Causal Inference
CS 477-677

Introduction To Counterfactuals, Randomization Based Inference, And Missing Data Problems

Ilya Shpitser

Outline

Getting Causality From A Statistical Model

- Can learn model parameters from data.
- Would be great if we could interpret them causally.
- Example: large coefficient in linear regression – large causal effect (guns cause murders, alcohol causes accidents, etc.)
- But everyone knows: association does not imply causation:

Getting Causality From A Statistical Model

- Can learn model parameters from data.
- Would be great if we could interpret them causally.
- Example: large coefficient in linear regression – large causal effect (guns cause murders, alcohol causes accidents, etc.)
- But everyone knows: association does not imply causation:
    - People in hospitals tend to be sick.

Getting Causality From A Statistical Model

- Can learn model parameters from data.
- Would be great if we could interpret them causally.
- Example: large coefficient in linear regression – large causal effect (guns cause murders, alcohol causes accidents, etc.)
- But everyone knows: association does not imply causation:
    - People in hospitals tend to be sick. Therefore I should not go to the hospital – it will make me sick.

Getting Causality From A Statistical Model

- Can learn model parameters from data.
- Would be great if we could interpret them causally.
- Example: large coefficient in linear regression – large causal effect (guns cause murders, alcohol causes accidents, etc.)
- But everyone knows: association does not imply causation:
  - People in hospitals tend to be sick. Therefore I should not go to the hospital – it will make me sick.
  - People who own olympic gold medals in running tend to be fast runners!

Getting Causality From A Statistical Model

- Can learn model parameters from data.
- Would be great if we could interpret them causally.
- Example: large coefficient in linear regression – large causal effect (guns cause murders, alcohol causes accidents, etc.)
- But everyone knows: association does not imply causation:
  - People in hospitals tend to be sick. Therefore I should not go to the hospital – it will make me sick.
  - People who own olympic gold medals in running tend to be fast runners! Therefore to become fast, I should buy a medal!

Getting Causality From A Statistical Model

- Can learn model parameters from data.
- Would be great if we could interpret them causally.
- Example: large coefficient in linear regression – large causal effect (guns cause murders, alcohol causes accidents, etc.)
- But everyone knows: association does not imply causation:
  - People in hospitals tend to be sick. Therefore I should not go to the hospital – it will make me sick.
  - People who own olympic gold medals in running tend to be fast runners! Therefore to become fast, I should buy a medal!
  - “Cargo cult” behavior.

Getting Causality From A Statistical Model

- Can learn model parameters from data.
- Would be great if we could interpret them causally.
- Example: large coefficient in linear regression – large causal effect (guns cause murders, alcohol causes accidents, etc.)
- But everyone knows: association does not imply causation:
  - People in hospitals tend to be sick. Therefore I should not go to the hospital – it will make me sick.
  - People who own olympic gold medals in running tend to be fast runners! Therefore to become fast, I should buy a medal!
  - "Cargo cult" behavior.
- When does association imply causation? Will talk about this today.

Hume's Definition

- Recall ("first" = cause, "second" = effect):

  all the objects, similar to the first, are followed by objects similar to the second, . . . where, if the first object had not been the second never had existed.

- This is a counterfactual definition.
- Let's try to think about this formally.

Counterfactuals

- Will need outcome Y (like in regression) and treatment or exposure A.
- Will define a potential outcome:

$$Y(a) \equiv \text{“}Y \text{ if } A\text{, possibly contrary to fact, had value } a\text{”}.$$

- What this is not (in general): Y conditional on $A = a$ $(Y \mid a)$.
- What this is:

Counterfactuals

- Will need outcome Y (like in regression) and treatment or exposure A.
- Will define a potential outcome:

$$Y(a) \equiv \text{“}Y \text{ if } A\text{, possibly contrary to fact, had value } a\text{”}.$$

- What this is not (in general): Y conditional on $A = a$ ($Y \mid a$).
- What this is:
  - A is input to a procedure in Java, Y is the output. $Y(a)$: output of procedure if we stopped execution in debugger, and set input to a.

Counterfactuals

- Will need outcome Y (like in regression) and treatment or exposure A.
- Will define a potential outcome:

$$Y(a) \equiv \text{“}Y \text{ if } A\text{, possibly contrary to fact, had value } a\text{”}.$$

- What this is not (in general): Y conditional on $A = a$ $(Y \mid a)$.
- What this is:
  - A is input to a procedure in Java, Y is the output. $Y(a)$: output of procedure if we stopped execution in debugger, and set input to a.
  - $p(Y \mid a)$: probability of rain if my lawn is wet.

Counterfactuals

- Will need outcome Y (like in regression) and treatment or exposure A.
- Will define a potential outcome:

$$Y(a) \equiv \text{“}Y \text{ if } A\text{, possibly contrary to fact, had value } a\text{”}.$$

- What this is not (in general): Y conditional on $A = a$ $(Y \mid a)$.
- What this is:
  - A is input to a procedure in Java, Y is the output. $Y(a)$: output of procedure if we stopped execution in debugger, and set input to a.
  - $p(Y \mid a)$: probability of rain if my lawn is wet.
  - $p(Y(a))$: probability of rain if I sprayed my lawn with a hose.

Encoding Hume's Definition

- $A = 1$: fire, $Y = 1$: smoke, $A = 0$: no fire, $Y = 0$: no smoke.
- Smoke follows fire: $Y(A = 1) = 1$.
- If there had been no fire, there would have been no smoke: $Y(A = 0) = 0$.
- Can establish causality by comparing $Y(a)$ for different a:

$$Y(A = 1) - Y(A = 0).$$

- This is called a causal effect or causal contrast.
- Shorthand: $Y(1) \equiv Y(A = 1)$, if A is understood.

Linking The Counterfactual And The Factual

- We are not (just) doing philosophy, we want to do data analysis!
- Data records what actually happened.
- What we want is something that did not happen.
- We need to link counterfactuals and observed data.
- Standard assumption is called consistency: $Y(A) = Y$. Read:

"Observed Y and Y if we were to set A to whatever value it was observed are the same variable."

Linking The Counterfactual And The Factual

- We are not (just) doing philosophy, we want to do data analysis!
- Data records what actually happened.
- What we want is something that did not happen.
- We need to link counterfactuals and observed data.
- Standard assumption is called consistency: $Y(A) = Y$. Read:

"Observed Y and Y if we were to set A to whatever value it was observed are the same variable."

On the board

Linking The Counterfactual And The Factual

- We are not (just) doing philosophy, we want to do data analysis!
- Data records what actually happened.
- What we want is something that did not happen.
- We need to link counterfactuals and observed data.
- Standard assumption is called consistency: $Y(A) = Y$. Read:

  "Observed Y and Y if we were to set A to whatever value it was observed are the same variable."
- If A was observed to be a, we can get $Y(a)$ as Y. But what if A were something else?

Fundamental Problem of Causal Inference

For every row, only see one outcome (Y^{obs})!

	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$	A	Y^{obs}
	1.1	2.3	-1.2	1	1.1
	1.8	0.3	1.5	0	0.3
	2.0	2.1	-0.1	0	2.1
	0.1	1.3	-1.2	1	0.1
mean	1.25	1.5	-0.25		

Fundamental Problem of Causal Inference

For every row, only see one outcome (Y^{obs})!

	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$	A	Y^{obs}
	1.1	2.3	-1.2	1	1.1
	1.8	0.3	1.5	0	0.3
	2.0	2.1	-0.1	0	2.1
	0.1	1.3	-1.2	1	0.1
mean	1.25	1.5	-0.25		

Need assumptions to link table on A, Y^{obs} and table on $Y(1), Y(0)$.

Ignorability

- A determines what treatment people get.
- Intuition: want A not to depend on potential outcome.
- Example: flip a coin, if heads $A = 1$, if tails, $A = 0$.
- Results in "fair" assignment, any difference in $Y(A)$ has to do with the person, not the assignment mechanism.
- False if e.g. sick people get $A = 1$, healthy people get $A = 0$.
- Compare to Lind's diary.

Ignorability

- A determines what treatment people get.
- Intuition: want A not to depend on potential outcome.
- Example: flip a coin, if heads $A = 1$, if tails, $A = 0$.
- Results in "fair" assignment, any difference in $Y(A)$ has to do with the person, not the assignment mechanism.
- False if e.g. sick people get $A = 1$, healthy people get $A = 0$.
- Compare to Lind's diary.
- Formally: $\{Y(1), Y(0)\} \perp\!\!\!\perp A$.
- Known as ignorability.

Consequences Of Ignorability

- Remember, want to compare, $Y(1)$ and $Y(0)$. Have data on $Y^{obs} = Y$ and A.
- Assume we had *infinite* amount of data, in fact we knew the underlying distribution $p(Y, A)$.
- Assume $\{Y(1), Y(0)\} \perp\!\!\!\perp A$, and consistency. Then

$$p(Y(1)) = p(Y(1)|A = 1) = p(Y|A = 1)$$
$$p(Y(0)) = p(Y(0)|A = 0) = p(Y|A = 0)$$

Consequences Of Ignorability

- Remember, want to compare, $Y(1)$ and $Y(0)$. Have data on $Y^{obs} = Y$ and A.
- Assume we had *infinite* amount of data, in fact we knew the underlying distribution $p(Y, A)$.
- Assume $\{Y(1), Y(0)\} \perp\!\!\!\perp A$, and consistency. Then

$$p(Y(1)) = p(Y(1)|A = 1) = p(Y|A = 1)$$
$$p(Y(0)) = p(Y(0)|A = 0) = p(Y|A = 0)$$

- Ignorability (random treatment assignment) means association is causation(!)
- Basis of the causal validity of randomized controlled trials.

Example

- Assume ignorablity, and our table:

	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$	A	Y^{obs}
	1.1	2.3	-1.2	1	1.1
	1.8	0.3	1.5	0	0.3
	2.0	2.1	-0.1	0	2.1
	0.1	1.3	-1.2	1	0.1
mean	1.25	1.5	-0.25		

- Then $E[Y(1)] = E[Y \mid A = 1]$, $E[Y(0)] = E[Y \mid A = 0]$.

Example

- Assume ignorablity, and our table:

	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$	A	Y^{obs}
	1.1	2.3	-1.2	1	1.1
	1.8	0.3	1.5	0	0.3
	2.0	2.1	-0.1	0	2.1
	0.1	1.3	-1.2	1	0.1
mean	1.25	1.5	-0.25		

- Then $E[Y(1)] = E[Y \mid A = 1]$, $E[Y(0)] = E[Y \mid A = 0]$.
- MLE: $E[Y \mid A = 1] \approx 0.6$, $E[Y \mid A = 0] \approx 1.2$, so

$$E[Y(1)] - E[Y(0)] \approx 0.6 - 1.2 = -0.6.$$

- Difference is called the average causal effect (ACE).
- May also want individual causal effect, e.g. $1.1 - 2.3 = -1.2$ for unit 1.

Neyman's Causal Effect

- For every unit i, $Y_i(1)$ and $Y_i(0)$ are (potentially unknown) fixed quantities.
- Randomness comes only from A (treatment assignment).
- Assume n total units, k assigned to $A = 1$.
- ACE estimate:

$$\widehat{\text{ACE}} = \left(\frac{1}{k}\sum_{i=1}^{n} Y_i \cdot A_i\right) - \left(\frac{1}{n-k}\sum_{i=1}^{n} Y_i \cdot (1 - A_i)\right)$$

- Can estimate variance of ACE also (for confidence intervals):

$$\widehat{\text{Var}}(\widehat{\text{ACE}}) = \frac{\text{Var}(Y_i \mid A = 1)}{k} + \frac{\text{Var}(Y_i \mid A = 0)}{n-k}.$$

- Can estimate reliability of estimator using this.

Confidence Intervals and Central Limit Theorem

- We measure estimator reliability using confidence intervals.
- Generate many estimates $\hat{\theta}(\vec{X}_i)$ where $\vec{X}_i$ is a bootstrap sample.
- Look at quantiles.
- What is the distribution of $\hat{\theta}(\vec{X}_i)$ $(i = 1, \ldots, k)$?

Confidence Intervals and Central Limit Theorem

- We measure estimator reliability using confidence intervals.
- Generate many estimates $\hat{\theta}(\vec{X}_i)$ where $\vec{X}_i$ is a bootstrap sample.
- Look at quantiles.
- What is the distribution of $\hat{\theta}(\vec{X}_i)$ $(i = 1, \ldots, k)$?
- For some $\hat{\theta}(\vec{X}_i)$ looks more and more like a Gaussian as $k \to \infty$.
- Due to Central Limit Theorem: mean of independent random variables is approximately Gaussian.

Confidence Intervals and Central Limit Theorem

- We measure estimator reliability using confidence intervals.
- Generate many estimates $\hat{\theta}(\vec{X}_i)$ where $\vec{X}_i$ is a bootstrap sample.
- Look at quantiles.
- What is the distribution of $\hat{\theta}(\vec{X}_i)$ $(i = 1, \ldots, k)$?
- For some $\hat{\theta}(\vec{X}_i)$ looks more and more like a Gaussian as $k \to \infty$.
- Due to Central Limit Theorem: mean of independent random variables is approximately Gaussian.
- 2.5% and 97.5% quantiles for $\mathcal{N}(\mu, \sigma^2)$ is at

$$\mu - 1.96 \cdot \sigma, \mu + 1.96 \cdot \sigma.$$

Confidence Intervals For Neyman's ACE

- Recall:

$$\widehat{\text{ACE}} = \left(\frac{1}{k}\sum_{i=1}^{n} Y_i \cdot A_i\right) - \left(\frac{1}{n-k}\sum_{i=1}^{n} Y_i \cdot (1 - A_i)\right)$$

$$\widehat{\text{Var}}(\widehat{\text{ACE}}) = \frac{\text{Var}(Y_i \mid A = 1)}{k} + \frac{\text{Var}(Y_i \mid A = 0)}{n-k}.$$

- 2.5%, 97.5% CI is:

$$\widehat{\text{ACE}} \pm 1.96 \cdot \sqrt{\frac{\text{Var}(Y_i \mid A = 1)}{k} + \frac{\text{Var}(Y_i \mid A = 0)}{n-k}}$$

Model Based Inference

- What if we have lots of binary treatments $A_1, \dots A_k$?
- All treatments randomized (need lots of units...)
- Can use any regression model (say linear):

$$E[Y \mid a_1, \dots a_k] = w_0 + \sum_{i=1}^{k} w_i \cdot a_k.$$

- Estimate as usual. Then $E[Y \mid a_1, \dots, a_k] = E[Y(a_1, \dots, a_k)]$, and

$$E[Y(A_1 = 1, a_2, \dots a_k)] - E[Y(A_1 = 0, a_2, \dots, a_k)] = w_1.$$

- Regression coefficients directly encode mean contrast for any treatment!
- Only holds if A_i were perfectly randomized: $Y(a_1, \dots, a_k) \perp\!\!\!\perp \{A_1, \dots, A_k\}$.

Checking For No Effect

- Question of substantive interest: is there an effect at all?
- Frequentist approach:
  - Assume no effect.
  - Calculate a statistic, see how surprising it is under assumption.
  - If very surprising, reject assumption.
- This is propositional logic (contrapositive) applied to probability.
- Problems with this.
- Example: false positives for rare events.

Cancer Screening Example

- Presence of rare cancer: $p(C=1)=0.00001$.
- Test false positive ("boy cries wolf"): $p(T=1|C=0)=0.01$.
- Test false negative ("a wolf is ignored"): $p(T=0|C=1)=0.001$.
- Oh no! Test came back positive ($T=1$)! Should we worry?

Cancer Screening Example

- Presence of rare cancer: $p(C = 1) = 0.00001$.
- Test false positive ("boy cries wolf"): $p(T = 1|C = 0) = 0.01$.
- Test false negative ("a wolf is ignored"): $p(T = 0|C = 1) = 0.001$.
- Oh no! Test came back positive ($T = 1$)! Should we worry?
- Bayes theorem gives answer directly:

$$p(C = 1|T = 1) = \frac{p(T = 1|C = 1)p(C = 1)}{p(T = 1)}$$
$$= \frac{0.999 \cdot 0.00001}{0.999 \cdot 0.00001 + 0.99 \cdot 0.99999} \approx 0.00001$$

Cancer Screening Example

- Presence of rare cancer: $p(C = 1) = 0.00001$.
- Test false positive ("boy cries wolf"): $p(T = 1|C = 0) = 0.01$.
- Test false negative ("a wolf is ignored"): $p(T = 0|C = 1) = 0.001$.
- Oh no! Test came back positive ($T = 1$)! Should we worry?
- Bayes theorem gives answer directly:

$$p(C = 1|T = 1) = \frac{p(T = 1|C = 1)p(C = 1)}{p(T = 1)}$$
$$= \frac{0.999 \cdot 0.00001}{0.999 \cdot 0.00001 + 0.99 \cdot 0.99999} \approx 0.00001$$

- Frequentist way is assume $C = 0$, see how surprised we are!
- $p(T = 1 \mid C = 0) = 0.01$ is surprising. So we start therapy...

Cancer Screening Example

- Presence of rare cancer: $p(C=1)=0.00001$.
- Test false positive ("boy cries wolf"): $p(T=1|C=0)=0.01$.
- Test false negative ("a wolf is ignored"): $p(T=0|C=1)=0.001$.
- Oh no! Test came back positive ($T=1$)! Should we worry?
- Bayes theorem gives answer directly:

$$p(C=1|T=1)=\frac{p(T=1|C=1)p(C=1)}{p(T=1)}$$
$$=\frac{0.999\cdot 0.00001}{0.999\cdot 0.00001+0.99\cdot 0.99999}\approx 0.00001$$

- Frequentist way is assume $C=0$, see how surprised we are!
- $p(T=1\mid C=0)=0.01$ is surprising. So we start therapy...
- Lesson: can't use logic if hypothesis probabilities are very uneven.

P-value: Measure Of Surprise

- Pick a statistic $\hat{\theta}(\vec{X})$.
- Derive a distribution $p_{\emptyset}(\hat{\theta}(\vec{X}))$ of $\hat{\theta}(\vec{X})$ (in closed form, by resampling, etc.) assuming no effect.
- p-value : probability $p_{\emptyset}(\hat{\theta}(\vec{X}) \geq \hat{\theta}(\tilde{\vec{X}}))$, where $\tilde{\vec{X}}$ is actual dataset.
- Throwing two dice example.

On the board

Fisher's Test (Background)

- Will use frequentist approach for checking hypothesis "ACE is 0."
- This means: $Y_i(1) = Y_i(0)$ for all i.
- What is the statistic?

Deriving The Statistic

- Simple example: binary treatment A, binary outcome Y:

	$Y=1$	$Y=0$	
$A=1$	$\sum_i A_i Y_i^{obs}$	$\sum_i A_i(1-Y_i^{obs})$	$\sum_i A_i$
$A=0$	$\sum_i (1-A_i) Y_i^{obs}$	$\sum_i (1-A_i)(1-Y_i^{obs})$	$\sum_i (1-A_i)$
	$\sum_i Y_i^{obs}$	$\sum_i (1-Y_i^{obs})$	N

- Fictitious example, based on data in (Vesikari, 1990): randomized N = 200 infants aged 2-5 months to rotavirus vaccine or placebo:

	$Y=1$	$Y=0$	
$A=1$	35	65	100
$A=0$	45	55	100
	80	120	200

Deriving The Statistic

- Total number of subjects: N.
- Number of cases (got the drug): n.
- Number of positive responses: K.
- Number of positive responses among cases: k.
- Remember, assumed no effect, so can redo table as:

	$Y=1$	$Y=0$	
$A=1$	$\sum_i A_i Y_i(0)$	$\sum_i A_i(1-Y_i(0))$	n
$A=0$	$\sum_i (1-A_i) Y_i(0)$	$\sum_i (1-A_i)(1-Y_i(0))$	$N-n$
	$K=\sum_i Y_i(0)$	$\sum_i (1-Y_i(0))$	N

- Note: everything except case/control assignment is fixed.
- Fisher derived the distribution of k successes out of n draws (w/o replacement) out of a finite pop. of size N with K successes.
- Called the hypergeometric distribution.

Fisher's Test

- Hypergeometric distribution:

$$p(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

k	$n-k$	n
$K-k$	$N-n-K+k$	$N-n$
K	$N-K$	N

- To calculate p-value, count valid table arrangements where top left cell counts are $\leq k$.
- Called Fisher's exact test.

Fisher's Test

- Hypergeometric distribution:

$$p(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

k	$n-k$	n
$K-k$	$N-n-K+k$	$N-n$
K	$N-K$	N

- To calculate p-value, count valid table arrangements where top left cell counts are $\leq k$.
- Called Fisher's exact test.
- In our example, $p(\sum_i A_i Y_i(0) \leq 35) = 0.1938$.

General View Of Missing Data

$$p(\vec{X})$$

$$\Big\downarrow \text{sampling process}$$

$$\begin{array}{cccc} x_1^1 & x_2^1 & x_3^1 & x_4^1 \\ x_1^2 & x_2^2 & x_3^2 & x_4^2 \\ x_1^3 & x_2^3 & x_3^3 & x_4^3 \\ x_1^4 & x_2^4 & x_3^4 & x_4^4 \\ \dots & \dots & \dots & \dots \end{array}$$

$$\Big\downarrow \text{missingness process}$$

$$\begin{array}{cccc} ? & ? & x_3^1 & x_4^1 \\ x_1^2 & x_2^2 & x_3^2 & x_4^2 \\ x_1^3 & x_2^3 & x_3^3 & ? \\ x_1^4 & x_2^4 & x_3^4 & x_4^4 \\ \dots & \dots & \dots & \dots \end{array}$$

General View Of Missing Data

$$
\begin{array}{c}
p(\vec{X}) \\
\text{sampling process} \downarrow \\
\begin{array}{cccc}
x_1^1 & x_2^1 & x_3^1 & x_4^1 \\
x_1^2 & x_2^2 & x_3^2 & x_4^2 \\
x_1^3 & x_2^3 & x_3^3 & x_4^3 \\
x_1^4 & x_2^4 & x_3^4 & x_4^4 \\
\dots & \dots & \dots & \dots
\end{array} \\
\text{missingness process} \downarrow \\
\begin{array}{cccc}
? & ? & x_3^1 & x_4^1 \\
x_1^2 & x_2^2 & x_3^2 & x_4^2 \\
x_1^3 & x_2^3 & x_3^3 & ? \\
x_1^4 & x_2^4 & x_3^4 & x_4^4 \\
\dots & \dots & \dots & \dots
\end{array}
\end{array}
$$

General View Of Missing Data

$p(\vec{X})$

sampling process $\downarrow$

$$\begin{matrix} x_1^1 & x_2^1 & x_3^1 & x_4^1 \\ x_1^2 & x_2^2 & x_3^2 & x_4^2 \\ x_1^3 & x_2^3 & x_3^3 & x_4^3 \\ x_1^4 & x_2^4 & x_3^4 & x_4^4 \\ \cdots & \cdots & \cdots & \cdots \end{matrix}$$

missingness process $\downarrow$

$$\begin{matrix} ? & ? & x_3^1 & x_4^1 \\ x_1^2 & x_2^2 & x_3^2 & x_4^2 \\ x_1^3 & x_2^3 & x_3^3 & ? \\ x_1^4 & x_2^4 & x_3^4 & x_4^4 \\ \cdots & \cdots & \cdots & \cdots \end{matrix}$$

model

$p(\vec{X} \mid \text{missing})$ $\rightarrow$

$$\begin{matrix} x_1^1 & x_2^1 & x_3^1 & x_4^1 \\ x_1^2 & x_2^2 & x_3^2 & x_4^2 \\ x_1^3 & x_2^3 & x_3^3 & x_4^3 \\ x_1^4 & x_2^4 & x_3^4 & x_4^4 \\ \cdots & \cdots & \cdots & \cdots \end{matrix}$$

model

$p(\text{missing} \mid \vec{X})$ $\rightarrow$

$$\begin{array}{cccc|c} x_1^2 & x_2^2 & x_3^2 & x_4^2 & w^2 \\ x_1^4 & x_2^4 & x_3^4 & x_4^4 & w^4 \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{array}$$

Missing Data Vs Causal Inference

- In causal inference, inferences about a counterfactual from the observed data law.
- In missing data, inferences about the full data law from the observed data law.
- Lots of similarities.
  - In both cases the link provided by an untestable model.
  - In both cases we are compensating for bias in the data.
- As we will see, similarities are not superficial.

Causal Inference As A Missing Data Problem

- Can think of causal inference as inference about missing variables.
- Full data distribution: $p(Y(1), Y(0), A)$. A is observed, $Y(1), Y(0)$ are missing!
- Observed data distribution: $p(Y, A)$, where Y is an always observed "coarsened version" of $Y(1)$ and $Y(0)$: $Y \equiv Y(1)A + Y(0)(1 - A)$ (by consistency!)
- Modeling assumption: $\{Y(1), Y(0)\} \perp\!\!\!\perp A$.

Missing Data as a Causal (Counterfactual) Problem

- Will have three types of variables: missing (first matrix), proxies (second matrix), and indicators (is variable missing or not?)
- Denote missing variable by $X_i(1)$, reads

  "the variable X_i if we could, hypothetically, see it."

- Every $X_i(1)$ has an indicator R_i and a factual (proxy) variable X_i.
- If $R_i = 1$, $X_i = X_i(1)$.
- If $R_i = 0$, $X_i = ?$ (or undefined).
- Problem is about how R_i and $X_i(1)$ are related.

Rubin's Missingness Hierarchy

- Donald Rubin did much of early work on causal inference and missing data. Also invented the EM algorithm!
- Established modern missingness hierarchy:
    - Missing Completely At Random (MCAR): presence of ? determined by an independent coin flip.
    - Missing At Random (MAR): presence of ? determined by a coin independent of underlying variable given observed data.
    - Missing Not At Random (MNAR): neither of the above.
- MCAR is very easy.
- Most missing data work assumes MAR, but it's unrealistic.
- Will use graphs to represent this hierarchy.

Missing Completely at Random

- MCAR: "events that lead to missingness occur independently of observed and unobserved data."
- Our translation: $X_1(1) \perp\!\!\!\perp R_1$.
- Then: $p(X_1(1)) = p(X_1(1) \mid R_1 = 1) = p(X_1 \mid R_1 = 1)$.
- The assumption and derivation should look very familiar.

Missing Completely at Random

- MCAR: "events that lead to missingness occur independently of observed and unobserved data."
- Our translation: $X_1(1) \perp\!\!\!\perp R_1$.
- Then: $p(X_1(1)) = p(X_1(1) \mid R_1 = 1) = p(X_1 \mid R_1 = 1)$.
- The assumption and derivation should look very familiar.
- MCAR is just the ignorable model, with the same derivation!
- Proxy is "outcome," indicator is "treatment."

Missing Completely at Random

- MCAR: "events that lead to missingness occur independently of observed and unobserved data."
- Our translation: $X_1(1) \perp\!\!\!\perp R_1$.
- Then: $p(X_1(1)) = p(X_1(1) \mid R_1 = 1) = p(X_1 \mid R_1 = 1)$.
- The assumption and derivation should look very familiar.
- MCAR is just the ignorable model, with the same derivation!
- Proxy is "outcome," indicator is "treatment."
- Under MCAR can use observed case analysis.
- In other words, do your analysis on fully observed rows only. Under MCAR this analysis is equivalent to one we would have done on full data.

Summary

- Operationalized Hume's definition using potential outcomes.
- Like Lind, define causal effect as a contrast: $Y(1) - Y(0)$.
- If treatment A is randomly assigned, effectively association is causation.
- Approaches for assessing causation under randomization:

- Introduced missing data problems.
- Causal inference is a type of missing data problem, missing data may be viewed counterfactually.

Summary

- Operationalized Hume's definition using potential outcomes.
- Like Lind, define causal effect as a contrast: $Y(1) - Y(0)$.
- If treatment A is randomly assigned, effectively association is causation.
- Approaches for assessing causation under randomization:
  - Neyman's conditional mean and variance.

- Introduced missing data problems.
- Causal inference is a type of missing data problem, missing data may be viewed counterfactually.

Summary

- Operationalized Hume's definition using potential outcomes.
- Like Lind, define causal effect as a contrast: $Y(1) - Y(0)$.
- If treatment A is randomly assigned, effectively association is causation.
- Approaches for assessing causation under randomization:
    - Neyman's conditional mean and variance.
    - Regression based approaches for multiple randomized treatments (lots more in design of experiments).

- Introduced missing data problems.
- Causal inference is a type of missing data problem, missing data may be viewed counterfactually.

Summary

- Operationalized Hume's definition using potential outcomes.
- Like Lind, define causal effect as a contrast: $Y(1) - Y(0)$.
- If treatment A is randomly assigned, effectively association is causation.
- Approaches for assessing causation under randomization:
  - Neyman's conditional mean and variance.
  - Regression based approaches for multiple randomized treatments (lots more in design of experiments).
  - Fisher's exact (permutation) test for no effect.
- Introduced missing data problems.
- Causal inference is a type of missing data problem, missing data may be viewed counterfactually.

Next time: Dealing With Confounding