Lecture 4: Observational Studies

Tianpei Xie

Sep. 21st., 2022

Contents

1 From Randomized Experiments to Observational Studies

- Randomized experiments may be ***unavailable***:
  - Randomized experiments are ***expensive***;
  - It may be ***unethical*** to conduct randomized experiments (e.g. effect of smoking)
  - Many people may refuse to participate in randomized experiments (due to strong regulation);
  - Randomized experiments take a long time; sometimes it would be too late for result to come out.
- An observational study is an empiric investigation of effects caused by treatments when randomized experimentation is ***unethical*** or ***infeasible***. Like randomized experiments, the data are collected on a common set of variables at planned times. Outcomes are carefully measured with protocals. On the other hand, the regulation is much weaker since there is no intervention in an observational study. It is usually available for larger group of people.

  There are ***active*** data collection process and ***passive*** data collection process. The latter refer to the process when the researcher has little control on the data collection process.
- In ***observational studies***, the treatment is almost always a function of some covariate(s), i.e. there always exist *confounders*.
- Due to the existence of confounder, the ***comparability*** ***and*** ***covariate balance*** between treatment and control groups do **not** necessarily **hold**.

  A **central concern** in observational studies: people who look comparable in the **observed** data may not actually be comparable; *they may **differ** in ways we did **not observe**.*
- Many observational studies on causal inference focus on **design** of ***treatment assignment mechanism*** so that the ***unconfoundness*** *assumption* holds.
- Since the treatment T may depend on some covariate, there might exist **back-door path** from outcome Y to the treatment T. In order to estimate the causal effect of T on Y, we need to ***control the confouding bias***.

2 Controlling confounding bias

- The key characteristic of ***covariates*** is that they are a ***priori*** *known* to be **unaffected** by the treatment assignment. This knowledge often comes from the fact that they are *permanent characteristics* of units, or that they took on their values *prior to the treatment* being assigned, as reflected in the label "***pre-treatment***" variables.
- ***Confounders*** is a set of covariates that affect **both** treatment assignments and outcomes.
- **Definition** (***Confounding***) [Peters et al., 2017]
  Consider an SCM $\mathfrak{C}$ over nodes $\mathcal{V}$ with a directed path from $X \rightarrow Y$, $X, Y \in \mathcal{V}$. The causal effect from X to Y is called ***confounded*** if

$$p^{\mathfrak{C}:do(X=x)}(y) \neq p^{\mathfrak{C}}(y|x). \tag{1}$$

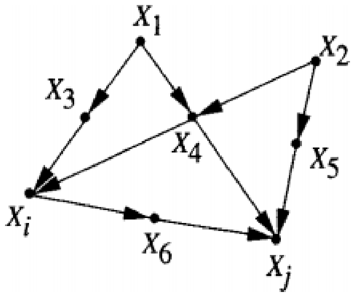


Figure 3.4 A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ (or $\{X_4, X_5\}$) yields a consistent estimate of $P(x_j \mid \hat{x}_i)$. Adjusting for $\{X_4\}$ or $\{X_6\}$ would yield a biased estimate.

Figure 1: The back-door adjustment [Pearl, 2000]

Otherwise, the causal effect is called ***unconfounded***.

- In order to account for the influence of confounder, we should **partition** the population into groups that are ***homogeneous*** relative to *confounder* Z, assessing the effect of X on Y in each homogeneous group, and then averaging the results. This process is called ***covariate adjustment***. This is the idea behind the ***Adjustment Formula*** [Imbens and Rubin, 2015].
- ***Confounder*** should be distinguished from the ***collider***: confounders ***need*** to be **controlled for** when estimating causal associations, while collider should be **avoided** during the conditioning.

This section discuss the process of **choosing** adjustment set using causal structure.

2.1 The Back-door Adjustment

Assume we are given a ***causal diagram*** $\mathcal{G}$, together with *nonexperimental* data on a subset V of observed variables in $\mathcal{G}$, and suppose we wish to estimate what effect the interventions $do(X = x)$ would have on a set of response variables Y, where X and Y are two subsets of V. In other words, we seek to estimate $P(y \mid do(x))$ from a sample estimate of $P(v)$, given the assumptions encoded in $\mathcal{G}$.

The ***back-door adjustment*** or *back-door criterion* [Pearl, 2000] is a simple graphical test that can be applied directly to the causal diagram in order to test if a set $Z \subseteq V$ of variables is sufficient for identifying $P(y \mid do(x))$.

- **Definition** (***Back-Door***) [Pearl, 2000]
  A set of variables Z satisfies the ***back-door criterion*** relative to an ***ordered*** pair of variables $(X_i \rightarrow X_j)$ in a DAG $\mathcal{G}$ if:

  1. ***no*** node in Z is a ***descendant*** of X_i; and
  2. Z ***blocks every path*** between X_i and X_j that contains an arrow ***into*** X_i.

  Similarly, if X and Y are two **disjoint** subsets of nodes in $\mathcal{G}$, then Z is said to satisfy ***the back-door criterion*** relative to (X, Y) if it satisfies the criterion relative to *any pair* (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$.

- Satisfying the back-door criterion makes Z a ***sufficient adjustment set***. The main insight of the graphical approach to covariate adjustment is that the adjustment set must **block all**

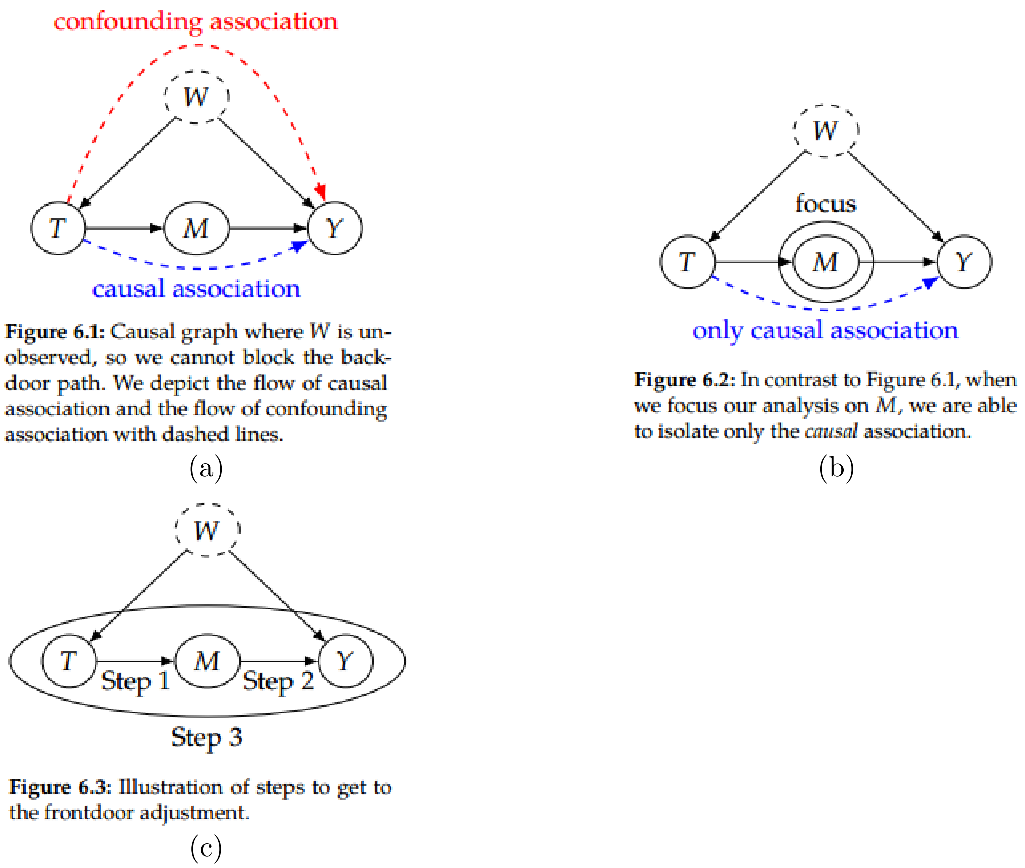


Figure 6.1: Causal graph where W is unobserved, so we cannot block the back-door path. We depict the flow of causal association and the flow of confounding association with dashed lines.

(a)

Figure 6.2: In contrast to Figure 6.1, when we focus our analysis on M, we are able to isolate only the *causal* association.

(b)

Figure 6.3: Illustration of steps to get to the frontdoor adjustment.

(c)

Figure 2: The front-door adjustment when the covariates in back-door path is not observed.

noncausal **paths without blocking** any ***causal*** paths between X and Y.

- **Theorem 2.1** *(**Back-Door Adjustment**) [Pearl, 2000, Neal, 2020]*
  *If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is **identifiable** and is given by the formula*

$$P(y \,|\, do(x)) = \sum_{z} P(y \,|\, x, z) P(z) \tag{2}$$

  To see why this works we need to know that $P(z|do(x)) = P(z)$ since by back-door criterion, Z has no descendant of X. Also $P(y \,|\, do(x), z) = P(y \,|\, x, z)$ since Z blocks all paths from X to Y, so by modularity

- Back-door criterion is equivalent to the ***conditional exchangeability / unconfoundedness*** assumption as in [Imbens and Rubin, 2015].

- The back-door adjustment is useful **only when** the ***causal DAG is available***.

2.2 The Front-door Adjustment

- The back-door criterion provides ***sufficient condition*** to identify adjustment set. If the covariates in back-door path is unobserved, it is not possible to block back-door path. Figure

2 (a).

- If there exists a ***mediator*** M in the path from T to Y (i.e. $T \to M \to Y$), we can isolate the association that flows through M by focusing our statistical analysis on M, and the only association that flows through M is causal association.

- We will focus our analysis on M using a **three step procedure** (see Figure 2 (c) for our corresponding illustration):

    1. **Identify the causal effect of T on M, i.e. $\underline{p(m|\, do(t))}$.** From Figure 2 (b), we see that $M \to Y \leftarrow W$, i.e. Y is a collider between W and M. This means that the back-door path to T from M is blocked. So T satisfies back-door criterion and $T \to M$ is a causal association. Thus the adjustment set is empty and by back-door adjustment

    $$p(m|\, do(t)) = p(m|\, t)$$

    2. **Identify the causal effect of M on Y, i.e. compute $\underline{p(y|\, do(m))}$.** Because T blocks the backdoor path $Y \leftarrow W \to T \to M$, we can simply adjust for T. Using back-door adjustment

    $$p(y|\, do(m)) = \sum_{t'} p(y|\, m, t')\, p(t')$$

    3. **Combine the above steps to identify the causal effect of T on Y.**

    $$\begin{aligned} p(y|\, do(t)) &= \sum_{m} p(m|\, do(t))\, p(y|\, do(m)) \\ &= \sum_{m} p(m|\, t) \sum_{t'} p(y|\, m, t')\, p(t') \end{aligned}$$

    The first factor on the right-hand side corresponds to setting T to t and observing the resulting value of M. The second factor corresponds to **setting** M to **exactly** the value m that **resulted from setting** T and then observing what value of Y results.

- **Definition** (***Front-Door***) [Pearl, 2000]
  A set of variables M is said to satisfy the ***front-door criterion*** relative to an ***ordered*** pair of variables (T, Y) if:

    1. M **intercepts all** directed paths from T to Y;
    2. there is ***no unblocked back-door path*** from T to M; and
    3. ***all back-door paths*** from M to Y are ***blocked*** by T.

- A set of variables M ***completely mediates*** the effect of T on Y if all causal (directed) paths from T to Y go through M. If M satisfies the front-door criterion, M is a set of complete mediators.

- **Theorem 2.2** *(**Front-Door Adjustment**) [Pearl, 2000, Neal, 2020]*
  *If M satisfies the front-door criterion relative to (T, Y) and if $P(t, m) > 0$, then the causal effect of T on Y is **identifiable** and is given by the formula*

  $$P(y|\, do(t)) = \sum_{m} P(m|\, t) \sum_{t'} P(y|\, m, t')\; P(t') \tag{3}$$

Proof: We can prove this theorm using Rules from *do*-calculus.

$$\begin{aligned}
P(y|\,do(t)) &= \sum_m P(y|\,do(t), m)P(m|\,do(t)) \quad \text{(marginalization trick)} \\
&= \sum_m P(y|\,do(t), m)P(m|\,t) \quad (\textbf{action to observation on } t) \\
&= \sum_m P(y|\,do(t), do(m))P(m|\,t) \quad (\textbf{observation to action exchange on } m) \\
&= \sum_m P(y|\,do(m))P(m|\,t) \quad (\textbf{deletion of actions } do(t)) \\
&= \sum_m P(m|\,t) \sum_{t'} P(y|\,do(m), t')\,p(t'|\,do(m)) \quad \text{(marginalization trick)} \\
&= \sum_m P(m|\,t) \sum_{t'} P(y|\,m, t')\,p(t'|\,do(m)) \quad (\textbf{action to observation on } m) \\
&= \sum_m P(m|\,t) \sum_{t'} P(y|\,m, t')\,p(t') \quad (\textbf{deletion of actions } do(m)) \quad \blacksquare
\end{aligned}$$

- Because the backdoor path from T to M is blocked by the collider Y, all of the association that flows from T to M is causal. Thus $P(m|\,do(t)) = P(m|\,t)$
- Under $do(T = t)$, there is no back-door path from M to T in induced graph. So $P(y|\,do(t), m) = P(y|\,do(t), do(m))$.
- Note that M is a fully mediator. Given $do(M = m)$, we can remove the link between T to M and $Y \perp\!\!\!\perp T|M$ in the induced graph. Thus we drop $do(T = t)$.
- We then introduce T by conditioning and marginalizing on T. Given $T = t'$, T blocks all back-door from M to Y so $M \rightarrow T$ is causal link, i.e. $P(y|\,do(m), t') = P(y|\,m, t')$
- Finally, M has no causal effect on T since the back-door path is blocked by Y. So $p(t'|\,do(m)) = p(t')$.

2.3 Propensity scores

- Given a high dimensional confounder set W that satisfies the backdoor criterion (or, equivalently, that $T \perp\!\!\!\perp Y|W$), it is not necessary to control on all of covariates in W.
- **Definition** (***Propensity Scores***) [Rosenbaum, 2017, Neal, 2020]
  If W satisfies ***unconfoundedness*** and ***positivity***, the ***propensity score*** is defined as

$$e(W) = P(T = 1\,|\,W). \tag{4}$$

  It is the probabiilty of receiving treatment ***within levels of*** W.

- **Theorem 2.3** *(**Propensity Score Theorem**) [Neal, 2020]*
  *Given positivity, unconfoundedness given W, implies **unconfoundedness** given the propensity score $e(W)$.*

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \;\;\Rightarrow\;\; (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W) \tag{5}$$

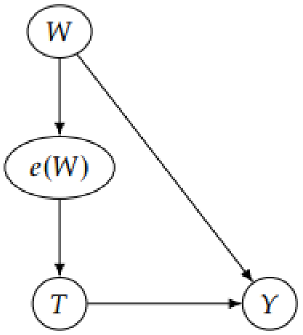


Figure 7.4: Graph illustrating that $e(W)$ blocks the backdoor path(s) that W blocks.

Figure 3: The propensity score blocks the backdoor paths [Neal, 2020]

- To prove this theorem, we note that $e(W)$ completely determines the edge $W \to T$, since it completely describes the mechanism $P(T|W)$ with binary $T \in \{0,1\}$. Thus we can think of the *propensity score* as a ***full mediator*** of the effect of W on T. Figure 3 indicates that $(e(W))$ blocks the back-door path that W blocks. Therefore, we have a graphical proof of the propensity score theorem using the backdoor adjustment.
- Recall The ***Positivity-Unconfoundedness Tradeoff***. As the dimensionality of W increases, the confoundedness will decrease, but the propensity score $e(W)$ will decrease, since the ***overlap*** between control and treatment group will decrease. This is also effect of ***the curse of dimensionality***.
- The propensity score $e(W)$ is usually unknown and we need to estimate it by training a model (e.g. *logistic regression*) to predict T from W.

3 Controlling covariate balance by matching

- A critical difference between randomized experiments and observational studies is that the covariate balance is not guaranteed in observation study. It puts in question the comparablity between treatment and control groups.
- In order to obtain covariate balance, we can select sub-populations of treatments and control group so that their (observed) covariate distributions are close (***stochastic balancing***). This process is called ***matching*** or ***treatment selection***.
- Matching is a naive approach to address the ***comparability*** issue: it says that people who *look comparable* are comparable.
- The **advantages** for matching include:
  - ***Controlling for confounder*** is achieved **at design phase** *without looking at outcomes.*
  - Matching will reveal ***lack of overlap*** in *covariate distribution.* That is, matching guarantees that the ***positivity assumption*** holds for both control and treatment group.

  - Once matched, we can **treat it as randomized experiments**.
  - Outcome analysis could be simple.

- Usually we care about causal treatment effects ***on the treated*** so we choose a subset of control group to **match the treatment group**.

3.1 Matching directly on confounders

- In order to find matches between treatment and control, we need to define some distance metrics to define *closeness*.

  1. The ***Mahalanobis distances*** between covariate of i, $\boldsymbol{w}_i$ and covariate of j, $\boldsymbol{w}_j$ is

     $$d_{\boldsymbol{\Sigma}}(\boldsymbol{w}_i, \boldsymbol{w}_j) = \left((\boldsymbol{w}_i - \boldsymbol{w}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{w}_i - \boldsymbol{w}_j) \right)^{\frac{1}{2}}$$

     for some positive definite matrix $\boldsymbol{\Sigma}$. A common choice of $\boldsymbol{\Sigma} = \widehat{\mathrm{Cov}}(W)$ is the estimated **covariance matrix** for W.

  2. The ***Robust Mahalanobis distances*** [Rosenbaum, 2017] is proposed to handle **outliers** in the covariates to avoid they increase the distance too much while the other covariates remain close.

     The idea to use ***order statistics (ranks)*** instead of absolute values for each $\boldsymbol{w}_i$.

     $$d_{robust}(\boldsymbol{w}_i, \boldsymbol{w}_j) = \left\| \boldsymbol{r}(\boldsymbol{w}_i) - \boldsymbol{r}(\boldsymbol{w}_j) \right\|_2$$

     where $\boldsymbol{r}(\boldsymbol{w}) = [r(w_1), \ldots, r(w_k)]$ is a list of ranks for each covariate.

  3. The distance between ***propensity scores*** $d_e(\boldsymbol{w}_i, \boldsymbol{w}_j) := \left\| e(\boldsymbol{w}_i) - e(\boldsymbol{w}_j) \right\|_2$.

- During ***pair matching (i.e. one-to-one matching)***, we select subject i with $\boldsymbol{w}_i$ *in the control group* whose distance $d(\boldsymbol{w}_i, \boldsymbol{w}_r) \le \epsilon$ for some $\epsilon > 0$, where $\boldsymbol{w}_r$ corresponds to a reference subject r in the treatment group.

3.2 Greedy matching by nearest neighbor

- Instead of matching using fixed threshold ϵ (ϵ-ball matching), we can determine the definition of closeness in nonparametric way.

- **Definition** (***Nearest neighbor matching (Greedy matching)***)
  Assume that W is a ***sufficient adjustment set*** that satisfies back-door criterion, i.e. the conditional ignorablity holds given W. Let $d_{i,j} := d(\boldsymbol{w}_i, \boldsymbol{w}_j)$, for $\boldsymbol{w}_i, \boldsymbol{w}_j \in W$ be pairwise distance between covariates of subjects in treatment and control groups.

  The ***nearest neighbor matching*** (***greedy matching***) has the following steps:

  1. Randomly order list of subjects in control set C and treatment set T;

  2. (**Greedy match**) Select a treated subject $r \in T$. Choose a control subject $i \in C$ so that

     $$i \in \arg\min_{j \in C} d(\boldsymbol{w}_r, \boldsymbol{w}_j);$$

3. Remove matched control i from C, i.e. $C \leftarrow C - \{i\}$ and $T \leftarrow T - \{r\}$;
4. Loop above steps until $T = \emptyset$ assuming more controls than treatments $|C| > |T|$.

- The benefits for greedy match are
  1. **Intutive**.
  2. **Computational fast** if implemented with fast approximate nearest neighbor algorithms, locality sensitive hashing etc.
- On the other hand, the disadvantages are
  1. **Not invariant to initial ordering**. Note that the previous match will affect the latter match since it will take out the candidates.
  2. **Not optimal**. It is a greedy algorithm, i.e. it does not guarantee that the closest distance match locally is the closest distance match globally. This is because this algorithm use ***hard match assignment*** instead of *soft match assignement* (i.e. with probability).
- An alternative to pair matching is ***many-to-one matching ($k : 1$ matching)***: after every one has 1 match, we loop back to the first treated in the list and find the second best match of it. Loop until every one has exactly k matches.
- There are several **tradeoffs** (**Bias-variance tradeoff**):
  - For pair matching, the matching is closer, thus has **low bias**, but have high variance and **less data efficient** since we are dropping data. It also has higher **computational speed**;
  - For many-to-one matching, it has **larger sample size** so it has smaller variance. It will add more bias since we are not selecting the best matches.
- We might want to discard treated subjects when there is no good match. We referred to the ***maximum acceptable distance*** as a ***caliper*** [Rosenbaum, 2017].
  - Only match a treated subject if the best match has distance less than the caliper.
  - If no match within caliper, then the **positivity assumption** would be violated. Excluding these subjects makes the positivity assumption more realistic.
  - The drawback is that the **definition** of population is **unclear** after removing treated subjects.

3.3 Optimal pair matching

- Greedy matching is not *global optimal*, i.e. $\sum_r d(\boldsymbol{w}_r, \boldsymbol{w}_{nn(r)}) \neq \min_{\{j_1,\ldots,j_T\}} \sum_r d(\boldsymbol{w}_r, \boldsymbol{w}_{j_r})$.
- In optimal pair matching, the objective is to pairs (r, j_r) for all $r \in T$ and $j_r \in C$ so that

$$\min_{\{j_1,\ldots,j_T\} \in C} \sum_{r=1}^{T} d(\boldsymbol{w}_r, \boldsymbol{w}_{j_r})$$

- It is an ***optimal assignment problem*** [Peyr and Cuturi, 2019], where the pairwise distance is the cost for each pair assignment. This is also a ***minimum flow problem*** in **network**

flow optimization [Rosenbaum, 2017]. In particular, we can drop the integer constraint to obtain a *linear programming relaxation* as

$$\begin{aligned} \min_{f(u,v)\,\forall(u,v)\in\mathcal{E}_{C,T}} \quad & d(\boldsymbol{w}_u, \boldsymbol{w}_v) f(u,v) \\ \text{s.t.} \quad & \sum_{v\in\mathcal{V}_T} f(u,v) = 1 \\ & \sum_{u\in\mathcal{V}_C} f(u,v) = 1 \\ & f(u,v) \ge 0, \quad \forall(u,v) \in \mathcal{E}_{C,T} \end{aligned}$$

where $\mathcal{G} = (\mathcal{V}_C \cup \mathcal{V}_T, \mathcal{E}_{C,T})$ is a bipartite graph between treatment and control group index set.

- Whether or not it is feasible to perform optimal matching depends on the **size of the problem**. **Size** is typically measured by *number of possible treatment-control pairing*, i.e. $|\mathcal{E}_{C,T}|$.
- Contraints can be imposed to make optimal matching computationally feasible for larger data set, i.e. reduce the edge set $\mathcal{E}_{C,T}$. This is known as ***sparse matching***.

3.4 Matching by propensity score

- The propensity score $e(W)$ computes the probability of receiving treatment under levels of W. In ***completely randomized experiments***, it is pre-determined $e(\boldsymbol{w}) = \frac{1}{2}$ for all $\boldsymbol{w}$.
- If two subjects have the **same propensity score**, $e(\boldsymbol{w})$, they may have different values of $\boldsymbol{w}$. However, their effects on the treatment assignment T is equal.
- The *goal* is to choose the subpopulation with estimated propensity score $\hat{e}(\boldsymbol{w}_n)$ **matches** between treatment and control groups.

  We can use distance-based matching as discussed above.
- The ***advantages*** of matching by propensity score:
  - Matching on $e(\boldsymbol{w})$ is often ***practical*** even when there are many covariates in $\boldsymbol{w}$ because $e(\boldsymbol{w})$ is a **single** variable;
  - ***Matching on*** $e(\boldsymbol{w})$ ***tends to balance all of w*** (i.e. *propensity score is a* ***matching score***);

    Note that $e(W)$ is a full mediator of W to T. So the covariate balance between control and treatment group given that the propensity score is matched.

    $$W \perp\!\!\!\perp T \,|\, e(W) \Leftrightarrow \quad p(W|\, e(W) = e,\, T = 0) = p(W|\, e(W) = e,\, T = 1)$$
  - Failure to balance $e(\boldsymbol{w})$ implies that $\boldsymbol{w}$ is not balanced.
- Given estimated propensity score, it is useful to first check the **overlap of support** for **propensity score distributions** $p(\hat{e}(W)|T)$ between treatment and control before matching. This is to verify the positivity assumption.

- If there lacks of overlap, **trimming the tails** is an option, i.e. to remove subjects with extreme scores. It would prevent ***extrapolations***.
- In practice, the ***log-odds (logit) of propensity score*** $\operatorname{logit}(e(W)) := \log\left(\frac{e(W)}{1-e(W)}\right)$ is often used as an alternative unbounded score which preserves ranks.
- In practice, the ***caliper*** is chosen as $0.2 \times \sigma\left(\operatorname{logit}(e(W))\right)$.
- It is important to understand that this is a true statement about **observed covariates** $\boldsymbol{w}$, and **only** about observed covariates, whether or not treatment assignment also depends on covariates that were **not measured**.

  On the negative side, success in balancing the observed covariates, $\boldsymbol{w}$, provides ***no assurance*** that **unmeasured covariates** are balanced [Rosenbaum, 2017].

3.5 Accessing balance

- After the matching is completed, we can evaluate its performance by checking on the covariate distributions for treatment and control groups.
- Hypothesis testing (two-sample t-test) and make decision based on p-values. Note that p-value is determined by the sample size as well.
- We can use a table that looks at pre-matching vs. post-matching covariate balance. In each part, we look at the ***standardized mean differences***

$$\text{smd} = \frac{\bar{m}_T - \bar{m}_C}{\sqrt{\frac{\hat{\sigma}_T^2 + \hat{\sigma}_C^2}{2}}}$$

  We can compute the standardized differences for each covariate. Usually choose threshold as ≤ 0.1.

4 Unobserved Confounding

- In observational studies, there could also be some ***unobserved confounder(s)***. Therefore, wed like to know how **robust** our estimates are to unobserved confounding.
  - The first way we can do is by getting an *upper and lower* ***bound*** on the causal effect using credible **assumptions**.
  - Another way we can do this is by simulating **how strong** the confounders effect on the treatment and the confounders effect on the outcome need to be to make *the true causal effect* ***substantially different*** *from our estimate* (i.e. ***sensitivity analysis***).
  - We can also use the ***instrumental variables***, which will be discussed in its own chapter.

4.1 Bounds

- Depending on what assumptions we are willing to make, we can derive various nonparametric bounds on causal effects. For example, the unconfoundedness assumption before might be unrealistic given unobserved confounder exists. In this section, we use assumptions weaker than the unconfoundedness assumption.
- Rather than telling us exactly what the causal effect must be, we can give **an interval** that the causal effect must be in.
- Under the unconfoundedness assumption, the interval will collapse to be a single point.

4.1.1 No-assumptions bound

- **Assumption 4.1** ***(Bounded Potential Outcomes)*** *[Neal, 2020]*

$$a \leq Y(0), Y(1) \leq b \tag{6}$$

- We can find a decomposition of the **Average Treatment Effect (ATE)**

  Proposition 4.2 ***(Observational-Counterfactual Decomposition)*** *[Neal, 2020]*

$$\begin{aligned}\mathbb{E}\left[Y(1)-Y(0)\right] = &\{\pi\,\mathbb{E}\left[Y \mid T=1\right] - (1-\pi)\,\mathbb{E}\left[Y \mid T=0\right]\} \\ &+ \{(1-\pi)\,\mathbb{E}\left[Y(1) \mid T=0\right] - \pi\,\mathbb{E}\left[Y(0) \mid T=1\right]\}\end{aligned} \tag{7}$$

  where $\pi = P(T=1)$.

  Proof: Note

$$\begin{aligned}\mathbb{E}\left[Y(1)-Y(0)\right] &= p(T=1)\mathbb{E}\left[Y(1)-Y(0)|T=1\right] + p(T=0)\mathbb{E}\left[Y(1)-Y(0)|T=0\right] \\ &= \pi\mathbb{E}\left[Y(1) \mid T=1\right] - \pi\mathbb{E}\left[Y(0)|T=1\right] \\ &\quad + (1-\pi)\mathbb{E}\left[Y(1) \mid T=0\right] - (1-\pi)\mathbb{E}\left[Y(0) \mid T=0\right] \\ &= \pi\mathbb{E}\left[Y \mid T=1\right] - \pi\mathbb{E}\left[Y(0)|T=1\right] \qquad \text{(by consistency)} \\ &\quad + (1-\pi)\mathbb{E}\left[Y(1) \mid T=0\right] - (1-\pi)\mathbb{E}\left[Y \mid T=0\right] \qquad \blacksquare\end{aligned}$$

- The first two terms in (7) are ***observational*** and the last two term are ***counterfactual***. Since the last two terms are not observed, we need to bound the difference using bounded potential outcome assumption (6).
- **Proposition 4.3** ***(No-Assumptions Bound)*** *[Neal, 2020]*
  Let π denote $P(T=1)$, where T is a binary random variable. Given that the outcome Y is bounded between a and b (Assumption 4.1), we have the following ***upper and lower bounds*** *on the ATE:*

$$\mathbb{E}\left[Y(1)-Y(0)\right] \leq \{\pi\,\mathbb{E}\left[Y \mid T=1\right] - (1-\pi)\,\mathbb{E}\left[Y \mid T=0\right]\} + \left[(1-\pi)\,b - \pi\,a\right] \tag{8}$$

$$\mathbb{E}\left[Y(1)-Y(0)\right] \geq \{\pi\,\mathbb{E}\left[Y \mid T=1\right] - (1-\pi)\,\mathbb{E}\left[Y \mid T=0\right]\} + \left[(1-\pi)\,a - \pi\,b\right] \tag{9}$$

  More importantly, the interval where $\mathbb{E}\left[Y(1)-Y(0)\right]$ lies has length $(b-a)/2$.

4.1.2 Monotone treatment response

- In order to improve the no-assumption bounds, we need to consider additional assumptions on the counterfactual difference terms in (7).
- **Assumption 4.4** ***(Nonnegative Monotone Treatment Response)*** *[Neal, 2020]*

$$\forall\, i, \quad Y_i(1) \geq Y_i(0). \tag{10}$$

- This means that every ITE is nonnegative, i.e. ***the treatment cannot hurt anyone***. It is easy to see that the ATE is above zero. This can be derived by seeing

$$\mathbb{E}\left[Y(1) \,|\, T=0\right] \geq \mathbb{E}\left[Y(0)|T=0\right] = \mathbb{E}\left[Y|T=0\right]$$
$$-\mathbb{E}\left[Y(0)|T=1\right] \geq -\mathbb{E}\left[Y(1) \,|\, T=1\right] = -\mathbb{E}\left[Y \,|\, T=1\right]$$

  Substituting this into (7), we can have the bound

- **Proposition 4.5** ***(Nonnegative MTR Lower Bound)*** *[Neal, 2020]*
  Under the nonnegative MTR assumption, the ATE is bounded from below by 0. Mathematically,

$$\mathbb{E}\left[Y(1) - Y(0)\right] \geq 0$$

- We can switch the sign if we have *Nonpositive Monotone Treatment Response.*

4.1.3 Monotone treatment selection

- **Assumption 4.6** ***(Monotone Treatment Selection)*** *[Neal, 2020]*

$$\begin{aligned}\mathbb{E}\left[Y(1) \,|\, T=1\right] &\geq \mathbb{E}\left[Y(1) \,|\, T=0\right]\\ \mathbb{E}\left[Y(0) \,|\, T=1\right] &\geq \mathbb{E}\left[Y(0) \,|\, T=0\right]\end{aligned} \tag{11}$$

- This assumes that the people who **selected treatment** would have **better outcomes** than those who didnt select treatment, under either treatment scenario.
- **Proposition 4.7** ***(Monotone Treatment Selection Upper Bound)*** *[Neal, 2020]*
  Under the MTS assumption, the ATE is bounded from above by the ***associational difference****. Mathematically,*

$$\mathbb{E}\left[Y(1) - Y(0)\right] \leq \mathbb{E}\left[Y \,|\, T=1\right] - \mathbb{E}\left[Y \,|\, T=0\right] \tag{12}$$

  Proof:

$$\begin{aligned}\mathbb{E}\left[Y(1) - Y(0)\right] &= \{\pi\, \mathbb{E}\left[Y \,|\, T=1\right] - (1-\pi)\, \mathbb{E}\left[Y \,|\, T=0\right]\}\\ &\quad + \{(1-\pi)\, \mathbb{E}\left[Y(1) \,|\, T=0\right] - \pi\, \mathbb{E}\left[Y(0) \,|\, T=1\right]\}\\ &\leq \{\pi\, \mathbb{E}\left[Y \,|\, T=1\right] - (1-\pi)\, \mathbb{E}\left[Y \,|\, T=0\right]\}\\ &\quad + \{(1-\pi)\, \mathbb{E}\left[Y(1) \,|\, T=1\right] - \pi\, \mathbb{E}\left[Y(0) \,|\, T=0\right]\}\\ &= \mathbb{E}\left[Y \,|\, T=1\right] - \mathbb{E}\left[Y \,|\, T=0\right] \quad \blacksquare\end{aligned}$$

4.1.4 Optimal treatment selection

- **Assumption 4.8** ***(Optimal Treatment Selection)*** *[Neal, 2020]*

$$T_i = 1 \Rightarrow Y_i(1) \geq Y_i(0), \quad \textit{and} \quad T_i = 0 \Rightarrow Y_i(0) > Y_i(1) \tag{13}$$

- This assumption means that the individuals always receive the treatment that is **best** for them (e.g. if an expert doctor is deciding which treatment to give people).

- From OTS, we see that

$$\begin{aligned} &\mathbb{E}\left[Y(0)\,|\,T=1\right] \leq \mathbb{E}\left[Y(1)\,|\,T=1\right] = \mathbb{E}\left[Y\,|\,T=1\right] \\ &\mathbb{E}\left[Y(1)\,|\,T=0\right] < \mathbb{E}\left[Y(0)\,|\,T=0\right] = \mathbb{E}\left[Y\,|\,T=0\right]. \end{aligned}$$

  We can combine OTS and bounded potential outcomes to have new bound

- **Proposition 4.9** ***(Optimal Treatment Selection Bound 1)*** *[Neal, 2020]*
  Let π denote $P(T=1)$, where T is a binary random variable. Given that the outcome Y is bounded from below by a (Assumption 4.1) and that the optimal treatment is always selection (Assumption 4.8), we have the following upper and lower bounds on the ATE:

$$\begin{aligned} \mathbb{E}\left[Y(1)-Y(0)\right] &< \pi\mathbb{E}\left[Y\,|\,T=1\right] - \pi\, a \qquad (14) \\ \mathbb{E}\left[Y(1)-Y(0)\right] &\geq (1-\pi)\, a - (1-\pi)\mathbb{E}\left[Y\,|\,T=0\right] \qquad (15) \\ (\textit{Interval Length}) &= \pi\mathbb{E}\left[Y\,|\,T=1\right] + (1-\pi)\mathbb{E}\left[Y\,|\,T=0\right] - a \qquad (16) \end{aligned}$$

- From the OTS, we can also derive

$$\begin{aligned} \mathbb{E}\left[Y(1)\,|\,T=0\right] &= \mathbb{E}\left[Y(1)\,|\,Y(0) > Y(1)\right] \quad \text{(by OTS)} \\ &< \mathbb{E}\left[Y(1)\,|\,Y(0) \leq Y(1)\right] \\ &\quad \text{since } \mathbb{E}\left[Y(1)\,|\,Y(0) > Y(1)\right] < \mathbb{E}\left[Y(0)\right] \leq \mathbb{E}\left[Y(1)\,|\,Y(1) \geq Y(0)\right] \\ &= \mathbb{E}\left[Y(1)\,|\,T=1\right] \quad \text{(by counterpositive argument of OTS)} \\ &= \mathbb{E}\left[Y\,|\,T=1\right] \qquad (17) \end{aligned}$$

  In other word, the **counterfactual outcome** ***given not treated*** will still be ***worse*** than the **actual outcome** ***given treated***. (actual outcome better than speculated outcome)

  Substitute this bound to (7), we have the following bound

- **Proposition 4.10** ***(Optimal Treatment Selection Bound 2)*** *[Neal, 2020]*
  Let π denote $P(T=1)$, where T is a binary random variable. Given that the outcome Y is bounded from below by a (Assumption 4.1) and that the optimal treatment is always selection (Assumption 4.8), we have the following upper and lower bounds on the ATE:

$$\begin{aligned} \mathbb{E}\left[Y(1)-Y(0)\right] &< \mathbb{E}\left[Y\,|\,T=1\right] - \pi\, a - (1-\pi)\mathbb{E}\left[Y\,|\,T=0\right] \qquad (18) \\ \mathbb{E}\left[Y(1)-Y(0)\right] &\geq \pi\mathbb{E}\left[Y\,|\,T=1\right] + (1-\pi)\, a - \mathbb{E}\left[Y\,|\,T=0\right] \qquad (19) \\ (\textit{Interval Length}) &= (1-\pi)\mathbb{E}\left[Y\,|\,T=1\right] + \pi\mathbb{E}\left[Y\,|\,T=0\right] - a \qquad (20) \end{aligned}$$

  This interval (20) is usually larger than that in (16). But the lower bound can be above zero. And we can take intersection between the intervals from OTS bound 1 and bound 2.

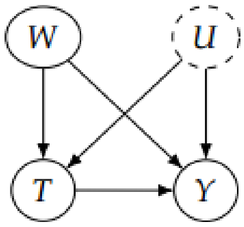


Figure 8.3: Simple causal structure where W is the observed confounders and U is the unobserved confounders.

Figure 4: The linear structural equations with an unobserved confounder U for sensitivity analysis [Neal, 2020]

- Some important takeaways:
  - Different bounds are better in **different cases**.
  - Different bounds can be better in different ways (e.g., *identifying the sign* vs. getting *a smaller interval*).

4.2 Sensitivity analysis

4.2.1 Overview

- Matching aims to achieve covariate balance on *observed covariates.* If there was imbalance on observed covariates, we have **overt bias** (i.e. we do not have full control over these covariates).
- When the ***unobserved confounder*** exists, we have ***hidden bias***. In this situtation, the causal effect is ***not identifiable***.
- The **main idea** of ***sensitivity analysis*** is to determine how severe the hidden bias has to be in order to **change the conclusion** e.g. change the *statistical significance* of the results or change the *direction* of effect.

4.2.2 Sensitivity basics in linear setting

- Assume a ***noiseless linear*** structrual causal equation:

$$T := \alpha_w W + \alpha_u U \tag{21}$$
$$Y := \beta_w W + \beta_u U + \delta\, T \tag{22}$$

  where W and U are both confounders for the causal effect of T on Y. See Figure 4.

- From this equation, we see that the ATE is determined by δ. By back-door assignment,

$$P(Y|\,do(T=t)\,) = \sum_{w,u} P(Y\,|\,T=t, w,\, u)P(w,u)$$

$$\begin{aligned}\Rightarrow \mathbb{E}\left[Y(1)-Y(0)\right] &= \mathbb{E}\left[Y \mid do(T=1)\,\right] - \mathbb{E}\left[Y \mid do(T=0)\,\right]\\ &= \mathbb{E}_{W,U}\left[\mathbb{E}\left[Y \mid T=1, W, U\right] - \mathbb{E}\left[Y \mid T=0, W, U\right]\right] = \delta\end{aligned}$$

- But because U isnt observed, the best we can do is adjust for only W. This leads to a ***confounding bias (hidden bias)*** of $\frac{\beta_u}{\alpha_u}$
- **Proposition 4.11** *[Neal, 2020]*
  *When T and U are generated by the **noiseless linear process** in Equations (21) and (22), the **confounding bias** of adjusting for **just** W (and not U) is $\frac{\beta_u}{\alpha_u}$. Mathematically:*

$$\begin{aligned}&\mathbb{E}_W\left[\mathbb{E}\left[Y \mid T=1, W\right] - \mathbb{E}\left[Y \mid T=0, W\right]\right]\\ &-\mathbb{E}_{W,U}\left[\mathbb{E}\left[Y \mid T=1, W, U\right] - \mathbb{E}\left[Y \mid T=0, W, U\right]\right] = \frac{\beta_u}{\alpha_u}\end{aligned} \tag{23}$$

Proof: Well prove Proposition in 3 steps:

1. Get a closed-form expression for $\mathbb{E}_W\left[\mathbb{E}\left[Y \mid T=t, W\right]\right]$ in terms of α_w, α_u, β_w and β_u.
2. Use step 1 to get a closed-form expression for the difference $\mathbb{E}_W\left[\mathbb{E}\left[Y \mid T=1, W\right] - \mathbb{E}\left[Y \mid T=0, W\right]\right]$
3. Subtract off $\mathbb{E}_{W,U}\left[\mathbb{E}\left[Y \mid T=1, W, U\right] - \mathbb{E}\left[Y \mid T=0, W, U\right]\right] = \delta$

The first step can be obtained by substituting the SCM (22).

$$\begin{aligned}\mathbb{E}_W\left[\mathbb{E}\left[Y \mid T=t, W\right]\right] &= \mathbb{E}_W\left[\mathbb{E}\left[\beta_w W + \beta_u U + \delta\, T \mid T=t, W\right]\right]\\ &= \mathbb{E}_W\left[\beta_w W + \delta\, t + \beta_u \mathbb{E}\left[U \mid T=t, W\right]\right]\end{aligned} \tag{24}$$

From (21), we have

$$U = \frac{T - \alpha_w\, W}{\alpha_u}. \tag{25}$$

So substituting it into (24), we have

$$\begin{aligned}\mathbb{E}_W\left[\mathbb{E}\left[Y \mid T=t, W\right]\right] &= \mathbb{E}_W\left[\beta_w W + \delta\, t + \beta_u \frac{t - \alpha_w\, W}{\alpha_u}\right]\\ &= \left(\delta + \frac{\beta_u}{\alpha_u}\right) t + \left(\beta_w - \frac{\beta_u\, \alpha_w}{\alpha_u}\right) \mathbb{E}\left[W\right]\end{aligned} \tag{26}$$

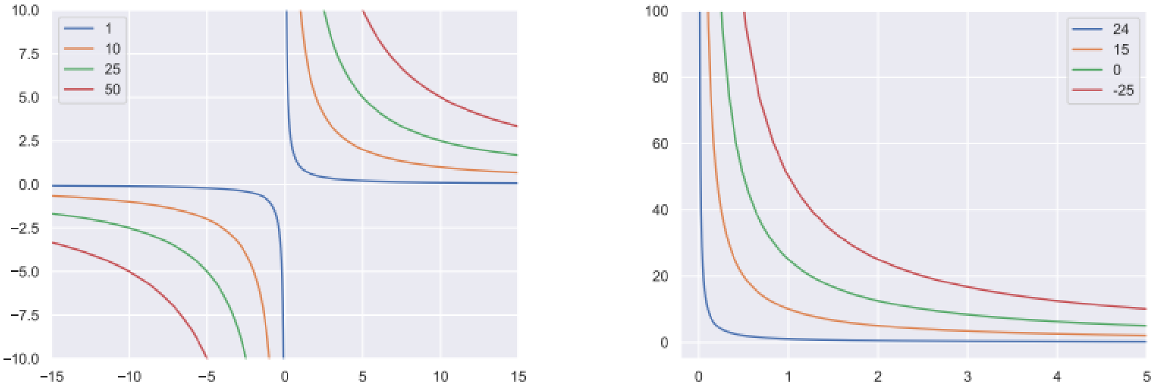
In the second step, substituting (26) into difference equation

$$\mathbb{E}_W\left[\mathbb{E}\left[Y \mid T=1, W\right]\right] - \mathbb{E}_W\left[\mathbb{E}\left[Y \mid T=0, W\right]\right] = \left(\delta + \frac{\beta_u}{\alpha_u}\right) \tag{27}$$

Finally, substracting $\mathbb{E}_{W,U}\left[\mathbb{E}\left[Y \mid T=1, W, U\right] - \mathbb{E}\left[Y \mid T=0, W, U\right]\right] = \delta$,

$$\mathbb{E}_W\left[\mathbb{E}\left[Y \mid T=1, W\right]\right] - \mathbb{E}_W\left[\mathbb{E}\left[Y \mid T=0, W\right]\right] - \mathbb{E}_{W,U}\left[\mathbb{E}\left[Y \mid T=1, W, U\right] - \mathbb{E}\left[Y \mid T=0, W, U\right]\right] = \frac{\beta_u}{\alpha_u}$$

■



(a) Contours of confounding bias $\frac{\beta_u}{\alpha_u}$

(b) Contours of the true ATE δ, given that $\mathbb{E}_W[\mathbb{E}[Y \mid T=1, W] - \mathbb{E}[Y \mid T=0, W]] = 25$

Figure 8.5: Contour plots for sensitivity where the x-axis for both is $\frac{1}{\alpha_u}$ and the y-axis is β_u. There is a color-coded correspondence between the curves in the upper right of Figure 8.5b and the curves in Figure 8.5

Figure 5: The sensitivity curve for $\frac{\beta_u}{\alpha_u}$ under different choice of (α_u, β_u) [Neal, 2020]

4.2.3 Sensitivity contour plots

- We can rearrange the equation (23)

$$\delta = \mathbb{E}_W \left[\mathbb{E}\left[Y \mid T = 1, W \right] - \mathbb{E}\left[Y \mid T = 0, W \right] \right] - \frac{\beta_u}{\alpha_u}$$

  Thus so for given values of α_u and β_u, we can compute the **true ATE** δ, from the *observational quantity* $\mathbb{E}_W \left[\mathbb{E}\left[Y \mid T = 1, W \right] - \mathbb{E}\left[Y \mid T = 0, W \right] \right]$

- This allows us to get ***sensitivity curves*** that allow us to know how ***robust*** conclusions like "$\mathbb{E}_W \left[\mathbb{E}\left[Y \mid T = 1, W \right] - \mathbb{E}\left[Y \mid T = 0, W \right] \right] = 25$ is positive, so δ is likely positive" are to *unobserved confounding*. See Figure 5.

- In other word, the observed quantity $\mathbb{E}_W \left[\mathbb{E}\left[Y \mid T = 1, W \right] - \mathbb{E}\left[Y \mid T = 0, W \right] \right]$ need to be *at least* $\frac{\beta_u}{\alpha_u}$ in order to ***change the sign*** of ATE.

- Note that α_u describes the **strength of influence** for the *unobserved confounder* U on **treatment** T and β_u describes the **strength of influence** for U on **outcome** Y.

$$\text{Confounding bias} = \frac{\beta_u}{\alpha_u} = \frac{\text{strength}(U \rightarrow Y)}{\text{strength}(U \rightarrow T)}$$

  Thus **confounding bias**, i.e. the gap $\frac{\beta_u}{\alpha_u}$ between ATE and the ***associational difference*** by controlling **observed confounder** W is ***propotional*** to the *strength of influence* of unobserved confounder U on **outcome** and ***inversely propotional*** to the *strength of influence* of U on the **treatment** T.

References

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Brady Neal. Introduction to causal inference from a machine learning perspective. *Course Lecture Notes (draft)*, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Gabriel Peyr and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237.

Paul Rosenbaum. *Observation and Experiment: An Introduction to Causal Inference*. Harvard University Press, 2017.