causal ratio: $\frac{P(Y^{a=1}=1)}{P(Y^{a=0}=1)} = \frac{P(Y=1|A=1)}{P(Y=1|A=0)}$ **associational risk ratio** holds since exchangeability

Way to compute causal risk ratio in conditional randomized experiment:
Recall: conditionally randomized experiment is combination of marginal randomized experiments.
$\because$ association is causation in each subset

1. If $P(Y^{a=1}=1|L=1)/P(Y^{a=0}=1|L=1) = P(Y=1|L=1,A=1)/P(Y=1|L=1,A=0)$ and $P(Y^{a=1}=1|L=0)/P(Y^{a=0}=0|L=0) = P(Y=1|L=0,A=1)/P(Y=1|L=0,A=0)$ are different $\implies$ stratification: effect modification by L or treatment effect heterogeneity across levels of L.
2. Average causal effect: $P(Y^{a=1}=1)/P(Y^{a=0}=1)$

$$P(Y^a=1) = w_0 \cdot P(Y^a=1|L=0) + w_1 \cdot P(Y^a=1|L=1)$$

weighted average of stratum-specific risks

$$\implies P(Y^a=1) = P(Y^a=1|L=0)\cdot P(L=0) + P(Y^a=1|L=1)\cdot P(L=1)$$

Or in general,

$$P(Y^a=1) = \sum_l P(Y^a=1|L=l)P(L=l)$$

Under conditional exchangeability,

$$P(Y^a=1) = \sum_l P(Y=1|L=l,A=a)P(L=l)$$

(Here just replace $P(Y^a=1|L=l)$ with $P(Y=1|L=l,A=a)$)

If a counterfactual quantity can be expressed as a function of distribution (or probability) of observed data, we say the quantity is identified (or identifiable).

Standardization

$$\frac{P(Y^{a=1}=1)}{P(Y^{a=0}=1)} = \frac{\sum_L P(Y=1|L=l,A=1)P(L=l)}{\sum_L P(Y=1|L=l,A=0)P(L=l)}$$

Standardized risk: the "counterfactual" risk that would have been observed had all the individuals in the population been treated under conditional exchangeability

IP weights

$$W^A = 1/f(A|L)$$

Example: A treated individual with $L = l$ receives weight $\frac{1}{P(A=1|L=l)}$. An untreated individual with $L = l'$ receives weight $\frac{1}{P(A=0|L=l')}$.

Note:

- IP weighting: marginal probability of treatment A given covariate L.
- Standardization: probability of covariate L and conditional probability of outcome Y given A and L.
- In discrete case, these two are equivalent under positivity and conditional exchangeability and equal to $E(Y^a)$: $E(Y^a) = \sum_l E(Y|A = a, L = l)P(L = l) = E\left[\frac{I(A=a)Y}{f(A|L)}\right]$

If the number of individuals is multiplied times 2, then the number of deaths is also doubled.

Chapter 3: Observational Studies

Hesitation to empower observational associations with a causal interpretation is the lack of randomized treatment assignment.

Strategy: Analyze as if treatment is randomly assigned conditional on covariate L. $\rightarrow$ viewed as a conditionally randomized experiment if

1. consistency
2. exchangeability
3. positivity

holds.

From observational studies, causal inference requires data and **identifiability conditions** ((1)-(3) mentioned above).

Exchangeability: In case of measured covariates (L), it must be conditional exchangeable ($Y^a \perp\!\!\!\perp A|L$) within levels of L. This would not hold if there exist some unmeasured independent predictor U.

Positivity: For every combination of covariates, there must be a chance receiving any of the treatment. i.e. $P(A = a|L = l) > 0$ for all a with $P(L = l) > 0$.

Note: Positivity is only required when the variable (L) is required for exchangeability.

Consistency: $Y = Y^a = AY^{a=1} + (1 - A)Y^{a=0}$

1. precise definition of counterfactual outcome Y^a
2. linkage of counterfactual outcomes to the observed outcomes

For (1), sometimes this assumption does not hold since although intervention $a = 1$ is well defined, ($Y^{a=1}$ is also well defined), sometimes each experiment implements a different version of $a = 1$. The counterfactual outcome Y^a will differ in the setting of different versions of $a = 1$.

Note: In reality, these interventions are very hard to perfectly specify $\rightarrow$ the average causal effect varies across populations.

For (2), if it is an observational study, it is hard to have the data available. Then the well-defined counterfactual outcome is not necessarily equal to the individual's observed outcome.

Target trial framework: This involves:

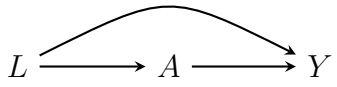
- Specifying a hypothetical randomized trial (the target trial) that you would ideally conduct to answer the causal question.

- Emulating this trial using the observational data.

This process forces investigators to explicitly define the key components of the study, such as eligibility criteria, interventions, outcomes, and follow-up. By doing so, it helps clarify the causal question and makes the necessary assumptions for consistency and exchangeability transparent. The framework helps prevent analyses that correspond to impossible or nonsensical interventions.

Chapter 7: Counfounding (With Part of Ch6: Graphical Representation of Causal Effects)

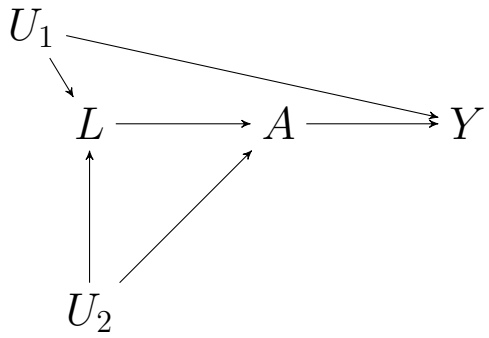
Counfounding: In obervational studies, treatment may be determined by many other factors.



From the image above, we can see the presence of the common cause L creates an additional source of association between the treatment A and the outcome Y, which we refer to as confounding for the effect of A on Y.

Note: Due to counfounding, associational risk ratio does not equal the causal risk ratio. i.e., association is not causation.

Collider: The definition of collider is path-specific: L is a collider on the path $A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$, but not on the path $A \leftarrow L \leftarrow U_1 \rightarrow Y$ in the following graph:



Note:

- Causal graphs theory shows that indeed conditioning on a collider opens the path, which was blocked when the collider was not conditioned on.
- There is an arrow $L \rightarrow A$. The presence of this arrow creates an open backdoor path $A \leftarrow L \leftarrow U_1 \rightarrow Y$ because U_1 is a common cause of A and Y, and so confounding exists. Conditioning on L would block that backdoor path but would simultaneously open a backdoor path on which L is a collider.

D-separation: We define a path to be either blocked or open according to the following graphical rules:

1. If there are no variables being conditioned on, a path is blocked if and only if two arrowheads on the path collide at some variable on the path
2. Any path that contains a non-collider that has been conditioned on is blocked
3. A collider that has been conditioned on does not block a path
4. A collider that has a descendant that has been conditioned on does not block a path

Backdoor Criterion: A set of covariates L satisfies the backdoor criterion if all backdoor paths between A and Y are blocked by conditioning on L and L contains no variables that are descendants

of treatment A.

Note:

- Under faithfulness and a further condition, conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ holds if and only if L satisfies the backdoor criterion.
- The two settings in which the backdoor criterion is satisfied are
  - No common causes of treatment and outcome: no backdoor paths that need to be blocked
  - No unmeasured confounding

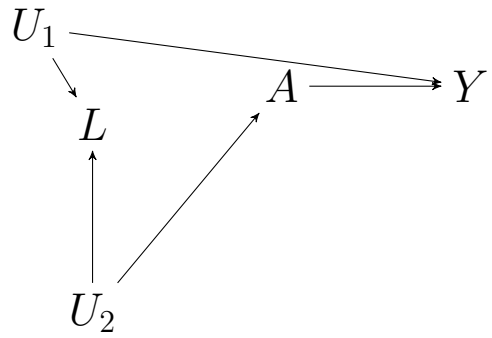
Two structural sources of lack of exchangeability:

- Confounding (any systematic bias that would be eliminated by randomized assignment of A): the presence of common causes of treatment and outcome—which creates an open backdoor path
- Selection bias: conditioning on a common effect—which may open a previously blocked backdoor path

The traditional approach to handling confounding is flawed because it relies on statistical associations rather than causal knowledge. This approach defines a confounder as a variable that is:

1. Associated with the treatment
2. Associated with the outcome (conditional on the treatment)
3. Not on the causal pathway between treatment and outcome ($A \to Y$)

For the graph below, where this traditional definition incorrectly identifies a non-confounder as a confounder. In this case, adjusting for the variable (L) would actually introduce a new bias, called selection bias, by opening a previously blocked backdoor path.

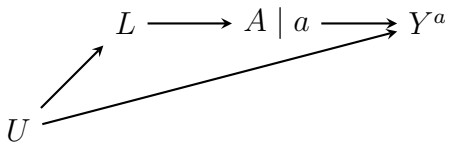


Structural Approach: A structural approach starts by explicitly identifying the sources of confounding—the common causes of treatment and outcome that, were they all measured, would be sufficient to adjust for confounding—and then identifies a sufficient set of adjustment variables.

Exchangeability is translated into graph language as the lack of open paths between the treatment A and outcom other than those originating from A, that would result in an association between A and Y.

Single-world Intervention Graphs (SWIGs): a graph that represents a counterfactual world created by a single intervention.

For example, the graph below represents represents a world in which all individuals have received an intervention that sets their treatment to the fixed value a:



Note: The key idea is that if the counterfactual outcome (Y^a) is d-separated from the natural treatment value (A) given a set of variables (L), then conditional exchangeability holds and confounding is eliminated.

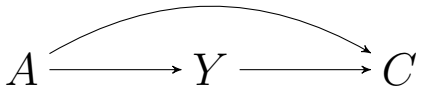
Confounding Adjustment: Methods that adjust for confounders L (a set of non-descendants of treatment A that includes enough variables to block all backdoor paths from A to Y) can be classified into two broad categories:

- G-methods: Standardization, IP weighting, and g-estimation: These methods (the "g" stands for "generalized") exploit conditional exchangeability given L to estimate the causal effect of A on Y in the entire population or in any subset of the population. (assume if backdoor paths involving the measured variables L did not exist)
- Conventional methods for stratification-based adjustment: Stratification (including restriction) and matching. These methods exploit conditional exchangeability given L to estimate the association between A and Y in subsets defined by L.

Note: All the above methods require conditional exchangeability given L. However, confounding can sometimes be handled by methods that do not require conditional exchangeability. Moreover, achieving conditional exchangeability may be an unrealistic goal in many observational studies but expert knowledge about the causal structure can be used to get as close as possible to that goal.

Chapter 8: Selection Bias

Only focus on selection bias under null: conditioning on common effects



treatment outcome common effect

For example: A study to estimate the effect of folic acid supplements given to pregnant women shortly after conception on the fetus's risk of developing a cardiac malformation during the first two months of pregnancy.

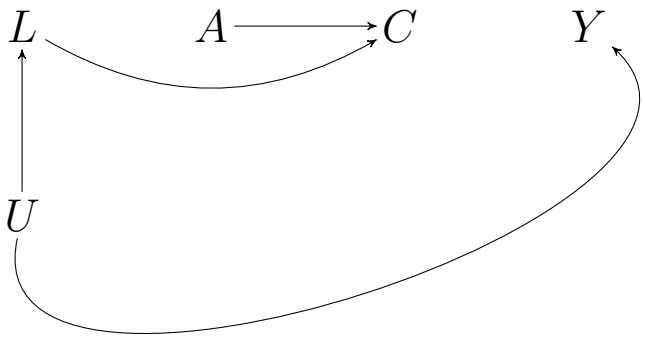
- A: folic acid supplement
- Y: cardiac malformation (1: yes, 0: no)
- C: death before birth

Source of association of treatment and outcome:

1. open path: $A \to Y$
2. open path: $A \to C \leftarrow Y$ if condition on C (association between A and Y)

Due to selection bias, association/risk ratio $\neq$ causal risk ratio $\implies$ association $\neq$ causation

Study of HIV infection



- A: antiretroviral treatment
- Y: 3-year risk of death
- U: high level of immunosuppression (1: yes, 0: no) (unmeasured variable)
- C: censored or not (1: yes, 0: no)
- L: symptoms, CD4 count etc

$A \to C$: treatment has side effects so patients want to dropout
C: analysis is restricted to those who are remain uncensored

$U \to Y$: individuals with $U = 1$ have a greater risk of death.

Based on d-separation, conditioning on C opens path, thus association flows from A to Y.

C: common effect of A and L (rather than Y)
Note: The bias is the result of selection on a common effect of two other variables in the diagram.

Selection Bias: Bias arises from conditioning on a common effect of two variables:

1. treatment or cause of treatment
2. outcome or cause of outcome.

Selection Bias Examples:

1. Differential loss to follow-up: informative censoring.
2. Missing Data Bias: restricting the analysis to individuals with complete data ($C = 0$) would result in bias (might be some reason why they are reluctant to provide information or miss study visits).
3. Healthy Worker Bias
4. Self-selection Bias (Volunteer Bias)
5. Selection affected by treatment received before study entry

Note:

1. Randomization protects against confounding, but not against selection bias when the selection occurs after randomization.
2. No bias arises in randomized experiments from selection into the study before treatment is assigned.

Comparison of Selection Bias and Confounding: Statisticians and econometricians often use the term "selection bias" to refer to both types of biases

- confounding: selection of individuals into analysis
- selection bias: selection of individuals into treatment

which would lead to lack of exchangeability.

Adjust of Selection Bias: IP weight could be used to adjust both confounding and selection bias.

- confounding: $W^A = 1/f(A|L)$
- selection bias: $W^C = 1/P(C = 0|A, L)$

$\implies$ construct pseudo-population of the same size as the original study population but in which nobody is lost to follow-up.

1. exchangeability

2. positivity
3. sufficiently well-defined intervention

Note: IP weighting appropriately adjusts for selection bias because this approach is not based on estimating effect measures conditional on the covariate L, but rather on estimating unconditional effect measures after reweighting the individuals according to their treatment and their values of L.

Chapter 12: IP Weighting and Marginal Structure Models

At start, the notations are defined as:

- L: covariates
- Y: outcome
- A: treatment

An Example

Aim: Estimate ATE of smoking cessation (treatment A) on weight gain (outcome Y).

Difficulty

ATE $= E[Y^{a=1}] - E[Y^{a=0}]$ ($a = 1$ is quit smoking, $a = 0$ is don't quit), which is **different from** what we have from the data, the associational difference: $E[Y|A = 1] - E[Y|A = 0]$

Why have this difference?

Age is a **confounder** of the effect $A \to Y$.

- Older people gain less weight than younger people no matter they quit smoking or not.
- Age need to adjust before analysis.

How do we adjust?

IP Weighting: Creates a pseudo-population where the contribution of each individual is re-weighted so that the distribution of L is independent of $A \Rightarrow$ Try to construct a dataset mimic randomized experiment.

- $A \perp L$
- $E_{ps}[Y|A = a] = \sum_l E[Y|A = a, L = l]P(L = l)$ (Pseudo-population mean equals to the standard mean in the actual population)

Note that these properties are still true if conditional exchangeability ($Y^a \perp A|L$) does not hold. If conditional exchangeability holds, then:

- mean of Y^a is the same in two populations.
- unconditional exchangeability (no confounding) holds in pseudo-population. ($Y^a \perp A$)
- $E(Y^a) = E_{ps}[Y|A = a]$
- **Association is causation in pseudo-population**

How can we estimate the IP weights?

Estimate weights via modeling. The individual-specific IP weights for treatment A are: $W^A = \frac{1}{f(A|L)}$. For non-parametric estimation of W^A, if we have high-dimensional L, obtaining a meaningful stratum-

specific estimate of W^A can be very difficult. For a parametric approach, we can fit a logistic regression:

$$\text{logit}(P(A=1|L)) = L^T\beta$$

After we have the estimation of IP weights, our next step is to compute $\hat{E}_{ps}[Y|A=1] - \hat{E}_{ps}[Y|A=0]$ in the pseudo-population.

The way to estimate $E_{ps}[Y|A=1] - E_{ps}[Y|A=0]$ is to fit a saturated linear mean model:

$$E[Y|A] = \theta_0 + \theta_1 A$$

by WLS (Weighted Least Squares) with weights $\hat{W}$:

$$\hat{W} = \begin{cases} \frac{1}{\hat{P}(A=1|L)} & \text{for quitters (A=1)} \\ \frac{1}{1-\hat{P}(A=1|L)} & \text{for non-quitters (A=0)} \end{cases}$$

The target is to minimize

$$\sum_i \hat{W}_i[Y_i - (\theta_0 + \theta_1 A_i)]^2$$

Note: This IP weighted mean for treatment a is equal to the counterfactual mean under positivity and exchangeability. i.e.,

$$E\left[\frac{I(A=a)Y}{f(A|L)}\right] = E[Y^a]$$

$E[Y^a]$ will be estimated by a consistent estimator by the Horvitz-Thompson approach [HT52]:

$$\hat{E}\left[\frac{I(A=a)Y}{f(A|L)}\right]$$

We need to assume $f(A|L)$ is known. However, the **Hájek estimator** [Háj71] is preferred [LD04], which is defined as:

$$\frac{\hat{E}\left[\frac{I(A=a)Y}{f(A|L)}\right]}{\hat{E}\left[\frac{I(A=a)}{f(A|L)}\right]}$$

It is preferred since it can be guaranteed to lie between 0 and 1 for dichotomous Y, even if $f(A|L)$ is unknown and replaced by $\hat{f}(A|L)$. If positivity is not held, the difference between Hájek estimators does not have a causal interpretation. Also, the variance is needed to construct a 95% C.I. There are normally three ways to do it:

1. Nonparametric bootstrapping (time-consuming, computation cost)
2. Derive variance estimator under statistical theory (not generally available)

3. Robust Variance Estimator: Standard option in most statistical software packages (Used for independent working correlation in GEE). Downside: conservative.

For example, if weight = 2, it means we form 2 copies of the original study population to produce the pseudo-population (one for treated and the other for untreated).

Better approach for obtaining weights?

Stabilized weights which is defined as:

$$SW^A = \frac{f(A)}{f(A|L)}$$

The properties are as follows:

- Can have narrower 95% C.I.
- This superiority can only occur when the model is not saturated.
- In data analysis, one should check whether SW has a mean of 1. If not, model misspecification or possible violations, or near violations, of positivity may be the issue.

Review: If the positivity assumption does not meet, it will signal these two problems:

1. Lack of Overlap: there's little to no empirical overlap between the treated and control groups, making causal comparison tenuous.
2. Model Misspecification: A misspecified propensity score model might generate extreme predicted probabilities.

How can we diagnose whether the IP weight estimation is good?

[AS15] suggest a framework for how we can use the IP weights and evaluate whether we have a balanced dataset or not and has not appear in the previous works.

1. Identify Confounders: draw DAGs, review literature to ensure conditional exchangeability
2. Specify Propensity Score Model: prior evidence show that it is preferable to include either prognostically important variables or confounding covariates in the model
3. Calculate Weights & Diagnose Weights: check the positivity assumption
    - The mean of stabilized weights should be close to 1.
    - Identify the maximum and minimum values of the weights.
4. Access Covariate Balance:
    (a) Quantitative: calculate standardized mean difference (SMD), which is defined as

$$\text{SMD} = \frac{\bar{x}_{treated} - \bar{x}_{control}}{\sqrt{\frac{s^2_{treated} + s^2_{control}}{2}}}$$

, where $\bar{x}_{weight} = \frac{\sum w_i x_i}{\sum w_i}$ and $s^2_{weight} = \frac{\sum w_i}{(\sum w_i)^2 - \sum w_i^2} \sum w_i (x_i - \bar{x}_{weight})^2$ for all covariates and interactions (note that SMDs is not influenced by sample size compare to t-test) and present it in a table such as in [Table 1] or absolute standardize difference graph as in [Figure 1].

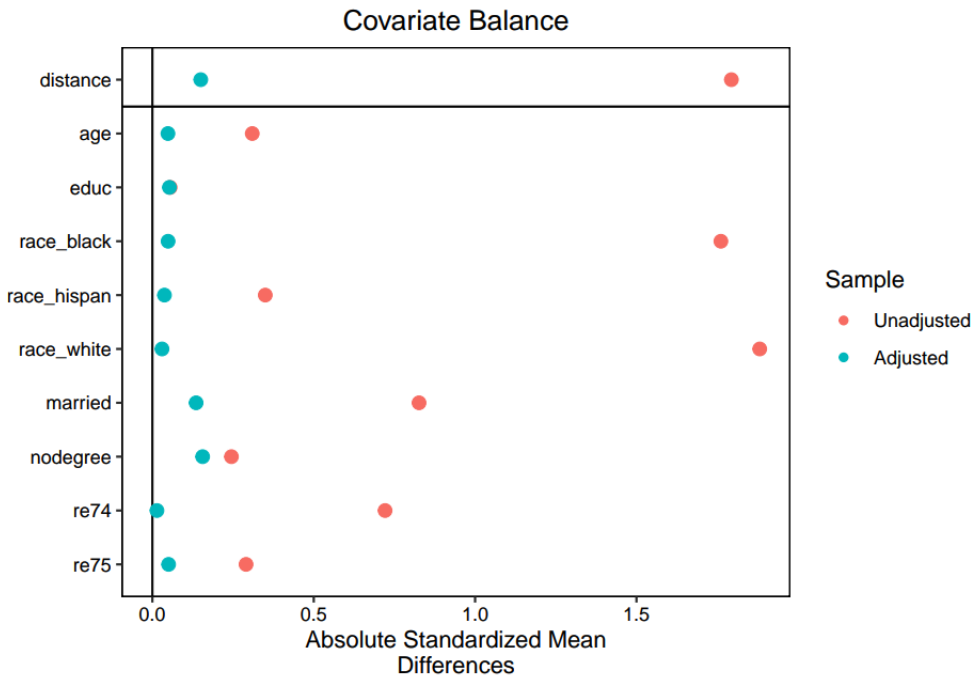
(b) Qualitative: graph boxplots and eCDF plots to inspect the distributional balance. For eCDFs, we can use Kolmogorov-Smirnov (K-S) test statistic to have a formal comparison of distributions. Example graphs are in [Figure 2].

5. Evaluate Balance and Iterate
6. Estimate Effect
7. Estimate Variance

Note: For the K-S test in detecting distributional difference, we use only the descriptive statistic rather than the p-value from the formal hypothesis testing. The reasons are:

- p-values are heavily influenced by sample size
- p-value answers questions on the hypothetical population, but in here we only cares about the specific dataset
- K-S test statistic can be computed on the weighted dataset, but p-value of this test assume you have an unweighted dataset

Figure 1: Absolute Standardized Mean Differences Graph using [LaL86].



Marginal Structural Model (MSM)

- **Model:** $E(Y^a) = \beta_0 + \beta_1 a$
  - We can have ATE $= E(Y^{a=1}) - E(Y^{a=0}) = \beta_1$

Figure 2: Box Plots and CDF Graphs to Show Pre- and Post Weighting [AS15]

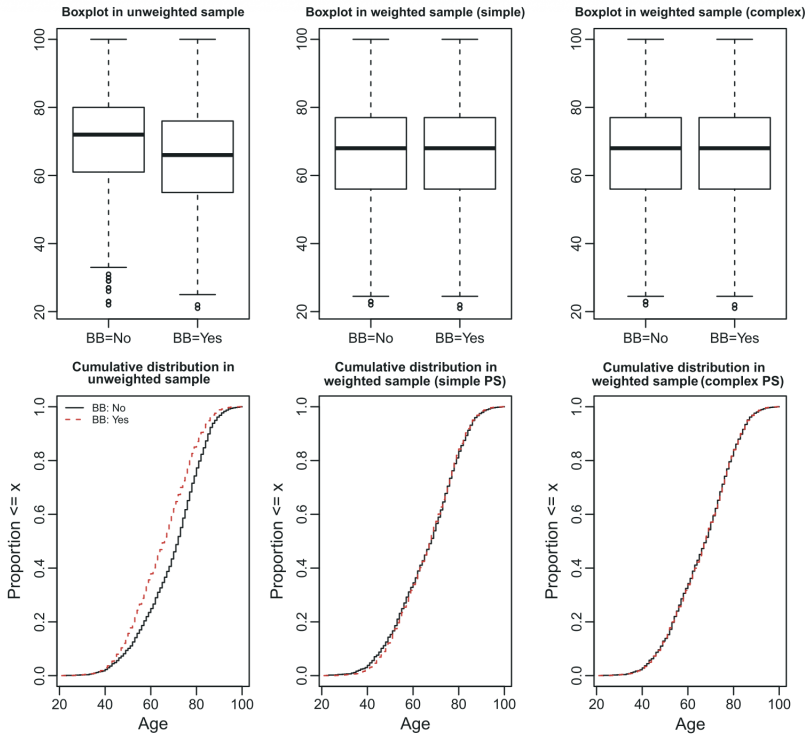


Table 1: Covariate Balance Before and After Inverse Probability of Treatment Weighting (Hypothetical Data)

Covariate	**Unweighted**			**Weighted**		
	Treated	**Control**	**SMD**	**Treated**	**Control**	**SMD**
Age (years), mean (SD)	55.2 (8.1)	51.5 (9.2)	0.42	53.1 (8.8)	53.0 (8.9)	0.01
Female, %	65.0	50.0	0.30	58.1	57.9	0.00
Prior MI, %	20.0	10.0	0.28	14.5	14.6	−0.00
BMI, mean (SD)	28.1 (4.5)	26.5 (4.8)	0.34	27.2 (4.6)	27.3 (4.7)	−0.02
Higher-Order Terms						
Age^2	3112.4	2736.5	0.39	2900.1	2888.7	0.01
Age × BMI	1551.1	1364.8	0.35	1449.6	1452.3	−0.00

- In **IP weighting**, we are fitting a weighted least square of $E(Y|A) = \theta_0 + \theta_1 A$ to the pseudo-population
  - Under assumption, association is causation in the pseudo-population
  - $\hat{\theta}_1$ has the same meaning as β_1
  - Having a consistent $\hat{\theta}_1$ of the association parameter in the pseudo-population is also a consistent estimator of causal effect β_1 in the population.
- Note that this is a saturated model.
  - For continuous treatment (or at least more than two treatment values), we need to fit a

non-saturated marginal structural mean model

  - Downside of IP weighting approach: for constructing pseudo-population, we need to estimate $SW^A = f(A)/f(A|L)$. For dichotomous treatment A, $f(A|L)$ can be estimated by logistic regression. However, for continuous treatment A, estimating pdf $f(A|L)$ is hard. Although you can assume the distribution of $f(A|L)$, the effect estimate will be sensitive. But still in the [RHB00], they suggest to assume to be normal distribution.

- MSM model on dichotomous outcome

$$\log\left(\frac{P(D^a=1)}{P(D^a=0)}\right) = \alpha_0 + \alpha_1 a, \quad a \in \{0,1\}$$

where $a = 1$ be getting heart death and $a = 0$ be no heart death. $\exp(\alpha_1)$: causal odd ratio of death for quitting smoke vs not quitting smoke.

- **Effect modification**
  - May want to add covariates V in a marginal structural model to access effect modification.

$$E(Y^a|V) = \beta_0 + \beta_1 a + \beta_2 V a + \beta_3 V$$

  - IP weighting based on $SW^A(V) = \frac{f(A|V)}{f(A|L)}$ generally gives smaller C.I. ($f(A|V)$ is estimated by logistic model by adding V as covariate)
  - Should choose V only for the investigator's **substantive interest**
    - If the investigator believe V is a "effect modifier" and has greater substantive interest in the causal effect of treatment within levels of the covariate V than in the population.

- **Time-Dependent Treatment MSM**
  - $\overline{A} = (A_1, A_2, ..., A_K)$: observed treatment history vector ; $\overline{a}$: potential treatment history vector ; $\overline{L}$: covariates history matrix
  - Assume no unmeasued confounders exists.
  - For time-dependent treatments, the counterfactual is $Y_{\overline{a}}$ (the outcome under treatment history $\overline{a}$). An MSM for this might be:

$$\text{logit}[P(Y_{\overline{a}}=1)] = \beta_0 + \beta_1 cum(\overline{a})$$

where $cum(\overline{a}) = \sum_{k=0}^{K} a_k$ is the cumulative treatment. This is an *unsaturated* (parsimonious) model assuming the causal effect depends only on the total cumulative treatment.

  - For the association model, we can use the similar formula:

$$\text{logit}[P(Y=1|\overline{A}=\overline{a})] = \beta_0' + \beta_1' cum(\overline{a})$$

with the stabilized weights to be:

$$sw_i = \prod_{k=0}^{K} \frac{P(A_k = a_{ki}|\overline{A}_{k-1} = \overline{a}_{(k-1)i})}{P(A_k = a_{ki}|\overline{A}_{k-1} = \overline{a}_{(k-1)i}, \overline{L}_k = \overline{l}_{ki})}$$

  - After adjusting for treatment A (treatment is now not confounded), $\beta_1 = \beta_1'$

- **Censoring and Missing Data**
  - No new idea is required since we can conceptually treat censoring as just another time-varying treatment. We can calculate the **Inverse-Probability-of-Censoring Weight (IPCW)**, $sw_i^\dagger$. Since the outcome Y is unobserved unless the subject does not drop out, our weighted model fit is restricted to subjects who were not censored:

$$sw_i^\dagger = \prod_{k=0}^{K+1} \frac{P(C_k = 0 | \overline{C}_{k-1} = 0, \overline{A}_{k-1})}{P(C_k = 0 | \overline{C}_{k-1} = 0, \overline{A}_{k-1}, \overline{L}_k)}.$$

    And the final weight for each subject is the product of the two: $sw_{final} = sw_i \times sw_i^\dagger$. The final MSM is then fit using this combined weight.

Chapter 13: Standardization and the Parametric G-Formula

For estimating the average treatment effect (ATE) $\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}]$, there are two ways to calculate:

1. IP weighting (Chapter 12)
2. Standardization (This chapter)

Under exchageability and positivity conditional on the variables L, we have the property of association equals to causation. For discrete treatment L, we can calculate the standardized mean in the uncensored who received treatment a:

$$\mathbb{E}[Y^{a,c=0}] = \sum_l \mathbb{E}[Y|A=a, C=0, L=l]\mathbb{P}[L=l]$$

and for continuous L, we can calculate the standardized mean in the uncensored who received treatment a:

$$\mathbb{E}[Y^{a,c=0}] = \int \mathbb{E}[Y|A=a, C=0, L=l]dF_L(l)$$

Estimating $\mathbb{E}[Y|A=a, C=0, L=l]$ would not be hard by non-parametric approach if L is discrete and the outcomes of L are not too many. If we have a high-dimensional data with many confounders, some of them with multiple levels, the non-parametric estimation would be hard and not accurate.

Instead, we can model $\mathbb{E}[Y|A=a, C=0, L=l]$ with a parametric model such as linear regression:

$$Y = \beta_0 + \beta_1 A + \beta_2 L + \epsilon$$

and have the the estimate $\hat{\mathbb{E}}[Y|A=a, C=0, L=l]$ for each level of L.

However, we don't have $\mathbb{P}[L=l]$. Fortunately, we have instead compute the average:

$$\hat{\mathbb{E}}[Y^{a,c=0}] = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbb{E}}[Y|A=a, C=0, L_i]$$

g-formula: standardization (plug-in g-formula), parametric g-formula

IP Weighting or Standardization?

A question would directly pop out that since IP weighting and standardization are two ways to estimate the ATE and has similar result, which one would we choose? First we need to understand the difference between IP weighting and standardization:

- IP weighting models $\mathbb{P}[A=a, C=0|L]$, which we can estimate by fitting logistic regressions to get the estimates of $\mathbb{P}[A=a|L]$ and $\mathbb{P}[C=0|A=a, L]$.
- Standardization models $\mathbb{E}[Y|A=a, C=0, L]$, which we can estimate by fitting the linear regression.

All practical models contain some degree of non-ignorable misspecification, which inevitably introduces bias. However, the bias resulting from misspecification in IP weighting and standardization is unlikely to be of the same magnitude or direction. Therefore, we should **compare the estimates from both IP weighting and standardization simultaneously** to assess the robustness of the findings.

- Big difference between those results will give us alert of the presense of model misspecification.
- Small difference between those results cannot garantee anything, but will be assuring us of the robustness. (Low probability of model misspecification at the same time for both IP weighting and standardization.)

Doubly Robust Estimator: A method that only requires a correct model for using IP weighting (modeling treatment A) or standardization (modeling outcome Y). Under the usual identifiability assumptions, a doubly robust estimator consistently estimates the causal effect if at least one of the two models is correct (and one need not know which of the two models is correct). The whole procedure would be as follows:

1. Estimate the IP weight $W^A = 1/f(A|L)$
2. Fit an outcome regression model with a canonical link for $\mathbb{E}[Y|A, L, R]$ that adds the covariate R, where $R = W^A$ if $A = 1$ and $R = -W^A$ if $A = 0$
3. Calculate the average causal effect by the difference of the two estimators

Augmented IP Weighted Estimator: For IP weighting method, we have the estimator in the form of $\mathbb{E}\left[\frac{AY}{\pi(L)}\right]$, where $\pi(L) = \mathbb{P}[A = 1|L]$. Also, For the standardization method, we have the estimator in the form of $\mathbb{E}[b(L)]$, where $b(L) = \mathbb{E}[Y|A = 1, L]$. The *augmented IP weighted estimator* is the average of the two estimators:

$$\begin{aligned}\hat{\mathbb{E}}[Y^{a=1}]_{DR} &= \frac{1}{n}\sum_{i=1}^{n}\left[\hat{b}(L_i) + \frac{A_i}{\hat{\pi}(L_i)}\left(Y_i - \hat{b}(L_i)\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i Y_i}{\hat{\pi}(L_i)} - \left(\frac{A_i}{\hat{\pi}(L_i)} - 1\right)\hat{b}(L_i)\right]\end{aligned}$$

Under exchageability and positivity,

$$\hat{\mathbb{E}}[Y^{a=1}]_{DR} - \mathbb{E}[Y^{a=1}] \xrightarrow{p} E\left[\pi(L)\left(\frac{1}{\pi(L)} - \frac{1}{\pi^*(L)}\right)(b(L) - b^*(L))\right],$$

where $\pi^*(L)$ and $b^*(L)$ are the probability limits of $\hat{\pi}(L)$ and $\hat{b}(L)$. It follows that doubly robust estimator is (asymptotically) unbiased when **either** the parametric **outcome model** is correct [so $b^*(L) = b(L)$] **or** the parametric **treatment model** is correct [so $\pi^*(L) = \pi(L)$]. Furthermore, we do not need to know which one of the two models is correct. Note that the bias of $\pi(L)$ and $b(L)$ can be small by using machine learning estimators.

Cautious for the estimating results

Causal analysis using observational data is best conducted by explicitly modeling a hypothetical target trial. Even when limiting the inference to the population under study (not transporting the results), the validity of the causal estimate requires several conditions:

- Identify the conditions of exchangeability (check confounding, selection bias), positivity (structural positivity) and consistency (check multiple version of treatments).
- Identify all the variables used in the analysis should be correctly measured (check measurement error, confounders).
- Check model misspecification, this is similar effect as measurement error in the confounders.

Structural Positivity: If the analytical model is perfectly specified, parametric or double robust methods can smooth over strata with structural zeros ("extrapolation"), the lack of positivity part in the dataset can be ignored. However, this introduces bias and reduces the standard error. Note that we use model to do extrapolation not because we lack enough data, but because we want to estimate a quantity that cannot be identified even with an infinite amount of data.

Data Analysis

The dataset we have is from National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS). Our goal is to estimate the average causal effect of smoking cessation (the treatment) on weight gain (the outcome).

The variables are as follows:

- **qsmk:** quit smoking between 1st questionnaire and 1982, 1:yes, 0:no
- **sex:** 0: male 1: female
- **race:** 0: white 1: black or other in 1971
- **age:** age in 1971
- **education:** amount of education by 1971: 1: 8th grade or less, 2: hs dropout, 3: hs, 4:college dropout, 5:college or more
- **smokeintensity:** number of cigarettes smoked per day in 1971
- **smokeyrs:** years of smoking
- **exercise:** in recreation, how much exercise? in 1971, 0:much exercise, 1:moderate exercise, 2:little or no exercise
- **active:** in your usual day, how active are you? in 1971, 0:very active, 1:moderately active, 2:inactive
- **wt71:** weight in kilograms in 1971

```
nhefs <- read.csv("nhefs.csv")
nhefs$cens <- ifelse(is.na(nhefs$wt82), 1, 0)

fit <- glm(wt82_71 ~ qsmk + sex + race + age + I(age*age) + as.factor(education)
           + smokeintensity + I(smokeintensity*smokeintensity) + smokeyrs
           + I(smokeyrs*smokeyrs) + as.factor(exercise) + as.factor(active)
           + wt71 + I(wt71*wt71) + qsmk*smokeintensity, data=nhefs)
summary(fit)
```

```
##
## Call:
## glm(formula = wt82_71 ~ qsmk + sex + race + age + I(age * age) +
##     as.factor(education) + smokeintensity + I(smokeintensity *
##     smokeintensity) + smokeyrs + I(smokeyrs * smokeyrs) + as.factor(exercise) +
##     as.factor(active) + wt71 + I(wt71 * wt71) + qsmk * smokeintensity,
##     data = nhefs)
```

```
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -1.5881657  4.3130359  -0.368 0.712756
## qsmk                              2.5595941  0.8091486   3.163 0.001590 **
## sex                              -1.4302717  0.4689576  -3.050 0.002328 **
## race                              0.5601096  0.5818888   0.963 0.335913
## age                               0.3596353  0.1633188   2.202 0.027809 *
## I(age * age)                     -0.0061010  0.0017261  -3.534 0.000421 ***
## as.factor(education)2             0.7904440  0.6070005   1.302 0.193038
## as.factor(education)3             0.5563124  0.5561016   1.000 0.317284
## as.factor(education)4             1.4915695  0.8322704   1.792 0.073301 .
## as.factor(education)5            -0.1949770  0.7413692  -0.263 0.792589
## smokeintensity                    0.0491365  0.0517254   0.950 0.342287
## I(smokeintensity * smokeintensity) -0.0009907  0.0009380  -1.056 0.291097
## smokeyrs                          0.1343686  0.0917122   1.465 0.143094
## I(smokeyrs * smokeyrs)           -0.0018664  0.0015437  -1.209 0.226830
## as.factor(exercise)1              0.2959754  0.5351533   0.553 0.580298
## as.factor(exercise)2              0.3539128  0.5588587   0.633 0.526646
## as.factor(active)1               -0.9475695  0.4099344  -2.312 0.020935 *
## as.factor(active)2               -0.2613779  0.6845577  -0.382 0.702647
## wt71                              0.0455018  0.0833709   0.546 0.585299
## I(wt71 * wt71)                   -0.0009653  0.0005247  -1.840 0.066001 .
## qsmk:smokeintensity               0.0466628  0.0351448   1.328 0.184463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 53.5683)
##
##     Null deviance: 97176  on 1565  degrees of freedom
## Residual deviance: 82763  on 1545  degrees of freedom
##   (63 observations deleted due to missingness)
## AIC: 10701
##
## Number of Fisher Scoring iterations: 2
```

```r
# create a dataset with 3 copies of each subject
nhefs$interv <- -1 # 1st copy: equal to original one

interv0 <- nhefs # 2nd copy: treatment set to 0, outcome to missing
interv0$interv <- 0
interv0$qsmk <- 0
interv0$wt82_71 <- NA

interv1 <- nhefs # 3rd copy: treatment set to 1, outcome to missing
interv1$interv <- 1
interv1$qsmk <- 1
interv1$wt82_71 <- NA
```

```r
onesample <- rbind(nhefs, interv0, interv1) # combining datasets

# linear model to estimate mean outcome conditional on treatment and confounders
# parameters are estimated using original observations only (nhefs)
# parameter estimates are used to predict mean outcome for observations with
# treatment set to 0 (interv=0) and to 1 (interv=1)

std <- glm(wt82_71 ~ qsmk + sex + race + age + I(age*age)
           + as.factor(education) + smokeintensity
           + I(smokeintensity*smokeintensity) + smokeyrs
           + I(smokeyrs*smokeyrs) + as.factor(exercise)
           + as.factor(active) + wt71 + I(wt71*wt71) + I(qsmk*smokeintensity),
           data=onesample)
onesample$predicted_meanY <- predict(std, onesample)

mean(onesample[which(onesample$interv==0),]$predicted_meanY)
```

```
## [1] 1.660267
```

```r
mean(onesample[which(onesample$interv==1),]$predicted_meanY)
```

```
## [1] 5.178841
```

$\hat{E}[Y^{a=1,c=0}]$ is 5.178841 and $\hat{E}[Y^{a=1,c=0}]$ is 1.660267. Thus, the causal effect of 3.5kg.

Simulation of Augmented IP Weighting Estimator

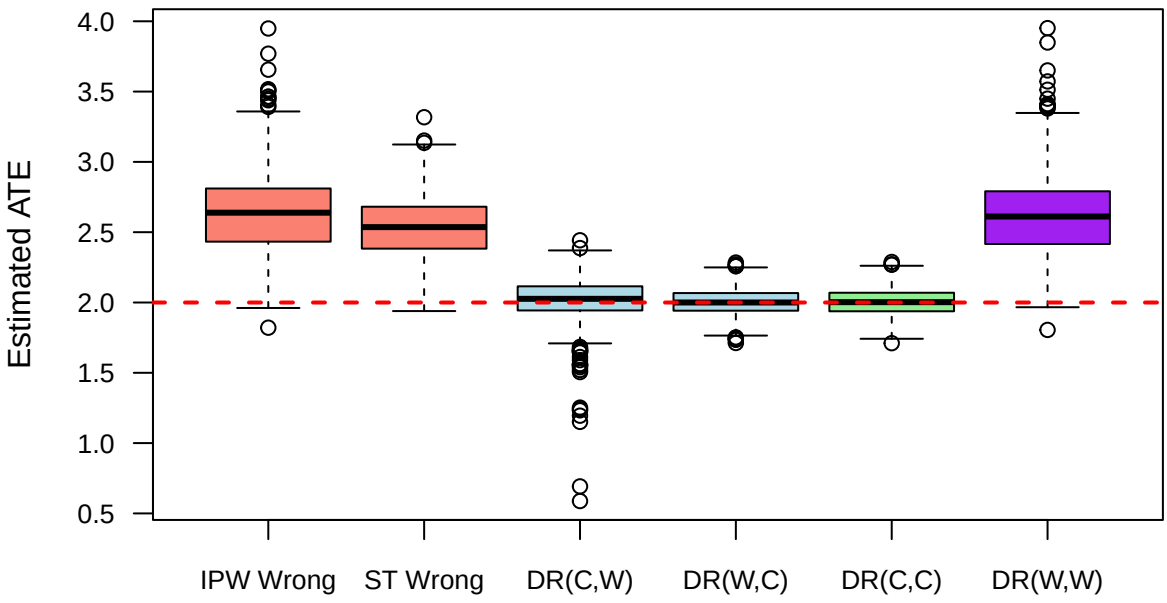
```r
set.seed(42)
n_sims <- 1000
n_sample <- 500
true_ate <- 2

results <- matrix(NA, nrow = n_sims, ncol = 6)
colnames(results) <- c("IPW Wrong", "ST Wrong", "DR(C,W)", "DR(W,C)", "DR Both Cor", "DR Bot

calc_dr <- function(Y, A_obs, pi_hat, b_a1_hat, b_a0_hat) {
  dr1 <- mean( (A_obs * Y) / pi_hat - ((A_obs - pi_hat) / pi_hat) * b_a1_hat)

  dr0 <- mean( ((1 - A_obs) * Y) / (1 - pi_hat) + ((A_obs - pi_hat) / (1 - pi_hat)) * b_a0_h

  return(dr1 - dr0)
}

for (i in 1:n_sims) {

  # === A. Data Generation (The "Truth") ===
  # Confounder X is N(0,1)
```

```r
X <- rnorm(n_sample, mean = 0, sd = 1)

# True $\pi$: Depends on X AND X^2 (Non-linear)
# logit(P(T=1)) = -0.5 + 0.5*X + 0.2*X^2
z <- -0.5 + 0.5 * X + 0.2 * (X^2)
true_pi <- 1 / (1 + exp(-z))
A_obs <- rbinom(n_sample, 1, true_pi)

# True $b(L)$: Depends on A, X, AND X^2
# Y = 2*A + 1 + 2*X + 1.5*X^2 + error
# The coefficient of T is 2, so True ATE = 2
Y <- 2 * A_obs + 1 + 2 * X + 1.5 * (X^2) + rnorm(n_sample)

# Create a data frame
df <- data.frame(Y = Y, T = A_obs, X = X, X_sq = X^2)


# === B. Model Fitting ===

# 1. \pi Estimation (IP Weighting)
# Wrong Model: Only sees X (misses X^2)
fit_ps_wrong <- glm(T ~ X, data = df, family = binomial)
ps_wrong <- predict(fit_ps_wrong, type = "response")

# Correct Model: Sees X and X^2
fit_ps_correct <- glm(T ~ X + X_sq, data = df, family = binomial)
ps_correct <- predict(fit_ps_correct, type = "response")


# 2. b(L) estimation (Standardization)
# --------------------------------
# Wrong Model: Only sees X (misses X^2)
fit_or_wrong <- lm(Y ~ T + X, data = df)
mu1_wrong <- predict(fit_or_wrong, newdata = transform(df, T = 1))
mu0_wrong <- predict(fit_or_wrong, newdata = transform(df, T = 0))

# Correct Model: Sees X and X^2
fit_or_correct <- lm(Y ~ T + X + X_sq, data = df)
mu1_correct <- predict(fit_or_correct, newdata = transform(df, T = 1))
mu0_correct <- predict(fit_or_correct, newdata = transform(df, T = 0))

# Scenario 1: IPW (Wrong Model)
# Formula: mean( TY/e - (1-T)Y/(1-e) )
ipw_est <- mean( (df$T * df$Y) / ps_wrong - ((1 - df$T) * df$Y) / (1 - ps_wrong) )
results[i, 1] <- ipw_est

# Scenario 2: Standardization (ST) (Wrong Model)
```

```r
  # Formula: mean( mu1 - mu0 )
  or_est <- mean(mu1_wrong - mu0_wrong)
  results[i, 2] <- or_est

  # Scenario 3: DR (IP Correct, ST Wrong)
  results[i, 3] <- calc_dr(df$Y, df$T, ps_correct, mu1_wrong, mu0_wrong)

  # Scenario 4: DR (IP Wrong, ST Correct)
  results[i, 4] <- calc_dr(df$Y, df$T, ps_wrong, mu1_correct, mu0_correct)

  # Scenario 5: DR (Both Correct)
  results[i, 5] <- calc_dr(df$Y, df$T, ps_correct, mu1_correct, mu0_correct)

  # Senario 6: DR (Both Wrong)
  results[i, 6] <- calc_dr(df$Y, df$T, ps_wrong, mu1_wrong, mu0_wrong)
}

# Calculate Bias and MSE
summary_table <- data.frame(
  Estimator = colnames(results),
  Mean_Estimate = colMeans(results),
  Bias = colMeans(results) - true_ate,
  MSE = colMeans((results - true_ate)^2)
)

rownames(summary_table) <- c()

print(summary_table, digits = 4)
```

```
##       Estimator Mean_Estimate     Bias      MSE
## 1     IPW Wrong         2.640 0.639562 0.493290
## 2      ST Wrong         2.535 0.535135 0.332257
## 3       DR(C,W)         2.016 0.016078 0.024198
## 4       DR(W,C)         2.004 0.003938 0.009011
## 5   DR Both Cor         2.003 0.003109 0.009054
## 6 DR Both Wrong         2.616 0.616402 0.461804
```

```r
boxplot(results,
        main = "Performance: IPW vs ST vs DR under Misspecification",
        ylab = "Estimated ATE",
        col = c("salmon", "salmon", "lightblue", "lightblue", "lightgreen", "purple"),
        names = c("IPW Wrong", "ST Wrong", "DR(C,W)", "DR(W,C)", "DR(C,C)", "DR(W,W)"),
        las = 1,
        cex.axis = 0.8) # Horizontal labels, slightly smaller text
abline(h = true_ate, col = "red", lwd = 2, lty = 2)
```


Performance: IPW vs ST vs DR under Misspecification
Estimated ATE
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
IPW Wrong
ST Wrong
DR(C,W)
DR(W,C)
DR(C,C)
DR(W,W)

References

[HT52] 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.

[Háj71] J. Hájek. "Comment on "An essay on the logical foundations of survey sampling" by D. Basu". In: *Foundations of Statistical Inference*. Ed. by V. P. Godambe and D. A. Sprott. Toronto, Ontario, Canada: Holt, Rinehart and Winston, 1971, p. 236.

[LaL86] Robert J LaLonde. "Evaluating the econometric evaluations of training programs with experimental data". In: *The American economic review* (1986), pp. 604–620.

[RHB00] James M Robins, Miguel Angel Hernan, and Babette Brumback. *Marginal structural models and causal inference in epidemiology*. 2000.

[LD04] Jared K Lunceford and Marie Davidian. "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study". In: *Statistics in medicine* 23.19 (2004), pp. 2937–2960.

[AS15] Peter C Austin and Elizabeth A Stuart. "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies". In: *Statistics in medicine* 34.28 (2015), pp. 3661–3679.