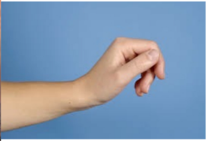
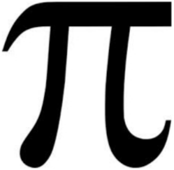
Gov 2002 - Causal Inference I

Matthew Blackwell   Arthur Spirling

September 25, 2014

Readings





What is association?

- Let's take two variables, an outcome, Y_i, and a treatment (or assignment/action), A_i.

What is association?

- Let's take two variables, an outcome, Y_i, and a treatment (or assignment/action), A_i.
- As a running example, let's use whether or not an incumbent candidate goes negative during the campaign as the treatment and the incumbent's share of the two party vote as the outcome.

What is association?

- Let's take two variables, an outcome, Y_i, and a treatment (or assignment/action), A_i.
- As a running example, let's use whether or not an incumbent candidate goes negative during the campaign as the treatment and the incumbent's share of the two party vote as the outcome.
- If Y_i and A_i are independent : $Y \perp\!\!\!\perp A$.
  $\Pr[Y = 1|A = 1] = \Pr[Y = 1|A = 0]$.

What is association?

- Let's take two variables, an outcome, Y_i, and a treatment (or assignment/action), A_i.
- As a running example, let's use whether or not an incumbent candidate goes negative during the campaign as the treatment and the incumbent's share of the two party vote as the outcome.
- If Y_i and A_i are independent : $Y \perp\!\!\!\perp A$.
  $\Pr[Y = 1|A = 1] = \Pr[Y = 1|A = 0]$.
- If the variables are not independent, we say they are dependent or associated: $\Pr[Y = 1|A = 1] \neq \Pr[Y = 1|A = 0]$.

What is association?

- Let's take two variables, an outcome, Y_i, and a treatment (or assignment/action), A_i.
- As a running example, let's use whether or not an incumbent candidate goes negative during the campaign as the treatment and the incumbent's share of the two party vote as the outcome.
- If Y_i and A_i are independent : $Y \perp\!\!\!\perp A$.
  $\Pr[Y = 1|A = 1] = \Pr[Y = 1|A = 0]$.
- If the variables are not independent, we say they are dependent or associated: $\Pr[Y = 1|A = 1] \neq \Pr[Y = 1|A = 0]$.
- Associations between variables, very famously, are not necessarily due to causation.

What is association?

- Let's take two variables, an outcome, Y_i, and a treatment (or assignment/action), A_i.
- As a running example, let's use whether or not an incumbent candidate goes negative during the campaign as the treatment and the incumbent's share of the two party vote as the outcome.
- If Y_i and A_i are independent : $Y \perp\!\!\!\perp A$.
  $\Pr[Y = 1|A = 1] = \Pr[Y = 1|A = 0]$.
- If the variables are not independent, we say they are dependent or associated: $\Pr[Y = 1|A = 1] \neq \Pr[Y = 1|A = 0]$.
- Associations between variables, very famously, are not necessarily due to causation.
- Is there a relationship between the number of swimming accidents on a given day and the total sales of ice cream cones on that day? Yes. Is that relationship causal? probably not.

Potential outcomes

- $Y_i(a = 1)$ or $Y_i(1)$ is value that Y would take if the incumbet went negative.

Potential outcomes

- $Y_i(a = 1)$ or $Y_i(1)$ is value that Y would take if the incumbet went negative.
- $Y_i(a = 0)$ or $Y_i(0)$ is the outcome when the incumbent stays positive.

Potential outcomes

- $Y_i(a = 1)$ or $Y_i(1)$ is value that Y would take if the incumbet went negative.
- $Y_i(a = 0)$ or $Y_i(0)$ is the outcome when the incumbent stays positive.
- For each unit, we observe one of these two possible potential outcomes. We can never observe both of the potential outcomes for the same unit. This is called **the fundamental problem of causal inference**.

Potential outcomes

- $Y_i(a = 1)$ or $Y_i(1)$ is value that Y would take if the incumbet went negative.
- $Y_i(a = 0)$ or $Y_i(0)$ is the outcome when the incumbent stays positive.
- For each unit, we observe one of these two possible potential outcomes. We can never observe both of the potential outcomes for the same unit. This is called **the fundamental problem of causal inference**.
- Here we have assumed that the treatment is binary, but we could generalize the potential outcomes to be a function of any value, $Y_i(a)$, where a can take any possible value.

Consistency/SUTVA

- We need some way of connecting these potential outcomes to the observed outcomes.

Consistency/SUTVA

- We need some way of connecting these potential outcomes to the observed outcomes.
- We will do this with a consistency assumption. This is what epidemiologists call it. Economists and statisticians call the "stable unit treatment value assumption".

$$Y_i(a) = Y_i \text{ if } A_i = a$$

Consistency/SUTVA

- We need some way of connecting these potential outcomes to the observed outcomes.
- We will do this with a consistency assumption. This is what epidemiologists call it. Economists and statisticians call the "stable unit treatment value assumption".

$$Y_i(a) = Y_i \text{ if } A_i = a$$

- Two key points here:

Consistency/SUTVA

- We need some way of connecting these potential outcomes to the observed outcomes.
- We will do this with a consistency assumption. This is what epidemiologists call it. Economists and statisticians call the “stable unit treatment value assumption”.

$$Y_i(a) = Y_i \text{ if } A_i = a$$

- Two key points here:

1. No interference between units.

Consistency/SUTVA

- We need some way of connecting these potential outcomes to the observed outcomes.
- We will do this with a consistency assumption. This is what epidemiologists call it. Economists and statisticians call the "stable unit treatment value assumption".

$$Y_i(a) = Y_i \text{ if } A_i = a$$

- Two key points here:

1. No interference between units.
2. Variation in the treatment is irrelevant.

A_i	Y_i	$Y_i(0)$	$Y_i(1)$
0	.63	.63	?
0	.52	.52	?
0	.55	.55	?
0	.47	.47	?
1	.49	?	.49
1	.51	?	.51
1	.43	?	.43
1	.52	?	.52

Estimands

- Suppose there are a population of units, $i = 1, \ldots, N$.

Estimands

- Suppose there are a population of units, $i = 1, \ldots, N$.
- Individual causal effect (ICE):

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

Estimands

- Suppose there are a population of units, $i = 1, \ldots, N$.
- Individual causal effect (ICE):

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

- Average treatment effect (ATE):

$$\tau = E[\tau_i] = \frac{1}{N} \sum_{i=1}^{N} Y_i(1) - Y_i(0)$$

- Conditional average treatment effect (CATE) for a subpopulation:

  $$\tau(x) = E[\tau_i | X_i = x] = \frac{1}{N_x} \sum_{i:X_i=x} Y_i(1) - Y_i(0),$$

  where N_x is the number of units in the subpopulation.

- Conditional average treatment effect (CATE) for a subpopulation:

$$\tau(x) = E[\tau_i | X_i = x] = \frac{1}{N_x} \sum_{i:X_i=x} Y_i(1) - Y_i(0),$$

where N_x is the number of units in the subpopulation.

- Average treatment effect on the treated (ATT):

$$\tau_{ATT} = E[\tau_i | A_i = 1] = \frac{1}{N_t} \sum_{i:A_i=1} Y_i(1) - Y_i(0),$$

where $N_t = \sum_i A_i$.

What is identification?

- If I gave you the entire population, so there is no sampling variation, could you estimate this quantity?

What is identification?

- If I gave you the entire population, so there is no sampling variation, could you estimate this quantity?
- **Nonparametric identification** means that we could estimate the parameter without placing any parametric models on the distribution of the data.

What is identification?

- If I gave you the entire population, so there is no sampling variation, could you estimate this quantity?
- **Nonparametric identification** means that we could estimate the parameter without placing any parametric models on the distribution of the data.
- **Parametric identification** generally refers to the situation where the estimand is identified under a certain parametric model for the distribution of the data, but is not identified otherwise.

What is identification?

- If I gave you the entire population, so there is no sampling variation, could you estimate this quantity?
- **Nonparametric identification** means that we could estimate the parameter without placing any parametric models on the distribution of the data.
- **Parametric identification** generally refers to the situation where the estimand is identified under a certain parametric model for the distribution of the data, but is not identified otherwise.
- A Heckman selection model is parametrically identified because estimating the causal effect in that case relies on the parametric assumption of normaly distributed errors.

Key to causal inference

- Data + assumptions = causal inference

Key to causal inference

- Data + assumptions = causal inference
- "What's your identification strategy?" = what are the assumptions that allow you to claim you've estimated a causal effect?

Key to causal inference

- Data + assumptions = causal inference
- "What's your identification strategy?" = what are the assumptions that allow you to claim you've estimated a causal effect?
- Estimation method (regression, matching, weighting, 2SLS, 3SLS, SEM, GMM, GEE, dynamic panel, etc) are secondary to the identification assumptions.

What is the selection problem?

- Start with *prima facie* effect, which is just the difference in means between those who take a treatment and those who do not.

$$E[Y_i|A_i=1]-E[Y_i|A_i=0]=E[Y_i(1)|A_i=1]-E[Y_i(0)|A_i=0]$$

What is the selection problem?

- Start with *prima facie* effect, which is just the difference in means between those who take a treatment and those who do not.

$$E[Y_i|A_i = 1] - E[Y_i|A_i = 0] = E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 0]$$

What is the selection problem?

- Start with *prima facie* effect, which is just the difference in means between those who take a treatment and those who do not.

$$\begin{aligned} E[Y_i|A_i=1]-E[Y_i|A_i=0] &= E[Y_i(1)|A_i=1]-E[Y_i(0)|A_i=0] \\ &= E[Y_i(1)|A_i=1]-E[Y_i(0)|A_i=1] \\ &+ E[Y_i(0)|A_i=1]-E[Y_i(0)|A_i=0] \end{aligned}$$

- The second line here is the average treatment effect on the treated

What is the selection problem?

- Start with *prima facie* effect, which is just the difference in means between those who take a treatment and those who do not.

$$\begin{aligned} E[Y_i|A_i=1]-E[Y_i|A_i=0] &= E[Y_i(1)|A_i=1]-E[Y_i(0)|A_i=0] \\ &= E[Y_i(1)|A_i=1]-E[Y_i(0)|A_i=1] \\ &+ E[Y_i(0)|A_i=1]-E[Y_i(0)|A_i=0] \end{aligned}$$

- The second line here is the average treatment effect on the treated
- The third line is what we call *selection bias*.

What is the selection problem?

- Start with *prima facie* effect, which is just the difference in means between those who take a treatment and those who do not.

$$\begin{aligned} E[Y_i|A_i=1]-E[Y_i|A_i=0] &= E[Y_i(1)|A_i=1]-E[Y_i(0)|A_i=0] \\ &= E[Y_i(1)|A_i=1]-E[Y_i(0)|A_i=1] \\ &+ E[Y_i(0)|A_i=1]-E[Y_i(0)|A_i=0] \end{aligned}$$

- The second line here is the average treatment effect on the treated
- The third line is what we call *selection bias*.
- Because of the selection bias, without further assumptions we say that the ATT is *unidentified*.

Randomization solves the selection problem

- Randomizing the treatment means that the treated group is a random sample from the population and the in-sample mean is equal to overall mean:

$$E[Y_i(0)|A_i = 0] = E[Y_i(0)] = E[Y_i(0)|A_i = 1]$$

Randomization solves the selection problem

- Randomizing the treatment means that the treated group is a random sample from the population and the in-sample mean is equal to overall mean:

$$E[Y_i(0)|A_i = 0] = E[Y_i(0)] = E[Y_i(0)|A_i = 1]$$

- Being a random sample, we know that those included in the sample are the same, on average, as those not included in the sample on any measure.

Randomization solves the selection problem

- Randomizing the treatment means that the treated group is a random sample from the population and the in-sample mean is equal to overall mean:

$$E[Y_i(0)|A_i = 0] = E[Y_i(0)] = E[Y_i(0)|A_i = 1]$$

- Being a random sample, we know that those included in the sample are the same, on average, as those not included in the sample on any measure.
- Specifically, randomization implies **ignorability**, which means the potential outcomes are independent of the treatments. We write ignorability like this:

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i$$

Randomization solves the selection problem

- Randomizing the treatment means that the treated group is a random sample from the population and the in-sample mean is equal to overall mean:

$$E[Y_i(0)|A_i = 0] = E[Y_i(0)] = E[Y_i(0)|A_i = 1]$$

- Being a random sample, we know that those included in the sample are the same, on average, as those not included in the sample on any measure.
- Specifically, randomization implies **ignorability**, which means the potential outcomes are independent of the treatments. We write ignorability like this:

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i$$

- This is not the same as the treatment being independent of the observed outcomes ($Y_i \perp\!\!\!\perp A_i$).

- How does randomization help indentify the causal effect? It ensures that there is no selection bias Note that, because of ingorability:

$$E[Y_i(0)|A_i = 1] - E[Y_i(0)|A_i = 0] = E[Y_i(0)] - E[Y_i(0)] = 0$$

- How does randomization help indentify the causal effect? It ensures that there is no selection bias Note that, because of ingorability:

$$E[Y_i(0)|A_i = 1] - E[Y_i(0)|A_i = 0] = E[Y_i(0)] - E[Y_i(0)] = 0$$

- Plugging this in above gives us:

$$\begin{aligned} E[Y_i|A_i = 1] - E[Y_i|A_i = 0] &= E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1] + \\ &= E[Y_i(1)] - E[Y_i(0)] = \tau \end{aligned}$$

- How does randomization help indentify the causal effect? It ensures that there is no selection bias Note that, because of ingorability:

$$E[Y_i(0)|A_i = 1] - E[Y_i(0)|A_i = 0] = E[Y_i(0)] - E[Y_i(0)] = 0$$

- Plugging this in above gives us:

$$\begin{aligned} E[Y_i|A_i = 1] - E[Y_i|A_i = 0] &= E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1] + \\ &= E[Y_i(1)] - E[Y_i(0)] = \tau \end{aligned}$$

- Thus, the ATE is nonparametrically identified: no matter what assumptions we make about the distribution of Y, we can always estimate it with the difference in means.

Samples versus Populations

- Above we defined all of the estimands in terms of the population. The ATE, τ_{ATE} was the average over the population, $\{1, \ldots, N\}$. Same with the ATT.

Samples versus Populations

- Above we defined all of the estimands in terms of the population. The ATE, τ_{ATE} was the average over the population, $\{1, \ldots, N\}$. Same with the ATT.
- Sometimes instead of making inference about a population, we would rather make inference about the sample that we actually observed.

Samples versus Populations

- Above we defined all of the estimands in terms of the population. The ATE, τ_{ATE} was the average over the population, $\{1, \ldots, N\}$. Same with the ATT.
- Sometimes instead of making inference about a population, we would rather make inference about the sample that we actually observed.
- This might make more sense in a lot of political science, where we don't have a larger super population in mind. This is similar to the arguments made about Bayesian inference.

- Suppose that we have a sample, S, of units, $i = 1, \ldots, n$ where n_t of the units are treated.

- Suppose that we have a sample, S, of units, $i = 1, \ldots, n$ where n_t of the units are treated.
- For this, we can define the **sample average treatment effect (SATE)** as the in-sample average of the potential outcomes:

$$SATE = \tau_S = \frac{1}{n} \sum_{i \in S} Y_i(1) - Y_i(0)$$

- Suppose that we have a sample, S, of units, $i = 1, \ldots, n$ where n_t of the units are treated.
- For this, we can define the **sample average treatment effect (SATE)** as the in-sample average of the potential outcomes:

$$SATE = \tau_S = \frac{1}{n} \sum_{i \in S} Y_i(1) - Y_i(0)$$

- The SATE is the in-sample version of the ATE (which we sometimes call the PATE to distinguish it from the SATE) and for any given sample, won't equal the PATE.

- Suppose that we have a sample, S, of units, $i = 1, \ldots, n$ where n_t of the units are treated.
- For this, we can define the **sample average treatment effect (SATE)** as the in-sample average of the potential outcomes:

$$SATE = \tau_S = \frac{1}{n} \sum_{i \in S} Y_i(1) - Y_i(0)$$

- The SATE is the in-sample version of the ATE (which we sometimes call the PATE to distinguish it from the SATE) and for any given sample, won't equal the PATE.
- SATE varies over samples from the population. What's this distribution called?

- If we are ignore the population here and condition on the sample, why is there any uncertainty in our estimate of the SATE? There is uncertainty in how the treatment was assigned.

- If we are ignore the population here and condition on the sample, why is there any uncertainty in our estimate of the SATE? There is uncertainty in how the treatment was assigned.
- The usual difference in means estimator can consistently estimate both the ATE and SATE, but the variance of that estimator is smaller when estimating the SATE. This makes sense as we are treating the sample as fixed, so that variation doesn't enter into the sampling distribution.

- If we are ignore the population here and condition on the sample, why is there any uncertainty in our estimate of the SATE? There is uncertainty in how the treatment was assigned.
- The usual difference in means estimator can consistently estimate both the ATE and SATE, but the variance of that estimator is smaller when estimating the SATE. This makes sense as we are treating the sample as fixed, so that variation doesn't enter into the sampling distribution.
- Unfortunately, it is usually impossible to estimate the variance of the sampling distribution for the SATE, but we know it's smaller than the variance for the ATE, so we can use that as a conservative estimator.

Observational studies and confounding

- An **observational study** is a study where the researcher *does not control the treatment assignment.*

Observational studies and confounding

- An **observational study** is a study where the researcher *does not control the treatment assignment.*
- No guarantee that the treatment and control groups are comparable.

Observational studies and confounding

- An **observational study** is a study where the researcher *does not control the treatment assignment.*
- No guarantee that the treatment and control groups are comparable.
- Need to justify our claims by assumption and by theory instead of by direct manipulation.

Selection on observables

- If ignorability doesn't hold by default, then what can we do?

Selection on observables

- If ignorability doesn't hold by default, then what can we do?
- Find a set of covariates such that, conditional on those covariates, the treatment is "as-if" randomized. Plausible?

Selection on observables

- If ignorability doesn't hold by default, then what can we do?
- Find a set of covariates such that, conditional on those covariates, the treatment is "as-if" randomized. Plausible?
- Basically, it says that selection into treatment is based only on observable data, X. Or, more specifically,

$$Y(a) \perp\!\!\!\perp A | X$$

Selection on observables

- If ignorability doesn't hold by default, then what can we do?
- Find a set of covariates such that, conditional on those covariates, the treatment is "as-if" randomized. Plausible?
- Basically, it says that selection into treatment is based only on observable data, X. Or, more specifically,

$$Y(a) \perp\!\!\!\perp A | X$$

- There are many names for this assumption and they vary by discipline. It is "selection on the observables" in economics, "no unmeasured confounders" in epidemiology, "exchangability" or "ingorability" in statistics, and "no omitted variables" in political science.

Selection on observables

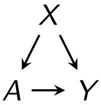
- If ignorability doesn't hold by default, then what can we do?
- Find a set of covariates such that, conditional on those covariates, the treatment is "as-if" randomized. Plausible?
- Basically, it says that selection into treatment is based only on observable data, X. Or, more specifically,

$$Y(a) \perp\!\!\!\perp A|X$$

- There are many names for this assumption and they vary by discipline. It is "selection on the observables" in economics, "no unmeasured confounders" in epidemiology, "exchangability" or "ingorability" in statistics, and "no omitted variables" in political science.
- How can we figure out if ignorability holds in some case? Untestable, but can infer from other assumptions. . .

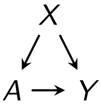
DAGs

- We can encode assumptions about causal relationships in what are called causal Directed Acyclic Graphs or DAGs. Here is an example:



DAGs

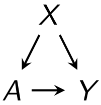
- We can encode assumptions about causal relationships in what are called causal Directed Acyclic Graphs or DAGs. Here is an example:



- Each arrow represents the presence of a direct causal effect (that is, an individual causal effect as above). The lack of an arrow represents the lack of a causal effect.

DAGs

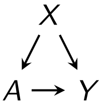
- We can encode assumptions about causal relationships in what are called causal Directed Acyclic Graphs or DAGs. Here is an example:



- Each arrow represents the presence of a direct causal effect (that is, an individual causal effect as above). The lack of an arrow represents the lack of a causal effect.
- These are *directed* because each arrow implies a direction (aspirin causes pain relief, not the other way around).

DAGs

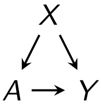
- We can encode assumptions about causal relationships in what are called causal Directed Acyclic Graphs or DAGs. Here is an example:



- Each arrow represents the presence of a direct causal effect (that is, an individual causal effect as above). The lack of an arrow represents the lack of a causal effect.
- These are *directed* because each arrow implies a direction (aspirin causes pain relief, not the other way around).
- They are *acyclic* because there are no cycles: a variable cannot cause itself, either directly or through cycles.

DAGs

- We can encode assumptions about causal relationships in what are called causal Directed Acyclic Graphs or DAGs. Here is an example:



- Each arrow represents the presence of a direct causal effect (that is, an individual causal effect as above). The lack of an arrow represents the lack of a causal effect.
- These are *directed* because each arrow implies a direction (aspirin causes pain relief, not the other way around).
- They are *acyclic* because there are no cycles: a variable cannot cause itself, either directly or through cycles.
- Causal Markov assumption: conditional on its direct causes, a variable V_j is independent of its non-descendents.

Causal DAGs and associations.

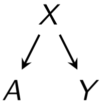
- DAGs are a convenient way to encode causal assumptions about the problem at hand, but they also can tell us about potential associations between variables in the graph.

Causal DAGs and associations.

- DAGs are a convenient way to encode causal assumptions about the problem at hand, but they also can tell us about potential associations between variables in the graph.
- A *path* between two variables (C and D) in a DAG is a route that connects the variables following nonintersecting edges.

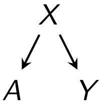
Causal DAGs and associations.

- DAGs are a convenient way to encode causal assumptions about the problem at hand, but they also can tell us about potential associations between variables in the graph.
- A *path* between two variables (C and D) in a DAG is a route that connects the variables following nonintersecting edges.
- A path is causal if those edges all have their arrows pointed in the same direction. Otherwise it is noncausal.



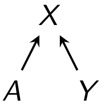
Causal DAGs and associations.

- DAGs are a convenient way to encode causal assumptions about the problem at hand, but they also can tell us about potential associations between variables in the graph.
- A *path* between two variables (C and D) in a DAG is a route that connects the variables following nonintersecting edges.
- A path is causal if those edges all have their arrows pointed in the same direction. Otherwise it is noncausal.

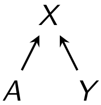


- Two variables connected by common causes will have a marginal associational relationship. That is, in the above example $\Pr[Y = 1|A = 1] \neq \Pr[Y = 1|A = 0]$.

- Let's look at another situation:

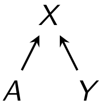


- Let's look at another situation:



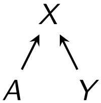
- Here, X is a *collider*: a node that two arrows point into.

- Let's look at another situation:



- Here, X is a *collider*: a node that two arrows point into.
- Are A and Y related? No.

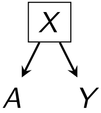
- Let's look at another situation:



- Here, X is a *collider*: a node that two arrows point into.
- Are A and Y related? No.
- Imagine that A is getting the flu and Y is getting hit by a bus. Both of these might cause us to be in the hospital, but knowing that I have the flu doesn't give me any information about whether or not I've been hit by a bus. The flow of association is blocked by a collider.

Conditioning on a confounder

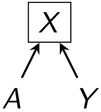
- Above we have shown how marginal associations flow over paths, but what about relationships between variables within levels of a third variable? We can represent conditioning on a variable by drawing a box around it.

  Conditioning on a variable is on a causal path or on a variable that is a common cause (above), will block the association that flows over that path.

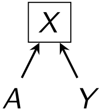
Conditioning on a collider

- Conditioning on a collider (a common consequence) actually opens the flow of association over that path, even though before there was none:



Conditioning on a collider

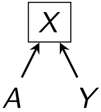
- Conditioning on a collider (a common consequence) actually opens the flow of association over that path, even though before there was none:



- To see why this is the case, let's go back to the flu, getting hit by a bus example.

Conditioning on a collider

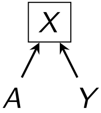
- Conditioning on a collider (a common consequence) actually opens the flow of association over that path, even though before there was none:



- To see why this is the case, let's go back to the flu, getting hit by a bus example.
- Conditional on being in the hospital, there is a negative relationship between the flu and getting hit by a bus.

Conditioning on a collider

- Conditioning on a collider (a common consequence) actually opens the flow of association over that path, even though before there was none:



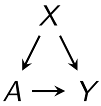
- To see why this is the case, let's go back to the flu, getting hit by a bus example.
- Conditional on being in the hospital, there is a negative relationship between the flu and getting hit by a bus.
- To sum up: associations flow over paths (causal or noncausal) that don't contain a collider. These associations can be blocked by conditioning a variable on the path that is not a collider. We'll come back to these properties later when we talk about the back-door criteria.

Backdoor paths and blocking paths

- A *backdoor path* is a non-causal path from A to Y. This is a path that would remain if we were to remove any arrows pointing out of A (these are the potentially causal paths from A, sometimes called *frontdoor paths*).

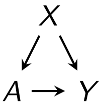
Backdoor paths and blocking paths

- A *backdoor path* is a non-causal path from A to Y. This is a path that would remain if we were to remove any arrows pointing out of A (these are the potentially causal paths from A, sometimes called *frontdoor paths*).
- Backdoor paths between A and Y generally indicate common causes of A and Y. The simplest possible backdoor path is the common confounding situation:



Backdoor paths and blocking paths

- A *backdoor path* is a non-causal path from A to Y. This is a path that would remain if we were to remove any arrows pointing out of A (these are the potentially causal paths from A, sometimes called *frontdoor paths*).
- Backdoor paths between A and Y generally indicate common causes of A and Y. The simplest possible backdoor path is the common confounding situation:



- Here there is a backdoor path $A \leftarrow X \rightarrow Y$, where X is a common cause for the treatment and the outcome.

- When there are unblocked backdoor paths, causal effect is muddled by spurious association.

- When there are unblocked backdoor paths, causal effect is muddled by spurious association.
- A path is *blocked* if (a) we control for or stratify a non-collider on that path OR (b) we do not control for a collider.

- When there are unblocked backdoor paths, causal effect is muddled by spurious association.
- A path is *blocked* if (a) we control for or stratify a non-collider on that path OR (b) we do not control for a collider.
- Thus, in the above sample, if we condition on X, then the backdoor path is blocked.

Backdoor criterion

- How to tell if an effect is identifiable from the graph? From Pearl ({2000}), we have the *backdoor criterion* which states that an effect of A on Y is identifiable if either:

Backdoor criterion

- How to tell if an effect is identifiable from the graph? From Pearl ({2000}), we have the *backdoor criterion* which states that an effect of A on Y is identifiable if either:

1. No backdoor paths from A to Y

Backdoor criterion

- How to tell if an effect is identifiable from the graph? From Pearl ({2000}), we have the *backdoor criterion* which states that an effect of A on Y is identifiable if either:

1. No backdoor paths from A to Y
2. Measured covariates are sufficient to block all backdoor paths from A to Y.

Backdoor criterion

- How to tell if an effect is identifiable from the graph? From Pearl ({2000}), we have the *backdoor criterion* which states that an effect of A on Y is identifiable if either:

1. No backdoor paths from A to Y
2. Measured covariates are sufficient to block all backdoor paths from A to Y.

- The first situation is only plausible in a randomized experiment, but the second might be plausible in observational studies as well.

Backdoor criterion

- How to tell if an effect is identifiable from the graph? From Pearl ({2000}), we have the *backdoor criterion* which states that an effect of A on Y is identifiable if either:

1. No backdoor paths from A to Y
2. Measured covariates are sufficient to block all backdoor paths from A to Y.

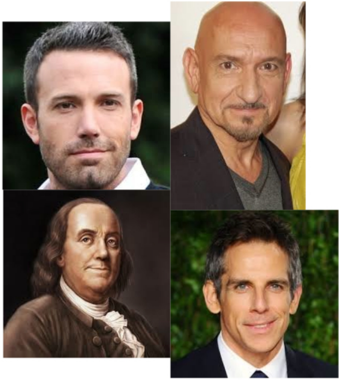
- The first situation is only plausible in a randomized experiment, but the second might be plausible in observational studies as well.
- The backdoor criterion is fairly powerful. It can tell us (1) is there confounding given this DAG, (2) if it is possible to removing the confounding, and (3) what variables to condition on to eliminate the confounding.

- How does the backdoor criterion relate to ignorability? If the graph is causal (in the sense that each of arrows represents a causal effect in the potential outcomes sense), then there is a specific relationship between the backdoor criterion and ignorability.

- How does the backdoor criterion relate to ignorability? If the graph is causal (in the sense that each of arrows represents a causal effect in the potential outcomes sense), then there is a specific relationship between the backdoor criterion and ignorability.
- “All backdoor paths blocked” $\equiv$ conditional ignorability

- How does the backdoor criterion relate to ignorability? If the graph is causal (in the sense that each of arrows represents a causal effect in the potential outcomes sense), then there is a specific relationship between the backdoor criterion and ignorability.
- “All backdoor paths blocked” $\equiv$ conditional ignorability
- **No free lunch**: DAG must be correctly specified

Readings



Estimating causal effects under no unmeasured confounders

- Another assumption we'll need here is the following overlap (or positivity) assumption: $0 < \Pr(A = 1|X) < 1$.

Estimating causal effects under no unmeasured confounders

- Another assumption we'll need here is the following overlap (or positivity) assumption: $0 < \Pr(A = 1|X) < 1$.
- The assumption of selection on the observables is what allows us to identify causal effects.

Estimating causal effects under no unmeasured confounders

- Another assumption we'll need here is the following overlap (or positivity) assumption: $0 < \Pr(A = 1|X) < 1$.
- The assumption of selection on the observables is what allows us to identify causal effects.
- But we still have to estimate them. And given ignorability, there are several choices we can make for the estimation of causal effects.

Conditional ignorability and identification

- We can identify the CATE with conditional ignorability:

$$\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$$

Conditional ignorability and identification

- We can identify the CATE with conditional ignorability:

$$\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$$

Conditional ignorability and identification

- We can identify the CATE with conditional ignorability:

$$\begin{aligned}\tau(x) &= E[Y_i(1) - Y_i(0)|X_i = x] \\ &= E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x]\end{aligned}$$

Conditional ignorability and identification

- We can identify the CATE with conditional ignorability:

$$\begin{aligned}\tau(x) &= E[Y_i(1) - Y_i(0)|X_i = x] \\ &= E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x] \\ &= E[Y_i(1)|A_i = 1, X_i = x] - E[Y_i(0)|A_i = 0, X_i = x]\end{aligned}$$

Conditional ignorability and identification

- We can identify the CATE with conditional ignorability:

$$
\begin{aligned}
\tau(x) &= E[Y_i(1) - Y_i(0)|X_i = x] \\
&= E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x] \\
&= E[Y_i(1)|A_i = 1, X_i = x] - E[Y_i(0)|A_i = 0, X_i = x] \\
&= E[Y_i|A_i = 1, X_i = x] - E[Y_i|A_i = 0, X_i = x]
\end{aligned}
$$

Conditional ignorability and identification

- We can identify the CATE with conditional ignorability:

$$\begin{aligned}
\tau(x) &= E[Y_i(1) - Y_i(0)|X_i = x] \\
&= E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x] \\
&= E[Y_i(1)|A_i = 1, X_i = x] - E[Y_i(0)|A_i = 0, X_i = x] \\
&= E[Y_i|A_i = 1, X_i = x] - E[Y_i|A_i = 0, X_i = x]
\end{aligned}$$

- We can just use the within-levels of X difference in means to estimate the CATE. There are a number of ways we could estimate those conditional expectations, though. We'll cover a few in this class.

Regression

- When we look at a textbook, we often see regression defined without respect to causality. There is talk of the $\hat{\beta}$ estimator being "biased," but it isn't always clear what the "correct" specification would look like. There is an implicit assumption of causality, but no formal definitions. This can obscure the identification of the causal effects of interest. Today, we'll see if we can estimate causal effects with regression.

Regression

- When we look at a textbook, we often see regression defined without respect to causality. There is talk of the $\hat{\beta}$ estimator being "biased," but it isn't always clear what the "correct" specification would look like. There is an implicit assumption of causality, but no formal definitions. This can obscure the identification of the causal effects of interest. Today, we'll see if we can estimate causal effects with regression.
- Angrist and Pischke argue that a regression is causal when the CEF it approximates is causal. Identification is king.

Regression

- When we look at a textbook, we often see regression defined without respect to causality. There is talk of the $\hat{\beta}$ estimator being "biased," but it isn't always clear what the "correct" specification would look like. There is an implicit assumption of causality, but no formal definitions. This can obscure the identification of the causal effects of interest. Today, we'll see if we can estimate causal effects with regression.
- Angrist and Pischke argue that a regression is causal when the CEF it approximates is causal. Identification is king.
- We will show that under certain conditions, a regression of the outcome on the treatment and the covariates can recover *a* causal parameter, but perhaps not the one in which we are interested.

Regression

- When we look at a textbook, we often see regression defined without respect to causality. There is talk of the $\hat{\beta}$ estimator being "biased," but it isn't always clear what the "correct" specification would look like. There is an implicit assumption of causality, but no formal definitions. This can obscure the identification of the causal effects of interest. Today, we'll see if we can estimate causal effects with regression.
- Angrist and Pischke argue that a regression is causal when the CEF it approximates is causal. Identification is king.
- We will show that under certain conditions, a regression of the outcome on the treatment and the covariates can recover *a* causal parameter, but perhaps not the one in which we are interested.
- We have shown in past weeks that these effects are identified when ignorability holds. Angrist and Pischke call this the conditional independence assumption (CIA).

Linear constant effects model, binary treatment

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$$

Linear constant effects model, binary treatment

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$$

Linear constant effects model, binary treatment

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

$$\begin{aligned} Y_i &= A_i Y_i(1) + (1 - A_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) A_i \end{aligned}$$

Linear constant effects model, binary treatment

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

$$\begin{aligned} Y_i &= A_i Y_i(1) + (1 - A_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) A_i \\ &= \mu^0 + \tau A_i + (Y_i(0) - \mu^0) \end{aligned}$$

Linear constant effects model, binary treatment

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

$$\begin{aligned} Y_i &= A_i Y_i(1) + (1 - A_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) A_i \\ &= \mu^0 + \tau A_i + (Y_i(0) - \mu^0) \\ &= \mu^0 + \tau A_i + v_i^0 \end{aligned}$$

Linear constant effects model, binary treatment

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

$$\begin{aligned} Y_i &= A_i Y_i(1) + (1 - A_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) A_i \\ &= \mu^0 + \tau A_i + (Y_i(0) - \mu^0) \\ &= \mu^0 + \tau A_i + v_i^0 \end{aligned}$$

- Note that if ignorability holds (as in an experiment) for $Y_i(0)$, then it will also hold for v_i^0, since μ^0 is constant. Thus, this satifies the usual assumptions for regression.

- Let's now say that ignorability holds only conditional the covariates, so $Y_i(a) \perp\!\!\!\perp A_i | X_i$. We will assume a linear model for the potential outcomes:

$$Y_i(a) = \alpha + \tau a + \eta_i$$

- Let's now say that ignorability holds only conditional the covariates, so $Y_i(a) \perp\!\!\!\perp A_i | X_i$. We will assume a linear model for the potential outcomes:

$$Y_i(a) = \alpha + \tau a + \eta_i$$

- Because we are assuming the effect of A is constant here, the η_i are the only source of individual variation and we have $E[\eta_i] = 0$. We can use the consistency assumption to write this as a linear regression model:

$$Y_i = \alpha + \tau A_i + \eta_i.$$

- Assume that η_i is linear in the covariates $\eta_i = X_i'\gamma + \nu_i$. (strong assumption)

$$E[Y_i|A_i, X_i] = E[Y_i(a)|X_i] = \alpha + \tau A_i + E[\eta_i|X_i]$$

- Assume that η_i is linear in the covariates $\eta_i = X_i'\gamma + \nu_i$. (strong assumption)

$$E[Y_i|A_i, X_i] = E[Y_i(a)|X_i] = \alpha + \tau A_i + E[\eta_i|X_i]$$

- Assume that η_i is linear in the covariates $\eta_i = X_i'\gamma + \nu_i$. (strong assumption)

$$\begin{aligned} E[Y_i|A_i, X_i] = E[Y_i(a)|X_i] &= \alpha + \tau A_i + E[\eta_i|X_i] \\ &= \alpha + \tau A_i + X_i'\gamma + E[\nu_i|X_i] \end{aligned}$$

- Assume that η_i is linear in the covariates $\eta_i = X_i'\gamma + \nu_i$. (strong assumption)

$$\begin{aligned} E[Y_i|A_i, X_i] = E[Y_i(a)|X_i] &= \alpha + \tau A_i + E[\eta_i|X_i] \\ &= \alpha + \tau A_i + X_i'\gamma + E[\nu_i|X_i] \\ &= \alpha + \tau A_i + X_i'\gamma \end{aligned}$$

- Assume that η_i is linear in the covariates $\eta_i = X_i'\gamma + \nu_i$. (strong assumption)

$$\begin{aligned} E[Y_i|A_i, X_i] = E[Y_i(a)|X_i] &= \alpha + \tau A_i + E[\eta_i|X_i] \\ &= \alpha + \tau A_i + X_i'\gamma + E[\nu_i|X_i] \\ &= \alpha + \tau A_i + X_i'\gamma \end{aligned}$$

- Thus, a regression where A_i and X_i enter linearly will correctly estimate the average treatment effect, τ, since the residual of the linear regression is independent of the covariates:

$$Y_i = \alpha + \tau A_i + X_i'\gamma + \nu_i$$

- Assume that η_i is linear in the covariates $\eta_i = X_i'\gamma + \nu_i$. (strong assumption)

$$\begin{aligned} E[Y_i|A_i, X_i] = E[Y_i(a)|X_i] &= \alpha + \tau A_i + E[\eta_i|X_i] \\ &= \alpha + \tau A_i + X_i'\gamma + E[\nu_i|X_i] \\ &= \alpha + \tau A_i + X_i'\gamma \end{aligned}$$

- Thus, a regression where A_i and X_i enter linearly will correctly estimate the average treatment effect, τ, since the residual of the linear regression is independent of the covariates:

$$Y_i = \alpha + \tau A_i + X_i'\gamma + \nu_i$$

- Note that nothing we have done changes if A_i were continuous or ordinal (so long as linearity holds)

Heterogeneous effects

- What if we allow for individual effects to vary, τ_i?

Heterogeneous effects

- What if we allow for individual effects to vary, τ_i?
- For the binary case with a randomized treatment, no problems.

Heterogeneous effects

- What if we allow for individual effects to vary, τ_i?
- For the binary case with a randomized treatment, no problems.
- When we have to condition on some variables, things get difficult.

Heterogeneous effects

- Focus on the case where X_i is univariate and binary and we can generalize from there.

$$Y_i = X_i \alpha_x + \tau_R A_i + e_i.$$

Heterogeneous effects

- Focus on the case where X_i is univariate and binary and we can generalize from there.

$$Y_i = X_i\alpha_x + \tau_R A_i + e_i.$$

- Is τ_R equal to the ATE or the ATT?

Heterogeneous effects

- Focus on the case where X_i is univariate and binary and we can generalize from there.

$$Y_i = X_i \alpha_x + \tau_R A_i + e_i.$$

- Is τ_R equal to the ATE or the ATT?
- How can we investigate τ_R? Well, we can rely on the regression anatomy:

$$\tau_R = \frac{\text{Cov}(Y_i, \tilde{A}_i)}{V(\tilde{A}_i)}$$

Heterogeneous effects

- Focus on the case where X_i is univariate and binary and we can generalize from there.

$$Y_i = X_i\alpha_x + \tau_R A_i + e_i.$$

- Is τ_R equal to the ATE or the ATT?
- How can we investigate τ_R? Well, we can rely on the regression anatomy:

$$\tau_R = \frac{\text{Cov}(Y_i, \tilde{A}_i)}{V(\tilde{A}_i)}$$

- $\tilde{A}_i$ is the residual from a regression of A_i on the X_i or $\tilde{A}_i = A_i - E[A_i|X_i]$.

Heterogeneous effects (cont'd)

- A regression of Y_i on the treatment and covariates is the same as a regression of the $E[Y_i|X_i, A_i]$ on the treatment and covariates. Thus, in the above expression, we can replace Y_i with $E[Y_i|X_i, A_i]$.

$$
\begin{aligned}
\tau_R &= \frac{\text{Cov}(E[Y_i|X_i, A_i], A_i - E[A_i|X_i])}{V(A_i - E[A_i|X_i])} \\
&= \frac{E\left\{E[Y_i|X_i, A_i](A_i - E[A_i|X_i])\right\}}{E[(A_i - E[A_i|X_i])^2]}
\end{aligned}
$$

Heterogeneous effects (cont'd)

- A regression of Y_i on the treatment and covariates is the same as a regression of the $E[Y_i|X_i, A_i]$ on the treatment and covariates. Thus, in the above expression, we can replace Y_i with $E[Y_i|X_i, A_i]$.

$$\begin{aligned}\tau_R &= \frac{\text{Cov}(E[Y_i|X_i, A_i], A_i - E[A_i|X_i])}{V(A_i - E[A_i|X_i])} \\ &= \frac{E\left\{E[Y_i|X_i, A_i](A_i - E[A_i|X_i])\right\}}{E[(A_i - E[A_i|X_i])^2]}\end{aligned}$$

- Why stop here? We can simplify the CEF a bit more:

$$E[Y_i|X_i, A_i] = E[A_iY_i(1) + (1 - A_i)Y_i(0)|X_i, A_i]$$

Heterogeneous effects (cont'd)

- A regression of Y_i on the treatment and covariates is the same as a regression of the $E[Y_i|X_i, A_i]$ on the treatment and covariates. Thus, in the above expression, we can replace Y_i with $E[Y_i|X_i, A_i]$.

$$\begin{aligned}\tau_R &= \frac{\text{Cov}(E[Y_i|X_i, A_i], A_i - E[A_i|X_i])}{V(A_i - E[A_i|X_i])} \\ &= \frac{E\left\{E[Y_i|X_i, A_i](A_i - E[A_i|X_i])\right\}}{E[(A_i - E[A_i|X_i])^2]}\end{aligned}$$

- Why stop here? We can simplify the CEF a bit more:

$$E[Y_i|X_i, A_i] = E[A_iY_i(1) + (1 - A_i)Y_i(0)|X_i, A_i]$$

Heterogeneous effects (cont'd)

- A regression of Y_i on the treatment and covariates is the same as a regression of the $E[Y_i|X_i, A_i]$ on the treatment and covariates. Thus, in the above expression, we can replace Y_i with $E[Y_i|X_i, A_i]$.

$$\begin{aligned}\tau_R &= \frac{\mathrm{Cov}(E[Y_i|X_i, A_i], A_i - E[A_i|X_i])}{V(A_i - E[A_i|X_i])} \\ &= \frac{E\left\{E[Y_i|X_i, A_i](A_i - E[A_i|X_i])\right\}}{E[(A_i - E[A_i|X_i])^2]}\end{aligned}$$

- Why stop here? We can simplify the CEF a bit more:

$$\begin{aligned}E[Y_i|X_i, A_i] &= E[A_iY_i(1) + (1 - A_i)Y_i(0)|X_i, A_i] \\ &= E[Y_i(0)|X_i, A_i = 0] + A_iE[Y_i(1) - Y_i(0)|X_i, A_i]\end{aligned}$$

Heterogeneous effects (cont'd)

- A regression of Y_i on the treatment and covariates is the same as a regression of the $E[Y_i|X_i, A_i]$ on the treatment and covariates. Thus, in the above expression, we can replace Y_i with $E[Y_i|X_i, A_i]$.

$$\begin{aligned}\tau_R &= \frac{\text{Cov}(E[Y_i|X_i, A_i], A_i - E[A_i|X_i])}{V(A_i - E[A_i|X_i])} \\ &= \frac{E\left\{E[Y_i|X_i, A_i](A_i - E[A_i|X_i])\right\}}{E[(A_i - E[A_i|X_i])^2]}\end{aligned}$$

- Why stop here? We can simplify the CEF a bit more:

$$\begin{aligned}E[Y_i|X_i, A_i] &= E[A_i Y_i(1) + (1 - A_i)Y_i(0)|X_i, A_i] \\ &= E[Y_i(0)|X_i, A_i = 0] + A_i E[Y_i(1) - Y_i(0)|X_i, A_i] \\ &= E[Y_i|X_i, A_i = 0] + \tau(X_i)A_i\end{aligned}$$

Heterogeneous effects (cont'd)

- We can plug this into the numerator of τ_R above. After a bit of simplification, we get the following:

$$\tau_R = \frac{E\left[\tau(x)(A_i - E[A_i|X_i])^2\right]}{E[(A_i - E[A_i|X_i])^2]} = \frac{E[\tau(X_i)\sigma_A^2(X_i)]}{E[\sigma_A^2]}$$

Heterogeneous effects (cont'd)

- We can plug this into the numerator of τ_R above. After a bit of simplification, we get the following:

$$\tau_R = \frac{E\left[\tau(x)(A_i - E[A_i|X_i])^2\right]}{E[(A_i - E[A_i|X_i])^2]} = \frac{E[\tau(X_i)\sigma_A^2(X_i)]}{E[\sigma_A^2]}$$

- Here, σ_A^2 is the variance of A_i conditional on X_i.

Heterogeneous effects (cont'd)

- We can plug this into the numerator of τ_R above. After a bit of simplification, we get the following:

$$\tau_R = \frac{E\left[\tau(x)(A_i - E[A_i|X_i])^2\right]}{E[(A_i - E[A_i|X_i])^2]} = \frac{E[\tau(X_i)\sigma_A^2(X_i)]}{E[\sigma_A^2]}$$

- Here, σ_A^2 is the variance of A_i conditional on X_i.
- Remember that

$$\tau = E[\tau_i(X_i)]$$

Heterogeneous effects (cont'd)

- We can plug this into the numerator of τ_R above. After a bit of simplification, we get the following:

$$\tau_R = \frac{E\left[\tau(x)(A_i - E[A_i|X_i])^2\right]}{E[(A_i - E[A_i|X_i])^2]} = \frac{E[\tau(X_i)\sigma_A^2(X_i)]}{E[\sigma_A^2]}$$

- Here, σ_A^2 is the variance of A_i conditional on X_i.
- Remember that

$$\tau = E[\tau_i(X_i)]$$

- For the ATE, we simply take the average of the CATEs over the distribution of X_i.

Heterogeneous effects (cont'd)

- We can plug this into the numerator of τ_R above. After a bit of simplification, we get the following:

$$\tau_R = \frac{E\left[\tau(x)(A_i - E[A_i|X_i])^2\right]}{E[(A_i - E[A_i|X_i])^2]} = \frac{E[\tau(X_i)\sigma_A^2(X_i)]}{E[\sigma_A^2]}$$

- Here, σ_A^2 is the variance of A_i conditional on X_i.
- Remember that

$$\tau = E[\tau_i(X_i)]$$

- For the ATE, we simply take the average of the CATEs over the distribution of X_i.
- For the regression coefficient, we take the average weighted by the conditional variance of treatment in that stratum.

- Why does the OLS estimator weight by the conditional variance of the treatment? OLS is a minimum-variance estimator.

- Why does the OLS estimator weight by the conditional variance of the treatment? OLS is a minimum-variance estimator.
- Gives more weight to strata with lower expected variance in their estimates. That is, it gives higher weight to more precise within-strata estimates. When are these estimates going to be more precise? When the treatment and control group are roughly the same size and so the variance is maximized.

- Why does the OLS estimator weight by the conditional variance of the treatment? OLS is a minimum-variance estimator.
- Gives more weight to strata with lower expected variance in their estimates. That is, it gives higher weight to more precise within-strata estimates. When are these estimates going to be more precise? When the treatment and control group are roughly the same size and so the variance is maximized.
- When does $\tau = \tau_R$? When $\tau(x) = \tau$ is constant across the strata of the covariates.

Nonparametric regression

- What do we do about the fact that the regression coefficient does not estimate the ATE or the ATT under heterogeneous effects? Do we have to abandon regression?

Nonparametric regression

- What do we do about the fact that the regression coefficient does not estimate the ATE or the ATT under heterogeneous effects? Do we have to abandon regression?
- An alternative regression estimator is sometimes called the imputation estimator. and

Nonparametric regression

- What do we do about the fact that the regression coefficient does not estimate the ATE or the ATT under heterogeneous effects? Do we have to abandon regression?
- An alternative regression estimator is sometimes called the imputation estimator. and
- Impute the values of $Y_i(1)$ and $Y_i(0)$ for each unit, using a regression, and then taking the average of the differences between these imputations as the estimator for the ATE.

Nonparametric regression

- Let $\hat{\mu}_a(x)$ be a consistent estimator for $\mu_a(x) = E[Y_i(a)|X = x]$. We could always run a saturated (in X_i) linear regression in the treated and control groups separately as this estimator.

$$\tau_{imp} = \frac{1}{N}\sum_i \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

Nonparametric regression

- Let $\hat{\mu}_a(x)$ be a consistent estimator for $\mu_a(x) = E[Y_i(a)|X = x]$. We could always run a saturated (in X_i) linear regression in the treated and control groups separately as this estimator.

$$\tau_{imp} = \frac{1}{N} \sum_i \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

- Thus, we use the regression(s) to predict values of the potential outcomes, then average across the imputed individuals treatment effects. Because each of the regression estimators are consistent, then the imputations estimator is consistent for the ATE as well.

Nonparametric regression

- Let $\hat{\mu}_a(x)$ be a consistent estimator for $\mu_a(x) = E[Y_i(a)|X = x]$. We could always run a saturated (in X_i) linear regression in the treated and control groups separately as this estimator.

$$\tau_{imp} = \frac{1}{N} \sum_i \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

- Thus, we use the regression(s) to predict values of the potential outcomes, then average across the imputed individuals treatment effects. Because each of the regression estimators are consistent, then the imputations estimator is consistent for the ATE as well.
- Can go even further by weakening parametric assumptions on $\mu_a(x)$.

Nonparametric regression R example

```
## load in lalonde data
data(LL, package = 'cem')

reg.0 <- lm(re78 ~ age + education + black + married,
            data = LL, subset = treated == 0)
reg.1 <- lm(re78 ~ age + education + black + married,
            data = LL, subset = treated == 1)

muhat0 <- predict(reg.0, newdata = LL)
muhat1 <- predict(reg.1, newdata = LL)

mean(muhat1 - muhat0)
```

```
## [1] 806.9
```

Subclassification/stratification

- Can we avoid the mismatch between the ATE and the regression coefficient in other ways?

Subclassification/stratification

- Can we avoid the mismatch between the ATE and the regression coefficient in other ways?
- Stratify the data based on X and calculate the condtional average treatment effect

$$\tau(x) = E[Y_i(1) - Y_i(0) | X_i = x].$$

Subclassification/stratification

- Can we avoid the mismatch between the ATE and the regression coefficient in other ways?
- Stratify the data based on X and calculate the condtional average treatment effect

$$\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x].$$

- Ignorability ensures that these conditional average treatment effects are identified.

Subclassification/stratification

- Can we avoid the mismatch between the ATE and the regression coefficient in other ways?
- Stratify the data based on X and calculate the condtional average treatment effect

$$\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x].$$

- Ignorability ensures that these conditional average treatment effects are identified.
- If X is discrete with only a few levels, can use the exact values of X.

Subclassification/stratification

- Can we avoid the mismatch between the ATE and the regression coefficient in other ways?
- Stratify the data based on X and calculate the condtional average treatment effect

$$\tau(x) = E[Y_i(1) - Y_i(0) | X_i = x].$$

- Ignorability ensures that these conditional average treatment effects are identified.
- If X is discrete with only a few levels, can use the exact values of X.
- Otherwise, we may have to subclassify/coarsen the data.

Stratification on the propensity score

- What about when X has has many dimensions?

Stratification on the propensity score

- What about when X has has many dimensions?
- Curse of dimensionality: there will be very few, if any, units in a given stratum of X_i.

Stratification on the propensity score

- What about when X has has many dimensions?
- Curse of dimensionality: there will be very few, if any, units in a given stratum of X_i.
- Stratify on a low-dimensional summary, the *propensity score*, which is the unit's individual probability of receiving treatment, condition on the covariates:

$$e_i = \Pr[A_i = 1 | X_i]$$

Stratification on the propensity score

- What about when X has has many dimensions?
- Curse of dimensionality: there will be very few, if any, units in a given stratum of X_i.
- Stratify on a low-dimensional summary, the *propensity score*, which is the unit's individual probability of receiving treatment, condition on the covariates:

$$e_i = \Pr[A_i = 1|X_i]$$

- Rosenbaum and Rubin (1983) showed that if we correctly estimate the e_i, stratifying on e_i is the same as stratifying on the full X_i.

Estimating the propensity score

- Of course, in observational studies, we don't know the propensity score.

Estimating the propensity score

- Of course, in observational studies, we don't know the propensity score.
- We would run a parametric model with parameters γ to estimate the propensity scores:

Estimating the propensity score

- Of course, in observational studies, we don't know the propensity score.
- We would run a parametric model with parameters γ to estimate the propensity scores:

1. Estimate $\hat{\gamma}$

Estimating the propensity score

- Of course, in observational studies, we don't know the propensity score.
- We would run a parametric model with parameters γ to estimate the propensity scores:

1. Estimate $\hat{\gamma}$
2. Create $\hat{e}_i = \Pr[A_i = 1 | X_i; \hat{\gamma}]$

Estimating the propensity score

- Of course, in observational studies, we don't know the propensity score.
- We would run a parametric model with parameters γ to estimate the propensity scores:

1. Estimate $\hat{\gamma}$
2. Create $\hat{e}_i = \Pr[A_i = 1 | X_i; \hat{\gamma}]$

- For instance, in R, we could easily calculate the propensity scores using the `glm` function:

Estimating the propensity score

- Of course, in observational studies, we don't know the propensity score.
- We would run a parametric model with parameters γ to estimate the propensity scores:

1. Estimate $\hat{\gamma}$
2. Create $\hat{e}_i = \Pr[A_i = 1 | X_i; \hat{\gamma}]$

- For instance, in R, we could easily calculate the propensity scores using the `glm` function:
- What variables do we include in the propensity score model? Any set of variables that blocks all the backdoor paths from A_i to Y_i.

Estimating the propensity score

- Of course, in observational studies, we don't know the propensity score.
- We would run a parametric model with parameters γ to estimate the propensity scores:

1. Estimate $\hat{\gamma}$
2. Create $\hat{e}_i = \Pr[A_i = 1 | X_i; \hat{\gamma}]$

- For instance, in R, we could easily calculate the propensity scores using the `glm` function:
- What variables do we include in the propensity score model? Any set of variables that blocks all the backdoor paths from A_i to Y_i.
- Check balance within strata of $\hat{e}_i$ or use automated/nonparametric tools for estimating $\hat{e}_i$.

Standardization/direct adjustment

- Above we calculated the CATE, $\tau(x)$, but what if we want the average treatment effect, τ?

Standardization/direct adjustment

- Above we calculated the CATE, $\tau(x)$, but what if we want the average treatment effect, τ?
- Take the average of the CATEs over the distribution of X:

$$\tau = \sum_x E[Y_i(1) - Y_i(0)|X_i = x] \Pr[X_i = x]$$

Standardization/direct adjustment

- Above we calculated the CATE, $\tau(x)$, but what if we want the average treatment effect, τ?
- Take the average of the CATEs over the distribution of X:

$$\tau = \sum_x E[Y_i(1) - Y_i(0)|X_i = x] \Pr[X_i = x]$$

- When X_i is low dimensional and discrete, we can easily calculate $\Pr[X_i = x]$ with its empirical distribution: $\frac{1}{N} \sum_i^N \mathbb{I}(X_i = x)$.

Standardization/direct adjustment

- Above we calculated the CATE, $\tau(x)$, but what if we want the average treatment effect, τ?
- Take the average of the CATEs over the distribution of X:

$$\tau = \sum_x E[Y_i(1) - Y_i(0)|X_i = x] \Pr[X_i = x]$$

- When X_i is low dimensional and discrete, we can easily calculate $\Pr[X_i = x]$ with its empirical distribution: $\frac{1}{N}\sum_i^N \mathbb{I}(X_i = x)$.
- For subclassification on the propensity score, you simply weight by the size of each stratum.

Matching

- Basic idea: find control units that are very similar to treated units on X_i.

Matching

- Basic idea: find control units that are very similar to treated units on X_i.
- One way to think of this approach is that we are “imputing” the missing values $Y_i(0)$ for the treated units, using control units with very similar values of X_i.

Matching

- Basic idea: find control units that are very similar to treated units on X_i.
- One way to think of this approach is that we are "imputing" the missing values $Y_i(0)$ for the treated units, using control units with very similar values of X_i.
- Remember that matching doesn't justify a causal effect, ignorability does.

Exact matching

- Let's say that for each treated unit we can find an *exact match*: a control unit with the same values of X_i and suppose we drop any control units that are not matched.

Exact matching

- Let's say that for each treated unit we can find an *exact match*: a control unit with the same values of X_i and suppose we drop any control units that are not matched.
- Exact balance: $\Pr(X_i = x | A_i = 1) = \Pr(X_i = x | A_i = 0)$ for all values of x.

Exact matching

- Let's say that for each treated unit we can find an *exact match*: a control unit with the same values of X_i and suppose we drop any control units that are not matched.
- Exact balance: $\Pr(X_i = x | A_i = 1) = \Pr(X_i = x | A_i = 0)$ for all values of x.
- This is because in the matched data, for every treated unit, there is one (and, in this case, only one) control unit with the same exact value of X_i. The two groups must have the same distribution in X_i.

Why exact matching works

$$\tau_{\text{ATT}} = E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1]$$

Why exact matching works

$$
\begin{aligned}
\tau_{\mathrm{ATT}} =& E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1] \\
=& E[Y_i|A_i = 1] - \sum_{x\in\mathcal{X}} E[Y_i(0)|X_i = x, A_i = 1] \Pr(X_i|A_i = 1)
\end{aligned}
$$

(Consistency & Interated Expectations)

Why exact matching works

$$\begin{aligned}\tau_{\text{ATT}} =& E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1]\\ =& E[Y_i|A_i = 1] - \sum_{x \in \mathcal{X}} E[Y_i(0)|X_i = x, A_i = 1]\,\Pr(X_i|A_i = 1)\end{aligned}$$

(Consistency & Interated Expectations)

$$= E[Y_i|A_i = 1] - \sum_{x \in \mathcal{X}} E[Y_i(0)|X_i = x, A_i = 0]\,\Pr(X_i|A_i = 1)$$

(Ignorability)

Why exact matching works

$$
\begin{aligned}
\tau_{\mathrm{ATT}} =& E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1] \\
=& E[Y_i|A_i = 1] - \sum_{x \in \mathcal{X}} E[Y_i(0)|X_i = x, A_i = 1] \Pr(X_i|A_i = 1) \\
& \quad (\text{Consistency \& Interated Expectations}) \\
=& E[Y_i|A_i = 1] - \sum_{x \in \mathcal{X}} E[Y_i(0)|X_i = x, A_i = 0] \Pr(X_i|A_i = 1) \\
& \quad (\text{Ignorability}) \\
=& E[Y_i|A_i = 1] - \sum_{x \in \mathcal{X}} E[Y_i|X_i = x, A_i = 0] \Pr(X_i|A_i = 1) \\
& \quad (\text{Consistency})
\end{aligned}
$$

Why exact matching works

$$\begin{aligned}
\tau_{\text{ATT}} =& E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1] \\
=& E[Y_i|A_i = 1] - \sum_{x\in\mathcal{X}} E[Y_i(0)|X_i = x, A_i = 1] \Pr(X_i|A_i = 1) \\
& \text{(Consistency \& Interated Expectations)} \\
=& E[Y_i|A_i = 1] - \sum_{x\in\mathcal{X}} E[Y_i(0)|X_i = x, A_i = 0] \Pr(X_i|A_i = 1) \\
& \text{(Ignorability)} \\
=& E[Y_i|A_i = 1] - \sum_{x\in\mathcal{X}} E[Y_i|X_i = x, A_i = 0] \Pr(X_i|A_i = 1) \\
& \text{(Consistency)} \\
=& E[Y_i|A_i = 1] - \sum_{x\in\mathcal{X}} E[Y_i|X_i = x, A_i = 0] \Pr(X_i|A_i = 0) \\
& \text{(Exactly Matched Data)}
\end{aligned}$$

Why exact matching works

$$
\begin{aligned}
\tau_{\mathrm{ATT}} =& E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1] \\
=& E[Y_i|A_i = 1] - \sum_{x \in \mathcal{X}} E[Y_i(0)|X_i = x, A_i = 1] \Pr(X_i|A_i = 1) \\
& \text{(Consistency \& Interated Expectations)} \\
=& E[Y_i|A_i = 1] - \sum_{x \in \mathcal{X}} E[Y_i(0)|X_i = x, A_i = 0] \Pr(X_i|A_i = 1) \\
& \text{(Ignorability)} \\
=& E[Y_i|A_i = 1] - \sum_{x \in \mathcal{X}} E[Y_i|X_i = x, A_i = 0] \Pr(X_i|A_i = 1) \\
& \text{(Consistency)} \\
=& E[Y_i|A_i = 1] - \sum_{x \in \mathcal{X}} E[Y_i|X_i = x, A_i = 0] \Pr(X_i|A_i = 0) \\
& \text{(Exactly Matched Data)} \\
=& E[Y_i|A_i = 1] - E[Y_i|A_i = 0] \\
& \text{(Iterated Expectations)}
\end{aligned}
$$

The matching procedure

1. Choose a number of matches (1 control:1 treated, 2:1, k:1, etc), whether to match with replacement or not, and a distance metric (propensity scores, Mahalanobis distance).

The matching procedure

1. Choose a number of matches (1 control:1 treated, 2:1, k:1, etc), whether to match with replacement or not, and a distance metric (propensity scores, Mahalanobis distance).
2. Choose a set of pre-treatment covariates that satify ignorability.

The matching procedure

1. Choose a number of matches (1 control:1 treated, 2:1, k:1, etc), whether to match with replacement or not, and a distance metric (propensity scores, Mahalanobis distance).
2. Choose a set of pre-treatment covariates that satify ignorability.
3. Find matches (nearest neighbor, GenMatch, optimal matching), dropping control units that are not matched.

The matching procedure

1. Choose a number of matches (1 control:1 treated, 2:1, k:1, etc), whether to match with replacement or not, and a distance metric (propensity scores, Mahalanobis distance).
2. Choose a set of pre-treatment covariates that satify ignorability.
3. Find matches (nearest neighbor, GenMatch, optimal matching), dropping control units that are not matched.
4. Check balance (difference-in-means, medians, eQQ, etc)

The matching procedure

1. Choose a number of matches (1 control:1 treated, 2:1, k:1, etc), whether to match with replacement or not, and a distance metric (propensity scores, Mahalanobis distance).
2. Choose a set of pre-treatment covariates that satify ignorability.
3. Find matches (nearest neighbor, GenMatch, optimal matching), dropping control units that are not matched.
4. Check balance (difference-in-means, medians, eQQ, etc)
5. Repeat (1)-(4) until balance is acceptable, adding variables or functions of variables to improve balance.

The matching procedure

1. Choose a number of matches (1 control:1 treated, 2:1, k:1, etc), whether to match with replacement or not, and a distance metric (propensity scores, Mahalanobis distance).
2. Choose a set of pre-treatment covariates that satify ignorability.
3. Find matches (nearest neighbor, GenMatch, optimal matching), dropping control units that are not matched.
4. Check balance (difference-in-means, medians, eQQ, etc)
5. Repeat (1)-(4) until balance is acceptable, adding variables or functions of variables to improve balance.
6. Calculate the effect of the treatment on the outcome in the matched datasets.

Matching notes

- As long as you only drop control units, matching will estimate the ATT. But if we drop any treatment units, then we are estimating a different quantity of interest depending on the sample that remains. Sometimes we call this the feasible ATE.

Matching notes

- As long as you only drop control units, matching will estimate the ATT. But if we drop any treatment units, then we are estimating a different quantity of interest depending on the sample that remains. Sometimes we call this the feasible ATE.
- There's a bias-variance tradeoff in the number of matches—more matches means the set of matches might be worse, but you have more of them so the estimates are better.

Weighting

- The intuition behind the weighting approach comes from a sampling mindset: each of the treated and control samples are unrepresentative of the overall population, which leads to imbalance in the covariates and the confounding.

Weighting

- The intuition behind the weighting approach comes from a sampling mindset: each of the treated and control samples are unrepresentative of the overall population, which leads to imbalance in the covariates and the confounding.
- What do we usually do with unrepresentative samples? Reweight them to be more representative.

Weighting

- The intuition behind the weighting approach comes from a sampling mindset: each of the treated and control samples are unrepresentative of the overall population, which leads to imbalance in the covariates and the confounding.
- What do we usually do with unrepresentative samples? Reweight them to be more representative.
- Turns out, we can weight by the inverse of the probability of receiving the treatment the unit actually received:

$$W_{ax} = 1/\Pr[A = a|X = x]$$

Weighting

- The intuition behind the weighting approach comes from a sampling mindset: each of the treated and control samples are unrepresentative of the overall population, which leads to imbalance in the covariates and the confounding.
- What do we usually do with unrepresentative samples? Reweight them to be more representative.
- Turns out, we can weight by the inverse of the probability of receiving the treatment the unit actually received:

$$W_{ax} = 1 / \Pr[A = a | X = x]$$

- Taking a weighted difference in means or using a WLS with these as weights can estimate the ATE.

Weighting

- The intuition behind the weighting approach comes from a sampling mindset: each of the treated and control samples are unrepresentative of the overall population, which leads to imbalance in the covariates and the confounding.
- What do we usually do with unrepresentative samples? Reweight them to be more representative.
- Turns out, we can weight by the inverse of the probability of receiving the treatment the unit actually received:

$$W_{ax} = 1/\Pr[A = a|X = x]$$

- Taking a weighted difference in means or using a WLS with these as weights can estimate the ATE.
- Nice because it avoids having to model the relationship between X and Y, but you do have to model the propensity score.

Wrap-up

- Randomization breaks selection bias.

Wrap-up

- Randomization breaks selection bias.
- Without randomization we have to rely on assumptions about conditional ingorability.

Wrap-up

- Randomization breaks selection bias.
- Without randomization we have to rely on assumptions about conditional ingorability.
- With this selection on observables assumption, we can use a couple of different techniques for estimating causal effects.

Next week

- What if selection on observables doesn't hold? Are we completely out of luck?

Next week

- What if selection on observables doesn't hold? Are we completely out of luck?
- Not necessarily. If we have access to "natural experiments," we can sometimes make more limited inferences.

Next week

- What if selection on observables doesn't hold? Are we completely out of luck?
- Not necessarily. If we have access to "natural experiments," we can sometimes make more limited inferences.
- We'll start down that path next week with instrumental variables.

SATE vs PATE (more info)

- Once we assign some groups to treatment and some to control we do not actually observe $Y_i(1)$ and $Y_i(0)$ and so we cannot actually observe SATE. We can, however, estimate it:

$$\hat{\tau}_S = \frac{1}{n_t} \sum_{i:A_i=1} Y_i - \frac{1}{n_c} \sum_{i:A_i=0} Y_i$$

- Note that, conditional on the sample, the only variation in $\hat{\tau}_S$ is from the treatment assignment. Unconditionally, there are two sources of variation: the treatment assignment and the sampling procedure.

- We can show that, with a completely randomized experiment assignment, $\hat{\tau}_S$ is unbiased for τ_S and, in fact, τ:

$$\begin{aligned}
E[\hat{\tau}_S|S] &= \frac{1}{n_t}\sum_{i:A_i=1} E[Y_i|A_i=1,S] - \frac{1}{n_c}\sum_{i:A_i=0} E[Y_i|A_i=0,S] \\
&= \frac{1}{n_t}\sum_{i:A_i=1} E[Y_i(1)|S] - \frac{1}{n_c}\sum_{i:A_i=0} E[Y_i(0)|S] \\
&= \frac{1}{n_t} n_t E[Y_i(1)|S] - \frac{1}{n_c} n_c E[Y_i(0)|S] \\
&= E[Y_i(1) - Y_i(0)|S] = \frac{1}{n}\sum_{i\in S} Y_i(1) - Y_i(0) = \tau_S
\end{aligned}$$

- We can show that, with a completely randomized experiment assignment, $\hat{\tau}_S$ is unbiased for τ_S and, in fact, τ:

$$\begin{aligned}
E[\hat{\tau}_S|S] &= \frac{1}{n_t}\sum_{i:A_i=1} E[Y_i|A_i=1,S] - \frac{1}{n_c}\sum_{i:A_i=0} E[Y_i|A_i=0,S] \\
&= \frac{1}{n_t}\sum_{i:A_i=1} E[Y_i(1)|S] - \frac{1}{n_c}\sum_{i:A_i=0} E[Y_i(0)|S] \\
&= \frac{1}{n_t} n_t E[Y_i(1)|S] - \frac{1}{n_c} n_c E[Y_i(0)|S] \\
&= E[Y_i(1) - Y_i(0)|S] = \frac{1}{n}\sum_{i\in S} Y_i(1) - Y_i(0) = \tau_S
\end{aligned}$$

- By the law of iterated expectations, we also know that $E[E[\hat{\tau}_S|S]] = E[\tau_S] = \tau$. Thus, the difference in means is also unbiased for the PATE.

- It turns out that the sampling variance of the difference in means estimator is:

$$V(\hat{\tau}_S|S) = \frac{S_c^2}{n_c} + \frac{S_t^2}{n_t} - \frac{S_{\tau_i}^2}{n},$$

- It turns out that the sampling variance of the difference in means estimator is:

$$V(\hat{\tau}_S|S) = \frac{S_c^2}{n_c} + \frac{S_t^2}{n_t} - \frac{S_{\tau_i}^2}{n},$$

- Here S_c^2 and S_t^2 are the in-sample variances of $Y_i(0)$ and $Y_i(1)$, respectively. We can use sample variances within levels of A_i to estimat these.

- It turns out that the sampling variance of the difference in means estimator is:

$$V(\hat{\tau}_S|S) = \frac{S_c^2}{n_c} + \frac{S_t^2}{n_t} - \frac{S_{\tau_i}^2}{n},$$

- Here S_c^2 and S_t^2 are the in-sample variances of $Y_i(0)$ and $Y_i(1)$, respectively. We can use sample variances within levels of A_i to estimat these.
- The last term, $S_{\tau_i}^2$ is the in-sample variance of the individual treatmente effects.

- It turns out that the sampling variance of the difference in means estimator is:

$$V(\hat{\tau}_S|S) = \frac{S_c^2}{n_c} + \frac{S_t^2}{n_t} - \frac{S_{\tau_i}^2}{n},$$

- Here S_c^2 and S_t^2 are the in-sample variances of $Y_i(0)$ and $Y_i(1)$, respectively. We can use sample variances within levels of A_i to estimat these.
- The last term, $S_{\tau_i}^2$ is the in-sample variance of the individual treatmente effects.
- Obviously, we don't observe any individual treatment effects, so we can't estimate a sample variance of this quantity. If the treatment effect is constant, then this term equals zero.

- It turns out that the overall variance of the estimator is simply:

$$V(\hat{\tau}_S) = \frac{\sigma_c^2}{n_c} + \frac{\sigma_t^2}{n_t},$$

which can be estimated with this simple variance estimator:

$$\hat{V} = \frac{\hat{s}_c^2}{n_c} + \frac{\hat{s}_t^2}{n_t}$$

Pearl, J. {2000}. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

- It turns out that the overall variance of the estimator is simply:

$$V(\hat{\tau}_S) = \frac{\sigma_c^2}{n_c} + \frac{\sigma_t^2}{n_t},$$

which can be estimated with this simple variance estimator:

$$\hat{V} = \frac{\hat{s}_c^2}{n_c} + \frac{\hat{s}_t^2}{n_t}$$

- This estimator is unbiased for the variance of the difference in means in the population OR a conservative estimate of the variance of the difference in means in the sample.

Pearl, J. {2000}. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.