6.3720/6.3722 Introduction to Statistical Data Analysis May 8, 2024

Lecture 25: Causal inference

Instructor: Christina Ji *Written by: Christina Ji*

After this lecture, you will be able to:

1. Set up a causal question and define the answer in terms of the (conditional) average treatment effect
2. Design an experiment or observational study that satisfies standard assumptions in causal inference
3. Run an A/B experiment and use hypothesis testing to determine whether the average treatment effect is significant
4. Estimate treatment effect from observational data using a regression or nearest neighbor matching

1 Motivation

Causal inference is motivated by the need to answer the question: What is the effect of a treatment T on an outcome Y? For example, if a person in a car accident is taken to the hospital in a helicopter instead of an ambulance, will that improve their chance of survival? The treatment is taking a helicopter. The outcome is survival. To answer this question, suppose we have the data in Table 1. 24% of people who are taken by ambulance die, while 32% of people who are taken by helicopter die. If we run a two-sample t-test or a GLRT, this difference is statistically significant. This result is quite surprising! We would expect that taking a helicopter would result in better outcomes since patients will arrive at the hospital and get treated earlier. Thus, we should be hesitant to draw the conclusion that taking an ambulance leads to better outcomes.

Overall	Died	Survived
Ambulance	260	840
Helicopter	64	136

Table 1: Patient outcomes for each mode of transportation to hospital.

To understand what happened here, we can examine the data in Table 2, where outcomes are tabulated separately for serious and lighter accidents. In both of these cases, the proportion of patients who die is lower among those who are transported by helicopter. This phenomenon we just observed where the trend is reversed when conditioning on another feature is called Simpson's paradox. To avoid drawing the wrong conclusion, we must account for seriousness of the accident when examining causal effect.

Serious	Died	Survived	Lighter	Died	Survived
Ambulance	60	40	Ambulance	200	800
Helicopter	48	52	Helicopter	16	84

Table 2: Table 1 separated into two cases: more serious accidents on the left and lighter accidents on the right.

2 Fundamental Causal Question

With this insight in mind, let us revise the causal question we set out to answer: **If a person with features X is given a treatment T, what outcome Y will happen? The fundamental question in causal**

inference has three components: features X, treatment T, and outcome Y. In the example above, transportation via helicopter or ambulance is the treatment T, and survival is the outcome Y. Some features X that may be relevant include patient age, whether the patient received a head injury, consciousness at the time of transportation, car speed at the time of collision, and whether the person was wearing a seat belt.

Here is another example of a causal question: If Facebook gives a user more friend recommendations, will the user spend more time on Facebook? The number of friend recommendations is the treatment T. The amount of time spent on Facebook afterwards is the outcome Y. The effect of friend recommendations may vary for different users based on current usage of Facebook and sociability in real life. The latter is a feature that Facebook does not observe. That makes answering the causal question more challenging.

2.1 Conditional Average Treatment Effect

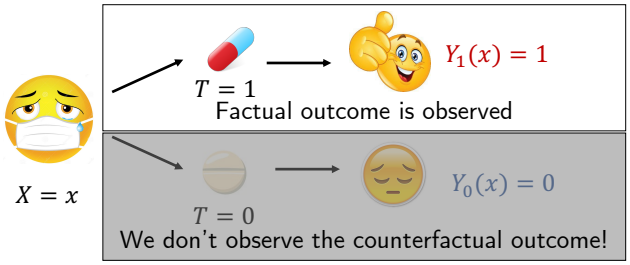

Figure 1: Causal inference set-up. $Y_1(x)$ and $Y_0(x)$ are potential outcomes.

Consider the set-up shown in Figure 1. If a patient with features $X = x$ is given treatment $T = 1$, the outcome would be $Y_1(x)$. If the same patient is given treatment $T = 0$, the outcome would be $Y_0(x)$.

Concept: Potential outcomes and conditional average treatment effect

Because the patient can only be given one treatment, only one outcome is observed. The observed outcome is called the **factual outcome**. Unobserved outcomes are called **counterfactual outcomes**. All outcomes are considered **potential outcomes**.

Definition 1 (Potential outcome)**.** *A potential outcome $Y_t(x)$ is the outcome that would be observed for a person with features x if treatment t is given.*

To answer the fundamental causal question, we can compute the **conditional average treatment effect**:

Definition 2 (Conditional average treatment effect (CATE))**.** *Conditional average treatment effect is the difference between potential outcomes with and without treatment:*

$$CATE := \mathbb{E}[Y_1(X) | X = x] - \mathbb{E}[Y_0(X) | X = x] \quad (1)$$

Estimating CATE is challenging because only one of the potential outcomes is observed for each person. If a person had a twin who received the opposite treatment, then we can estimate CATE using the factual outcomes from both twins. However, most people do not have a twin. Instead, we can try to estimate $\mathbb{E}[Y_t(x) | X = x] = f(x, t)$ using observations from many people. There are two additional challenges when estimating the expected potential outcome. First, f needs to be specified such that it can represent the true causal model. Second, we are using the factual outcomes $\mathbb{E}[Y | X = x, T = t]$ and assuming they are equal to the expected potential outcomes.

Practical guideline: Challenges in causal inference

To summarize the main challenges in causal inference,

1. Some features that affect both the treatment and outcome are not observed.
2. The counterfactual outcome is not observed.
3. The true causal model is unknown.
4. Assuming $\mathbb{E}\left[Y_t\left(x\right)|X=x\right]=\mathbb{E}\left[Y|X=x,T=t\right]$ might not hold if people who received $T=t$ have different potential outcomes compared to people who did not receive $T=t$.

2.2 Causal Assumptions

To address these challenges, there are three standard assumptions in causal inference:

1. **No unmeasured confounding:** A confounder is a factor affecting both treatment and outcome. All confounders need to be included among the features. An example of a confounder was given in lecture 21: When examining the effect of sleeping with a night light on development of myopia in children, a confounder is myopia in parents because parents who have myopia may be more inclined to leave a night light on (confounder affects treatment) and myopia is hereditary (confounder affects outcome). Thus, myopia among parents must be included as a feature in this analysis. This assumption is formally known as strong ignorability, that is, the potential outcomes are independent of the treatment conditioned on the features: $Y_0, Y_1 \perp\!\!\!\perp T|X$.
2. **Overlap between features of treated and control cohorts:** To compare the outcomes of patients in the two cohorts, we must observe similar patients in both cohorts: $\mathbb{P}\left(X=x|T=1\right)>0 \Leftrightarrow \mathbb{P}\left(X=x|T=0\right)>0$. As an example of when overlap might not hold is shown in Figure 2. Suppose we are analyzing the effect of a vasopressor (which constricts blood vessels to raise blood pressure) with patient age as a feature. Older patients may have lower blood pressure when untreated, but because older patients are almost always treated, we do not observe the potential outcome Y_0 and may underestimate treatment effect if we compare Y_1 among older patients and Y_0 among younger patients. With more features, this assumption is harder to satisfy.
3. **Stable unit treatment value:** This assumption consists of two parts: (i) **Causal consistency:** If a patient receives a treatment $T=t$, the corresponding potential outcome is observed: $\mathbb{E}\left[Y|X=x,T=t\right]=Y_t\left(x\right)$. That is, $T=t$ is the same treatment for all patients. This assumption would be violated if $T=1$ is recorded for any dosage. (ii) **No interference:** One person's treatment does not affect another person's outcome. This assumption may not hold when people are connected in a social network. For instance, if Facebook user A introduces a new friend recommendation to her existing friend Facebook user B, then user B's outcome is affected by user A's treatment.

It may be tempting to add as many features as possible to satisfy the no unmeasured confounding assumption. However, when there are too many features, no two patients are similar, and there would be little overlap between the treated and control cohorts. On the other hand, if the feature set is limited, there will be overlap, but some confounders may have been excluded, and the causal effect may be estimated incorrectly. The two assumptions can be balanced by selecting the necessary features with domain knowledge.

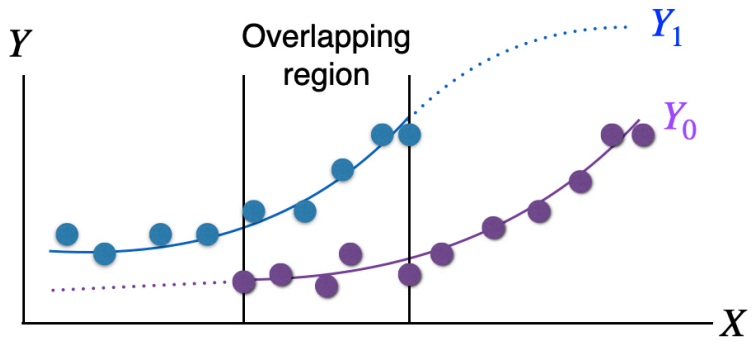

Figure 2: Solid lines denote where the potential outcomes $Y_0(x)$ and $Y_1(x)$ can be estimated from the factual outcomes. Dashed lines indicate the potential outcomes cannot be estimated. CATE can only be estimated in the overlapping region.

2.3 Summary: Setting up a Causal Analysis

Practical guideline: Setting up a causal analysis

When asking a causal question, we need to define the treatment, features, cohort, and outcome such that the causal assumptions hold:

1. All confounders are included among the features.
2. The cohort is defined in terms of features and restricted to include only people who might plausibly receive either treatment.
3. The treatment level is consistent across all people, and the treatment one person receives does not affect the outcome of another person.

When these assumptions are satisfied, we can estimate the conditional average treatment effect

$$CATE(x) = \mathbb{E}[Y_1(x)|X = x] - \mathbb{E}[Y_0(x)|X = x] \tag{2}$$

or the average treatment effect across a cohort with feature distribution $\mathbb{P}$:

$$ATE = \mathbb{E}_{X\sim\mathbb{P}}[\mathbb{E}[Y_1(X)|X] - \mathbb{E}[Y_0(X)|X]] \tag{3}$$

Example 3 (Setting up a causal analysis for effect of 6.3720 on salary)**.** What is effect of taking 6.3720 on a student's first post-graduation salary? Define the treatment, cohort, and covariates to answer this causal question. Show that your set-up satisfies the 3 assumptions.

Solution:

- Treatment: $T = 1$ can be defined as passing 6.3720. $T = 0$ can be defined as never enrolled or attended 6.3720. This definition satisfies causal consistency since everyone who received $T = 1$ took the entire class, and everyone who received $T = 0$ did not take the class at all. For no interference, we assume that 6.3720 students did not teach the course material to their friends outside the class.
- Outcome: We can define Y to be the total compensation in the first year after receiving the bachelor's degree.

- Features: To satisfy no unobserved confounding, some examples of features include major, prior coursework, and prior internships.
- Cohort: We can restrict to course 6 undergrads so students who do or do not take 6.3720 are more similar.

3 Experimentation: A/B Testing

When we have the ability to assign a treatment to each person, we can conduct an experiment to measure the average treatment effect and assess whether the effect is statistically significant. The average treatment effect is the difference between the mean outcomes in the treated and control groups. A classic example of an experiment is a clinical trial. Some patients are randomly assigned to receive the new treatment, while other patients are randomly assigned to receive the current standard of care or a placebo. Then, the outcomes in the treated and control groups are measured, and the new treatment is only approved if the difference is statistically significant.

3.1 Randomization

Random treatment assignment is an essential part of experiments because it ensures the causal assumptions are satisfied. For a feature to be a confounder, it must affect both the choice of treatment and the outcome. During randomization, no feature has an effect on the choice of treatment, so there are no confounders. In fact, we do not need to account for features when measuring the average treatment effect. Features may still be useful though when considering what kinds of people to include in the experiment. The overlap assumption is also satisfied perfectly. The feature distributions in the treated and control groups will be similar because every person has an equal probability of receiving each treatment.

Even when experiment protocols are closely followed, the assumptions may end up violated due to factors outside the control of the experimenter. One challenge is non-compliance to treatments. Subjects who are assigned to the treated cohort may opt to discontinue treatment due to adverse side effects. Alternatively, subjects who are assigned to the control cohort may seek treatment elsewhere. Even if the actual treatments are recorded, we cannot simply move a person to a different treatment group as the treatments would no longer be random and the assumptions may no longer be satisfied. If the experiment results are treated as observational data, one approach that can be used to analyze the data is instrumental variables. Intention to treat is an instrument for the actual treatment. Another challenge is non-random withdrawal from the trial. Subjects may choose to leave the trial for different reasons depending on which cohort they were assigned to. It is important to try to mitigate these issues when designing an experiment.

Example 4 (Setting up an experiment for Netflix recommendations)**.** Let's design an experiment to measure the effect of showing recommendations of shows to continue watching vs recommendations of new items to watch on time spent on Netflix.

1. How would you define the treatments?
2. How would you measure the outcome?
3. What restrictions do we want to place on the cohort?

Solution:

1. The main criterion for the treatments is consistency across all users. For example, we can define $T = 0$ as showing the 5 most recent shows at the top of the home screen every time a user opens

Netflix for 1 week. We can define $T = 1$ as showing the top 5 unseen recommendations at the top of the home screen during the same time period.

2. The main criterion for the outcome is we start measuring the outcome after treatment starts. For example, Y can be the number of minutes watched that week.

3. The two criteria to consider for the cohort are 1) the treatments must be possible and 2) who we expect to see an effect for. For example, we can include only users who have at least 5 ongoing shows and who opened Netflix at least once that week.

3.2 A/B Test

To determine whether the results from an experiment are statistically significant, we can run a two-sample t-test. The null hypothesis is the causal effect is 0, while the alternative hypothesis is the causal effect is positive. This is also known as an A/B test, where the default treatment is A and the new treatment being tested is B.

Example 5 (A/B test for Netflix experiment)**.** Suppose these were the results of the Netflix experiment proposed in Example 4. Is the effect of recommending new shows significantly greater than the effect of recommending continuations?

Treatment group	# people	Mean minutes watched	Std dev (minutes)
Continuing recs ($T = 0$)	1000	240	30
New recs ($T = 1$)	1000	300	60

Solution: We can perform a two-sample t-test where the null hypothesis is no difference in minutes watched between the two groups, and the alternative hypothesis is the group that received recommendations for new shows spent more time watching Netflix. The t-statistic is

$$t = \frac{300 - 240}{\sqrt{\frac{30^2}{1000} + \frac{60^2}{1000}}} \approx 28.3 \tag{4}$$

The p-value is close to 0, so recommending new shows is significantly more effective at increasing time spent watching Netflix.

Example 6 (Herceptin clinical trial)**.** Herceptin is a treatment for HER2-positive metastatic breast cancer. When it was in clinical trial, patients were given either herceptin and chemotherapy or only chemotherapy. In this example, we will look at one of the 4 endpoints that was measured: overall response rate—the proportion of patients whose tumors were destroyed or significantly reduced by treatment.[a] A new drug can only be approved if patients who take it have significantly better than outcomes. Based on these results, should Herceptin be approved?

Treatment group	Response	No response
Herceptin + chemotherapy	45	190
Only chemotherapy	29	205

Solution: We can perform an unpaired two-sample t-test with unequal variances. The t-statistic is

$$T = \frac{\widehat{\mu}_H - \widehat{\mu}_C}{\sqrt{\frac{\widehat{\sigma}_H^2}{n_H} + \frac{\widehat{\sigma}_C^2}{n_C}}} \tag{5}$$

Computing each statistic from the data:

$$\widehat{\mu}_H = \frac{45}{45+190} \approx .1915 \tag{6}$$

$$\widehat{\sigma^2}_H \approx \frac{45\,(1-.1915)^2 + 190\,(.1915)^2}{234} \approx .1555 \tag{7}$$

$$\widehat{\mu}_C = \frac{29}{29+205} \approx .1239 \tag{8}$$

$$\widehat{\sigma^2}_C \approx \frac{29\,(1-.1239)^2 + 205\,(.1239)^2}{233} \approx .1090 \tag{9}$$

Plugging these values in gives $T \approx 2.01$. The p-value is approximately 0.02. Based on these results, Herceptin was indeed approved.

[a] `https://www.gene.com/download/pdf/herceptin_prescribing.pdf`

3.3 Bonus: Sample Size Calculations

A critical step when designing an A/B testing procedure is determining the sample size prior to starting the experiment. Cycling between collecting more data and assessing whether the existing data is sufficient to conclude the effect is significant is not a valid approach. Because of the multiple comparison problem, this approach would likely lead to false rejection of the null hypothesis. Thus, the sample size must be set before starting the experiment. If the sample size is too small, the experiment may not be able to detect a significant result. If the sample size is too large, the experiment may be prohibitively expensive. To find a good balance, we can estimate the number of samples required by considering the expected effect size, the significance threshold for rejection, and the desired power of the test for detecting an effect if one exists.

Concept: Sample size calculation

When designing a test, we want to set the sample size to achieve a minimum required power while maintaining a desired type I error rate. The sample size can be estimated as follows:

1. Write the test statistic in terms of sample parameters (such as $\widehat{\mu}$ and $\widehat{\sigma}^2$) and sample size. These sample parameters will need to be estimated based on domain knowledge.
2. Compute the threshold for the test statistic required to achieve a type I error rate α using the distribution of the test statistic under the null.
3. Define the region of thresholds that achieve the desired power β using the distribution of the test statistic under the alternative. The distribution under the alternative will also need to be estimated based on domain knowledge.
4. Plug the threshold from step 2 into the inequality from step 3 to compute the sample size.

Deriving the required sample size calculation is a good way to improve your understanding of power from lecture 22. Let's illustrate this procedure with a two-sample t-test:

Example 7 (A/B test sample size calculation)**.** Many experiments involve running a two-sample t-test with unequal variances $\widehat{\sigma}_T^2$ and $\widehat{\sigma}_C^2$ in the treated (T) and control (C) cohorts, respectively. Let the null hypothesis H be $\mu_T - \mu_C = 0$. Let the alternative hypothesis K be $\mu_T - \mu_C = \gamma$, an effect size that is considered meaningful. Assume γ, $\widehat{\sigma}_T^2$, and $\widehat{\sigma}_C^2$ are quantities we can estimate based on domain knowledge.
The experiment will assign the same number of people to the treated and control group. If each group will have n people, how large does sample size n need to be for the test to have size α and power to be at least β?

Solution: As given in lecture 20, the t-statistic for a two-sample t-test is

$$T = \frac{\widehat{\mu}_T - \widehat{\mu}_C}{\sqrt{\frac{\widehat{\sigma}_T^2}{n} + \frac{\widehat{\sigma}_C^2}{n}}} \tag{10}$$

Instead of setting a threshold for the t-statistic, we will set a threshold d_{thr} for the difference between means so that the threshold is a function of n. A test has size α if under the null distribution $\mathbb{P}$

$$\mathbb{P}\left(\widehat{\mu}_T - \widehat{\mu}_C \geq d_{thr}\right) = \alpha \tag{11}$$

As the asymptotic distribution of the t-statistic under the null hypothesis is $\mathcal{N}(0,1)$, $\mathbb{P}$ is $\mathcal{N}\left(0, \frac{\widehat{\sigma}_T^2}{n} + \frac{\widehat{\sigma}_C^2}{n}\right)$. Let $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. The threshold for a size-α test is

$$d_{thr} = z_{1-\alpha}\sqrt{\frac{\widehat{\sigma}_T^2}{n} + \frac{\widehat{\sigma}_C^2}{n}} \tag{12}$$

Now let's consider the alternative distribution $\mathbb{Q}$. For this test to have power at least β, the threshold must satisfy

$$\mathbb{Q}\left(\widehat{\mu}_T - \widehat{\mu}_C \geq d_{thr}\right) \geq \beta \tag{13}$$

This holds if the area to the right of the Gaussian with mean shifted to γ is at least $1-\beta$. Let $z_{1-\beta} = \Phi^{-1}(1-\beta)$.

$$d_{thr} \leq \gamma + z_{1-\beta}\sqrt{\frac{\widehat{\sigma}_T^2}{n} + \frac{\widehat{\sigma}_C^2}{n}} \tag{14}$$

Solving for n gives the following minimum required sample size:

$$n \geq \frac{\left(z_{1-\alpha} - z_{1-\beta}\right)^2\left(\widehat{\sigma}_T^2 + \widehat{\sigma}_C^2\right)}{\gamma^2} \tag{15}$$

Experiments such as clinical trials are expensive to run. If the minimum required sample size is prohibitively large, some approaches can be considered to reduce sample size without compromising on power. One approach is to measure the effect on a different outcome that has larger effect size γ. If multiple treatments are being tested, another approach is to run an adaptive trial. Based on the confidence bounds for multi-arm bandits in lecture 17, a treatment can be eliminated from a trial when its effect is almost guaranteed to be worse than another treatment.

3.4 Summary: Experimentation

Practical guideline: Experiment procedure

To summarize, these are the main steps when running an experiment:

1. Define consistent treatments, outcomes, and the cohort.
2. Determine the number of people to enroll in the experiment.
3. Randomly assign each person to a treatment group.
4. Measure the outcomes in each group.
5. Run a two-sample t-test to determine whether the effect size is significant.

4 Observational Studies

Running an experiment is not always feasible due to cost, safety, or ethical reasons. For instance, in Example 3 where we designed a study to analyze the effect of taking 6.3720 on post-graduation salary, we could not run an experiment: We cannot randomly require some students to take 6.3720 and randomly prohibit other students from taking the class. In scenarios like this, we can leverage data from treatments that were given in the past to estimate treatment effect. Because treatments are no longer randomly assigned, we need to be more cautious about ensuring the assumptions specified in Section 2.2 hold.

4.1 Covariate Adjustment: Regression

Concept: Covariate adjustment: Estimating CATE with a regression

To estimate the conditional average treatment effect from observational data with a regression model:

1. Define treatment levels that are consistent across the entire population. Select all confounders as features. Define a cohort with overlap between the treated and control groups and no interference between members.
2. Choose a hypothesis class $\mathcal{F}$ for $Y_T = f(X, T)$. Any family of regression functions, such as a linear regression, causal forest, or neural network, can work for $\mathcal{F}$. Estimate $\hat{f}$.
3. To estimate the conditional average treatment effect of changing from treatment $T = 0$ to $T = 1$ for a person with features $X = x$, we can use the fitted regression

$$CATE = \hat{f}(x, 1) - \hat{f}(x, 0) \tag{16}$$

   Note that x must fall within the cohort definition from step 4. Otherwise, the prediction may be incorrect due to extrapolation.

4. To estimate the average treatment effect across samples $i = 1, \ldots, n$,

$$ATE = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}\left(x^{(i)}, 1\right) - \hat{f}\left(x^{(i)}, 0\right) \right) \tag{17}$$

To understand why the causal assumptions are essential, let us consider the special case of linear regression. In the unit on regression, we cautioned against claiming the coefficient was causal. In particular,

we said the coefficient reflected the *association* between a feature and the outcome holding other features constant. However, when we look at step 4, we see that the coefficient on the treatment term is the average treatment effect. Why does this hold under the 3 causal assumptions?

Let γ be the coefficient on T.

$$\gamma = \hat{f}(x,1) - \hat{f}(x,0) \tag{18}$$

$$\overset{(a)}{=} \underbrace{\mathbb{E}[Y|X=x,T=1] - \mathbb{E}[Y|X=x,T=0]}_{\text{Observed difference in outcomes}} \tag{19}$$

$$\overset{(b)}{=} \underbrace{\mathbb{E}[Y_1|X=x,T=1] - \mathbb{E}[Y_0|X=x,T=0]}_{\text{Difference in potential outcomes within treatment groups}} \tag{20}$$

$$\overset{(c)}{=} \underbrace{\mathbb{E}[Y_1|X=x] - \mathbb{E}[Y_0|X=x]}_{\text{Causal effect}} \tag{21}$$

Equality (a) follows from how the best prediction for a quadratic loss is the conditional mean. Overlap is required to estimate the two quantities in line 19 for all values of x. Equality (b) uses the causal consistency assumption so that the factual outcomes are the potential outcomes. Equality (c) depends on no unobserved confounding. $Y_0, Y_1 \perp\!\!\!\perp T|X$ implies $\mathbb{E}[Y_t|X=x,T] = \mathbb{E}[Y_t|X=x]$. Thus, all 3 assumptions are required for the coefficient to be causal.

To understand what happens when the no unobserved confounding assumption is violated, we can decompose the quantity in line 20:

$$\gamma = \underbrace{\mathbb{E}[Y_1|X=x,T=1] - \mathbb{E}[Y_0|X=x,T=0]}_{\text{Difference in potential outcomes within treatment groups}} \tag{22}$$

$$= \underbrace{\mathbb{E}[Y_1|X=x,T=1] - \mathbb{E}[Y_0|X=x,T=1]}_{\text{Conditional average treatment effect on the treated}} + \underbrace{\mathbb{E}[Y_0|X=x,T=1] - \mathbb{E}[Y_0|X=x,T=0]}_{\text{Selection bias}} \tag{23}$$

Because the treated and control groups are inherently different when confounders are present, they may have different CATEs. The selection bias term reflects how people who may have more adverse outcomes if left untreated are more likely to be given the treatment. This term is non-zero if X does not include all confounders.

4.2 Covariate Matching: Nearest Neighbors

Selecting a function f that reflects the true causal model is one challenge in causal analyses. Instead of using a parametric function, we may be able to compute the average treatment effect by taking the difference between the average outcomes in the treated and control cohorts if the two cohorts are similar enough. Another way to analyze observational data is to build a restricted cohort where the two groups are similar.

Concept: Covariate matching

The goal of covariate matching is to find a hypothetical twin for each treated sample among the control samples.

1. For each treated sample, find the nearest control sample. If the two samples are close enough, add the pair to the restricted cohort. Once a control sample has been added, it cannot be selected as the neighbor for another treated sample. This pairing is illustrated in Figure 3.
2. Discard treated samples without close neighbors and unmatched control samples.
3. Estimate the average treatment effect as the difference between the average factual outcomes

in the restricted cohort:

$$ATE = \frac{1}{n_1} \sum_{i:T^{(i)}=1} Y^{(i)} - \frac{1}{n_0} \sum_{i:T^{(i)}=0} Y^{(i)} \tag{24}$$

where n_1 and n_0 are the number of samples in the restricted treated and control cohorts, respectively.

4. When reporting the average treatment effect, describe the restricted cohort. **The estimate only applies to the restricted cohort.**

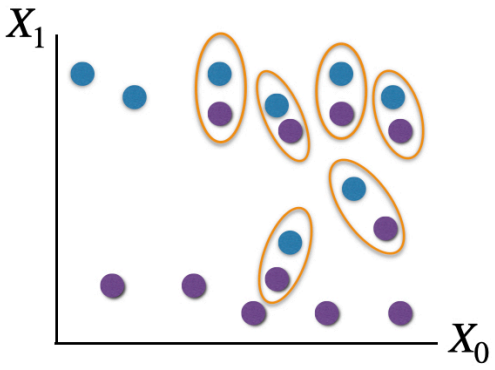

Figure 3: Covariate matching selects the closest control sample to each treated sample. Only the samples enclosed in the orange ellipses are included in the restricted cohort.

Example 8 (Covariate matching for Netflix recommendations)**.** Let's evaluate the effect of recommending new Netflix shows versus continuations using the observational data below.

Viewer	Genre	Prior # hours	Rec	Post # hours
1	Action	8	New	14
2	Action	10	New	10
3	Comedy	5	New	5
4	Comedy	7	New	9
5	Horror	2	New	1
6	Action	6	Continue	8
7	Action	10	Continue	8
8	Comedy	9	Continue	8
9	Comedy	12	Continue	10
10	Horror	10	Continue	2

Which viewers would you pair up for covariate matching based on genre and prior number of hours? What is the average treatment effect?

Solution: We can pair up viewers 1 & 6, 2 & 7, and 4 & 8. The other viewers would be discarded.

The average treatment effect is

$$\frac{1}{3}(14+10+9)-\frac{1}{3}(8+8+8)=3 \tag{25}$$

Recommending new shows instead of continuations leads to 3 additional hours of viewing in the following week. This estimate only applies to action and comedy viewers since no horror viewers are in the restricted cohort.

To assess whether the treated and control cohorts overlap more within the restricted cohort, we can perform a KS test for each continuous feature. The null hypothesis states the treated and control cohorts have the same feature distribution. After performing covariate matching, the KS test should no longer reject the null hypothesis.

A disadvantage of covariate matching is many control samples may be unnecessary excluded if the control cohort is much larger than the treated cohort. To use more control samples, we can perform k-to-1 matching: Similar to k-nearest neighbors, the k samples in the control cohort that are the most similar to each treated sample are included. Another alternative is to include all samples in the control cohort with higher weights given to samples more similar to the treated cohort. This is the underlying idea for another approach called inverse propensity weighting.

4.3 Summary: Observational Studies

Now that we have covered two approaches for estimating treatment effects from observational data, we will briefly discuss how to select a method. As shown in Figure 4, covariate adjustment can be used to estimate CATE or ATE, and it can handle a large set of features. Covariate matching can only be used to estimate ATE, and samples can only be matched when considering a small set of features. There are many other causal inference approaches beyond these two methods, such as inverse propensity weighting, doubly robust estimation, instrumental variables, panel data, and differences-in-differences. The choice of method depends on the type of data available.

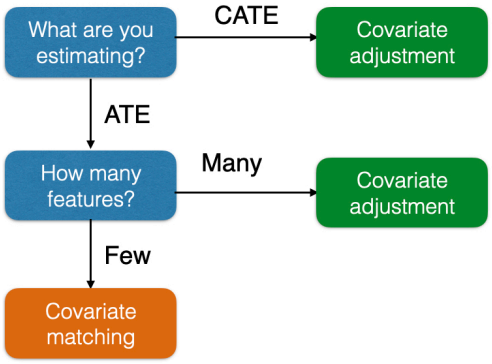

Figure 4: A framework for deciding whether to use covariate adjustment or covariate matching for observational studies. Note that other approaches exist as well.

5 Conclusion

In this lecture, we first defined the following causal question: For a person with features X, what is the effect of a treatment T on the outcome Y? Then, we emphasized the importance of carefully selecting the features, treatment, outcome, and cohort when designing a study to satisfy the following three assumptions: i) no

unobserved confounding, ii) overlap between the treated and control cohorts, and iii) consistent treatments with no interference. To answer the causal question, we can either run an experiment or an analysis with observational data. For an experiment, random treatment assignment is essential to satisfy the assumptions. Then, we can analyze whether the results are significant with a two-sample t-test. With observational data, there are many methods that can be used to analyze the data. We learned two today: We can fit a regression $Y_t = f(X, t)$ (covariate adjustment) or pair neighboring treated and control samples (covariate matching).