Causal Inference
CS 477-677

Dealing With (Simple) Confounding

Ilya Shpitser

Outline

Last Time

- Formalized "causal effect":

$$E[Y(1)] - E[Y(0)].$$

- Linked with observed data using:
  - Consistency: $Y(A) = Y$.
  - Ignorability: $Y(a) \perp\!\!\!\perp A$ for all a.
- Assessing effect size (non-parametric for one treatment, parametric for many).
- Test for no effect (Fisher's test).
- Introduced the "missing completely at random" model.
- Today: what if A is not randomly assigned? Most data is like this!

Introducing: Causal Graphs

- Will think about causal relationships using graphs.
- For now: nodes are variables, $\rightarrow$ means "directly causes."
- Will make more precise as we go.
- Absences of nodes and edges are important.

Introducing: Causal Graphs

- Will think about causal relationships using graphs.
- For now: nodes are variables, $\rightarrow$ means "directly causes."
- Will make more precise as we go.
- Absences of nodes and edges are important.
- Randomization example (one treatment A, one outcome Y).
- Observed situation:

Introducing: Causal Graphs

- Will think about causal relationships using graphs.
- For now: nodes are variables, $\rightarrow$ means "directly causes."
- Will make more precise as we go.
- Absences of nodes and edges are important.
- Randomization example (one treatment A, one outcome Y).
- Observed situation:

$$A \longrightarrow Y$$

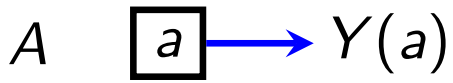
- Hypothetical situation (what if A were a):

$$A \quad \boxed{a} \longrightarrow Y(a)$$

- Think of hypothetical $A = a$ as setting a variable to a value in a debugger. Sometimes called an intervention.
- Differentiate between what variable does normally (A in the graph) and the hypothetical value (a in the graph).

Introducing: Causal Graphs

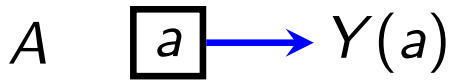
- Note: no path from A to $Y(a)$, which means (will make precise later) $A \perp\!\!\!\perp Y(a)$.



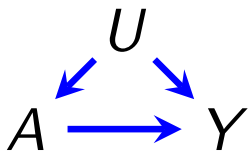
- General method of constructing graphs like this, and reading independences off introduced later.
- How do we represent A not being randomly assigned graphically?

Introducing: Causal Graphs

- Note: no path from A to $Y(a)$, which means (will make precise later) $A \perp\!\!\!\perp Y(a)$.



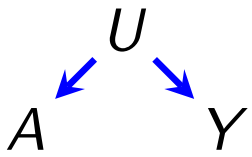
- General method of constructing graphs like this, and reading independences off introduced later.
- How do we represent A not being randomly assigned graphically?



- Intuition: there is a common cause U of both A and Y!
- What happens if U exists and is unobserved.

Unobserved Confounders

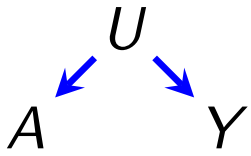
- Simplest example where association is not causation. Why? Imagine A does not cause Y at all!



- Example: "smoking lesion problem" in philosophy (Andy Egan's phrasing):

Unobserved Confounders

- Simplest example where association is not causation. Why? Imagine A does not cause Y at all!



- Example: "smoking lesion problem" in philosophy (Andy Egan's phrasing):

> Susan is debating whether or not to smoke. She knows that smoking is strongly correlated with lung cancer, but only because there is a common cause – a condition that tends to cause both smoking and cancer. Once we fix the presence or absence of this condition, there is no additional correlation between smoking and cancer. Susan prefers smoking without cancer to not smoking without cancer, and prefers smoking with cancer to not smoking with cancer. Should Susan smoke?

Decision Theory For Smoking Lesion

- The point of smoking lesion is that association of A and Y can be very high! That is:

$$p(Y = \text{cancer} \mid A = \text{smoke}) = \text{high}$$
$$p(Y = \text{no cancer} \mid A = \text{do not smoke}) = \text{high}$$

- Most probable states of the world: cancer and smoking, and no cancer and no smoking.
- Second world has more utility for Susan (having cancer is much worse than pleasure from smoking).

Decision Theory For Smoking Lesion

- The point of smoking lesion is that association of A and Y can be very high! That is:

$$p(Y = \text{cancer} \mid A = \text{smoke}) = \text{high}$$
$$p(Y = \text{no cancer} \mid A = \text{do not smoke}) = \text{high}$$

- Most probable states of the world: cancer and smoking, and no cancer and no smoking.
- Second world has more utility for Susan (having cancer is much worse than pleasure from smoking).
- So if Susan uses conditional probabilities to decide, she will not smoke...

Decision Theory For Smoking Lesion

- The point of smoking lesion is that association of A and Y can be very high! That is:

$$p(Y = \text{cancer} \mid A = \text{smoke}) = \text{high}$$
$$p(Y = \text{no cancer} \mid A = \text{do not smoke}) = \text{high}$$

- Most probable states of the world: cancer and smoking, and no cancer and no smoking.
- Second world has more utility for Susan (having cancer is much worse than pleasure from smoking).
- So if Susan uses conditional probabilities to decide, she will not smoke...
- Intuitively this seems wrong. It's all about that U: once it is fixed, whether Susan smokes or not does not influence value of Y ($Y \perp\!\!\!\perp A \mid U$).

Evidential vs Causal Decision Theory

- Maximizing utility based on conditional probabilities is evidential decision theory (EDT).
- Maximizing utility using "causal considerations" is causal decision theory (CDT).
- In the smoking lesion example, CDT does the right thing, and EDT does not.
- There is a long argument about these problems in philosophy (leaving aside for now).
- In simple cases, the lesson is:

Evidential vs Causal Decision Theory

- Maximizing utility based on conditional probabilities is evidential decision theory (EDT).
- Maximizing utility using "causal considerations" is causal decision theory (CDT).
- In the smoking lesion example, CDT does the right thing, and EDT does not.
- There is a long argument about these problems in philosophy (leaving aside for now).
- In simple cases, the lesson is:

Do not act on perceived associations!

Evidential vs Causal Decision Theory

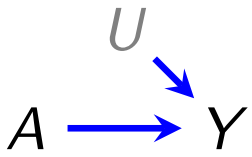
- Maximizing utility based on conditional probabilities is evidential decision theory (EDT).
- Maximizing utility using "causal considerations" is causal decision theory (CDT).
- In the smoking lesion example, CDT does the right thing, and EDT does not.
- There is a long argument about these problems in philosophy (leaving aside for now).
- In simple cases, the lesson is:

Do not act on perceived associations!

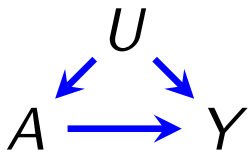
- Unless you can deal with the U somehow.

Dealing with Confounding (a Graphical View)

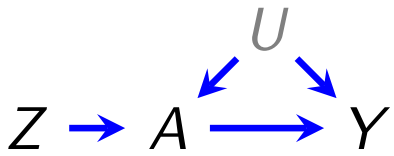
- Randomize (Daniel 1-15, Lind, Pierce, Neyman, Fisher):



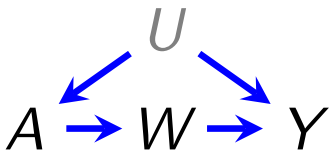
- Observe confounders/stratify, today:



- Instrumental variable + assumptions (P. Wright, 1928), next time:

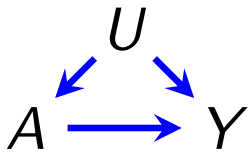


- Find a strong independent mediator (Pearl), later:



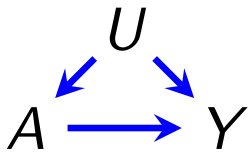
Dealing with Observed Confounders

- Observed situation:

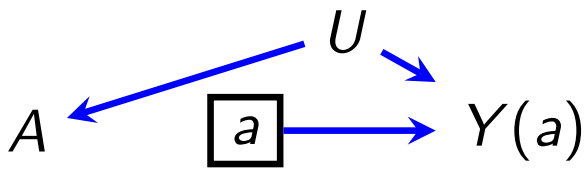


Dealing with Observed Confounders

- Observed situation:

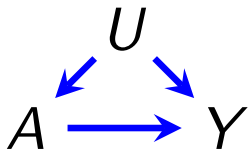


- Representing hypothetical $A = a$ as before:

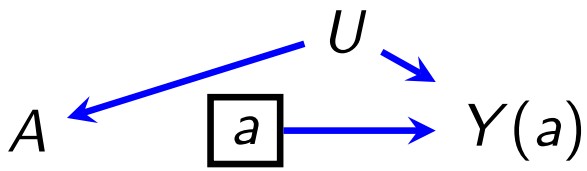


Dealing with Observed Confounders

- Observed situation:



- Representing hypothetical $A = a$ as before:



- Note: A and $Y(a)$ connected via U.
- Implies (will show later) that $A \not\perp\!\!\!\perp Y(a)$.
- Association is not causation: $p(Y \mid A = a) \neq p(Y(a))$!
- Can we get $p(Y(a))$ in some other way?

Ignorability Conditional On A Confounder

- Recall smoking lesion: if Susan had the gene ($U = 1$), $Y(a)$ fully determined by U.
- Same if Susan did not have the gene ($U = 0$).

Ignorability Conditional On A Confounder

- Recall smoking lesion: if Susan had the gene ($U = 1$), $Y(a)$ fully determined by U.
- Same if Susan did not have the gene ($U = 0$).
- In other words: $\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid U$.
- This is called conditional ignorability.

Ignorability Conditional On A Confounder

- Recall smoking lesion: if Susan had the gene ($U = 1$), $Y(a)$ fully determined by U.
- Same if Susan did not have the gene ($U = 0$).
- In other words: $\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid U$.
- This is called conditional ignorability.
- So, conditionally on U, can repeat earlier reasoning:

$$p(Y(a) \mid U = u) = p(Y(a) \mid U = u, A = a) = p(Y \mid U, A = a)$$

Ignorability Conditional On A Confounder

- Recall smoking lesion: if Susan had the gene ($U = 1$), $Y(a)$ fully determined by U.
- Same if Susan did not have the gene ($U = 0$).
- In other words: $\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid U$.
- This is called conditional ignorability.
- So, conditionally on U, can repeat earlier reasoning:

$$p(Y(a) \mid U = u) = p(Y(a) \mid U = u, A = a) = p(Y \mid U, A = a)$$

- But what if we don't know value of U?
- Can average across possible levels of U using prior probablity of observing $p(U = u)$.

Ignorability Conditional On A Confounder

- Recall smoking lesion: if Susan had the gene ($U = 1$), $Y(a)$ fully determined by U.
- Same if Susan did not have the gene ($U = 0$).
- In other words: $\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid U$.
- This is called conditional ignorability.
- So, conditionally on U, can repeat earlier reasoning:

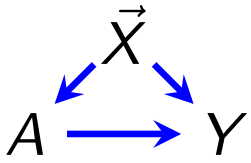
$$p(Y(a) \mid U = u) = p(Y(a) \mid U = u, A = a) = p(Y \mid U, A = a)$$

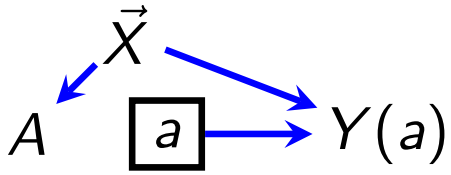
- But what if we don't know value of U?
- Can average across possible levels of U using prior probablity of observing $p(U = u)$.
- In other words, if smoking/cancer gene is rare $p(U = 1) = 0.01$, give most weight to "no gene.":

$$p(Y \mid U = 1, A = a) \cdot 0.01 + p(Y \mid U = 0, A = a) \cdot 0.99.$$

Conditional Ignorability in General

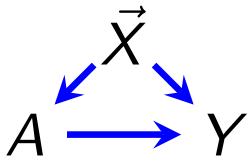
- Treatment A (usually binary, but not necessary).
- Outcome Y (discrete or continuous).
- A vector of baseline factors $\vec{X}$. Picture (observed and counterfactual):

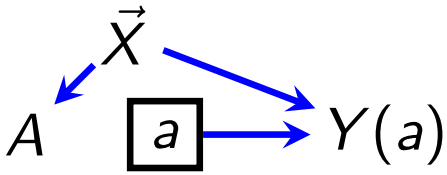




Conditional Ignorability in General

- Treatment A (usually binary, but not necessary).
- Outcome Y (discrete or continuous).
- A vector of baseline factors $\vec{X}$. Picture (observed and counterfactual):



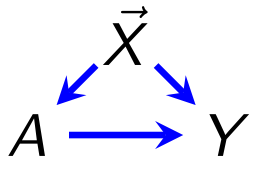


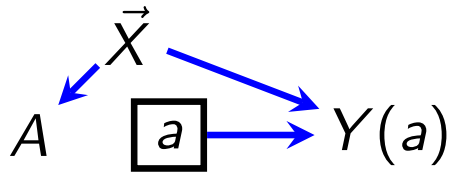
- Predicting what will happen to Y if $A = a$:

$$p(Y(a)) = \sum_{\vec{x}} p(Y \mid A = a, \vec{X} = \vec{x}) p(\vec{X} = \vec{x}).$$

Conditional Ignorability in General

- Treatment A (usually binary, but not necessary).
- Outcome Y (discrete or continuous).
- A vector of baseline factors $\vec{X}$. Picture (observed and counterfactual):





- Predicting what will happen to Y if $A = a$:

$$p(Y(a)) = \sum_{\vec{x}} p(Y \mid A = a, \vec{X} = \vec{x}) p(\vec{X} = \vec{x}).$$

- Average causal effect (ACE):

$$E[Y(1)] - E[Y(0)] = \sum_{\vec{x}} \left\{ E[Y \mid A = 1, \vec{X} = \vec{x}] - E[Y \mid A = 0, \vec{X} = \vec{x}] \right\} p(\vec{x}).$$

- This is called stratification or adjustment formula.
- Old idea: 'adjusting for," "controlling for" refers to this.

Formal Derivation Of The Adjustment Formula

$$
\begin{aligned}
p(Y(a)) &= \sum_{\vec{x}} p(Y(a) \mid \vec{x}) p(\vec{x}) \text{ by chain rule of probability} \\
&= \sum_{\vec{x}} p(Y(a) \mid A = a, \vec{x}) p(\vec{x}) \text{ since } A \perp\!\!\!\perp Y(a) \mid \vec{X} \\
&= \sum_{\vec{x}} p(Y \mid A = a, \vec{x}) p(\vec{x}) \text{ by consistency, } Y(A) = Y.
\end{aligned}
$$

Likelihood Modeling For Adjustment Formula

- Conditional ignorability gives us a target to estimate from observed data:

$$\sum_x \left\{ E[Y \mid A = 1, \vec{X} = \vec{x}] - E[Y \mid A = 0, \vec{X} = \vec{x}] \right\} p(\vec{x}).$$

- Remember the approach with a vector of treatments $\vec{A}$: model $E[Y \mid \vec{A}]$.

Likelihood Modeling For Adjustment Formula

- Conditional ignorability gives us a target to estimate from observed data:

$$\sum_x \left\{ E[Y \mid A = 1, \vec{X} = \vec{x}] - E[Y \mid A = 0, \vec{X} = \vec{x}] \right\} p(\vec{x}).$$

- Remember the approach with a vector of treatments $\vec{A}$: model $E[Y \mid \vec{A}]$.
- Can also do that here, put a model on $E[Y \mid A, \vec{X}]$ (our favorite regression model).
- Do not need to model $p(\vec{x})$, can use the empirical distribution ($1/n$ on every observed point).
- This is called the parametric g-formula.

Parametric g-formula (Step By Step Guide)

Given n data points on $A, Y, \vec{X}$, and assuming conditional ignorability and consistency:

1. Posit statistical model for $E[Y \mid A, \vec{X}; \alpha]$.
2. Fit model by MLE, yielding $\hat{\alpha}$.
3. Estimate ACE by:

$$\frac{1}{n}\left(\sum_i E[Y \mid A = 1, \vec{x}_i; \hat{\alpha}] - E[Y \mid A = 0, \vec{x}_i; \hat{\alpha}]\right)$$

4. Report confidence intervals using bootstrap.

Parametric g-formula (Step By Step Guide)

Given n data points on $A, Y, \vec{X}$, and assuming conditional ignorability and consistency:

1. Posit statistical model for $E[Y \mid A, \vec{X}; \alpha]$.
2. Fit model by MLE, yielding $\hat{\alpha}$.
3. Estimate ACE by:

$$\frac{1}{n}\left(\sum_i E[Y \mid A = 1, \vec{x}_i; \hat{\alpha}] - E[Y \mid A = 0, \vec{x}_i; \hat{\alpha}]\right)$$

4. Report confidence intervals using bootstrap.

Important: what happens if we regularize $E[Y \mid A, \vec{X}; \alpha]$?

Regularization In Causal Inference

- What happens if we regularize $E[Y \mid A, \vec{X}; \alpha]$ in estimator

$$\frac{1}{n}\left(\sum_i E[Y \mid A = 1, \vec{x}_i; \hat{\alpha}] - E[Y \mid A = 0, \vec{x}_i; \hat{\alpha}]\right)$$

- Imagine effect of A on Y is weak (usual case).

Regularization In Causal Inference

- What happens if we regularize $E[Y \mid A, \vec{X}; \alpha]$ in estimator

$$\frac{1}{n}\left(\sum_i E[Y \mid A = 1, \vec{x_i}; \hat{\alpha}] - E[Y \mid A = 0, \vec{x_i}; \hat{\alpha}]\right)$$

- Imagine effect of A on Y is weak (usual case).
- Regularizer will remove coefficient for A!
- We regularized our effect away to 0.

Regularization In Causal Inference

- What happens if we regularize $E[Y \mid A, \vec{X}; \alpha]$ in estimator

$$\frac{1}{n}\left(\sum_i E[Y \mid A = 1, \vec{x}_i; \hat{\alpha}] - E[Y \mid A = 0, \vec{x}_i; \hat{\alpha}]\right)$$

- Imagine effect of A on Y is weak (usual case).
- Regularizer will remove coefficient for A!
- We regularized our effect away to 0.
- Important lesson: regularization has to respect target of inference.
- In classical ML problems, $E[Y \mid \vec{X}]$ or $p(Y, \vec{X})$ is directly relevant.
- In causal inference $E[Y \mid \vec{X}]$ is a *nuisance model* (don't care about it directly, just need it to compute ACE).

Validating Answers

- In machine learning, can learn a very good predictor for Y in terms of $\vec{X}$.
- Can check quality directy using cross-validation.

Validating Answers

- In machine learning, can learn a very good predictor for Y in terms of $\vec{X}$.
- Can check quality directy using cross-validation.
- Cannot do that with ACE as it is a counterfactual quantity (unless A is randomized!)
- We are using observed data to predict result of a hypothetical experiment.

Validating Answers

- In machine learning, can learn a very good predictor for Y in terms of $\vec{X}$.
- Can check quality directy using cross-validation.
- Cannot do that with ACE as it is a counterfactual quantity (unless A is randomized!)
- We are using observed data to predict result of a hypothetical experiment.
- Unless we actually do an experiment and get data, cannot tell if we are right!
- This is why science is hard, we can't just use a fancy algorithm and some data to create science out of nothing!

Validating Answers

- In machine learning, can learn a very good predictor for Y in terms of $\vec{X}$.
- Can check quality directy using cross-validation.
- Cannot do that with ACE as it is a counterfactual quantity (unless A is randomized!)
- We are using observed data to predict result of a hypothetical experiment.
- Unless we actually do an experiment and get data, cannot tell if we are right!
- This is why science is hard, we can't just use a fancy algorithm and some data to create science out of nothing!
- Validation is a hard problem. Causal inference from observed data suggests but does not establish findings.

Ways For Causal Inferences To Go Wrong

- Say we estimate ACE $= 2.3 \pm 0.4$ using parametric g-formula, but experiment does not replicate, reports no significant effect (0).
- Why would that happen?

Ways For Causal Inferences To Go Wrong

- Say we estimate ACE $= 2.3 \pm 0.4$ using parametric g-formula, but experiment does not replicate, reports no significant effect (0).
- Why would that happen?
  - Causal model is wrong (we think all U are observed, but some U are not).

Ways For Causal Inferences To Go Wrong

- Say we estimate ACE $= 2.3 \pm 0.4$ using parametric g-formula, but experiment does not replicate, reports no significant effect (0).
- Why would that happen?
  - Causal model is wrong (we think all U are observed, but some U are not).
  - Causal model is right, but statistical model $E[Y \mid A, \vec{X}]$ is wrong.

Ways For Causal Inferences To Go Wrong

- Say we estimate ACE $= 2.3 \pm 0.4$ using parametric g-formula, but experiment does not replicate, reports no significant effect (0).
- Why would that happen?
  - Causal model is wrong (we think all U are observed, but some U are not).
  - Causal model is right, but statistical model $E[Y \mid A, \vec{X}]$ is wrong.
  - Causal model is right, statistical model is right, population in our analysis vs experiment is different (internal vs external validity).

Ways For Causal Inferences To Go Wrong

- Say we estimate ACE $= 2.3 \pm 0.4$ using parametric g-formula, but experiment does not replicate, reports no significant effect (0).
- Why would that happen?
  - Causal model is wrong (we think all U are observed, but some U are not).
  - Causal model is right, but statistical model $E[Y \mid A, \vec{X}]$ is wrong.
  - Causal model is right, statistical model is right, population in our analysis vs experiment is different (internal vs external validity).
- Cannot test causal model without experimentation.
- Later: learning models from data, ACE under different models.

Ways For Causal Inferences To Go Wrong

- Say we estimate ACE $= 2.3 \pm 0.4$ using parametric g-formula, but experiment does not replicate, reports no significant effect (0).
- Why would that happen?
  - Causal model is wrong (we think all U are observed, but some U are not).
  - Causal model is right, but statistical model $E[Y \mid A, \vec{X}]$ is wrong.
  - Causal model is right, statistical model is right, population in our analysis vs experiment is different (internal vs external validity).
- Cannot test causal model without experimentation.
- Later: learning models from data, ACE under different models.
- Can check suitability of statistical model. Can use flexible models. Can use robust models (may discuss this later). Alternatives to modeling $E[Y \mid A, \vec{X}]$ (next).

Ways For Causal Inferences To Go Wrong

- Say we estimate ACE $= 2.3 \pm 0.4$ using parametric g-formula, but experiment does not replicate, reports no significant effect (0).
- Why would that happen?
  - Causal model is wrong (we think all U are observed, but some U are not).
  - Causal model is right, but statistical model $E[Y \mid A, \vec{X}]$ is wrong.
  - Causal model is right, statistical model is right, population in our analysis vs experiment is different (internal vs external validity).
- Cannot test causal model without experimentation.
- Later: learning models from data, ACE under different models.
- Can check suitability of statistical model. Can use flexible models. Can use robust models (may discuss this later). Alternatives to modeling $E[Y \mid A, \vec{X}]$ (next).
- Movement towards larger, more representative studies to address validity.

Alternative View Of Conditional Ignorability

- Before: $Y(a)$ (causation) is $Y \mid a$ (association) if we condition on $\vec{X}$.
- Alternative: A is not randomized because people with different $\vec{X} = \vec{x}$ are preferentially assigned to 1 vs 0.
- Doctors more likely to give sick people treatment, people with free time are more likely to answer surveys, etc.
- Say we want $E[Y]$ as if $A = 1$.

$$E[Y \mid A = 1] = \sum_i \frac{Y_i \mathbb{I}(A_i = 1)}{\sum_i \mathbb{I}(A_i = 1)}?$$

Alternative View Of Conditional Ignorability

- Before: $Y(a)$ (causation) is $Y \mid a$ (association) if we condition on $\vec{X}$.
- Alternative: A is not randomized because people with different $\vec{X} = \vec{x}$ are preferentially assigned to 1 vs 0.
- Doctors more likely to give sick people treatment, people with free time are more likely to answer surveys, etc.
- Say we want $E[Y]$ as if $A = 1$.

$$E[Y \mid A = 1] = \sum_i \frac{Y_i \mathbb{I}(A_i = 1)}{\sum_i \mathbb{I}(A_i = 1)}?$$

- Wrong for usual reasons.

Alternative View Of Conditional Ignorability

- Before: $Y(a)$ (causation) is $Y \mid a$ (association) if we condition on $\vec{X}$.
- Alternative: A is not randomized because people with different $\vec{X} = \vec{x}$ are preferentially assigned to 1 vs 0.
- Doctors more likely to give sick people treatment, people with free time are more likely to answer surveys, etc.
- Say we want $E[Y]$ as if $A = 1$.

$$E[Y \mid A = 1] = \sum_i \frac{Y_i \mathbb{I}(A_i = 1)}{\sum_i \mathbb{I}(A_i = 1)}?$$

- Wrong for usual reasons.
- Want to compensate for $p(A = 1 \mid \vec{X} = \vec{x})$ being low for some $\vec{x}$.

Alternative View Of Conditional Ignorability

- Before: $Y(a)$ (causation) is $Y \mid a$ (association) if we condition on $\vec{X}$.
- Alternative: A is not randomized because people with different $\vec{X} = \vec{x}$ are preferentially assigned to 1 vs 0.
- Doctors more likely to give sick people treatment, people with free time are more likely to answer surveys, etc.
- Say we want $E[Y]$ as if $A = 1$.

$$E[Y \mid A = 1] = \sum_i \frac{Y_i \mathbb{I}(A_i = 1)}{\sum_i \mathbb{I}(A_i = 1)}?$$

- Wrong for usual reasons.
- Want to compensate for $p(A = 1 \mid \vec{X} = \vec{x})$ being low for some $\vec{x}$.
- So just weigh people more if their $p(A = 1 \mid \vec{X} = \vec{x})$ is low:

$$E[Y(a)] = \frac{1}{N} \sum_i Y_i \frac{\mathbb{I}(A_i = 1)}{p(A_i = 1 \mid \vec{x}_i)}$$

Alternative View Of Conditional Ignorability

- Before: $Y(a)$ (causation) is $Y \mid a$ (association) if we condition on $\vec{X}$.
- Alternative: A is not randomized because people with different $\vec{X} = \vec{x}$ are preferentially assigned to 1 vs 0.
- Doctors more likely to give sick people treatment, people with free time are more likely to answer surveys, etc.
- Say we want $E[Y]$ as if $A = 1$.

$$E[Y \mid A = 1] = \sum_i \frac{Y_i \mathbb{I}(A_i = 1)}{\sum_i \mathbb{I}(A_i = 1)}?$$

- Wrong for usual reasons.
- Want to compensate for $p(A = 1 \mid \vec{X} = \vec{x})$ being low for some $\vec{x}$.
- So just weigh people more if their $p(A = 1 \mid \vec{X} = \vec{x})$ is low:

$$E[Y(a)] = \frac{1}{N} \sum_i Y_i \frac{\mathbb{I}(A_i = 1)}{p(A_i = 1 \mid \vec{x}_i)}$$

- Why would this work?

Inverse Probability Weighting (Importance Sampling)

- Recall definition of $E[Y]$:

$$\sum_{y} y \cdot p(Y = y) = \sum_{y,\vec{x},a} y \cdot p(y \mid a, \vec{x}) p(a \mid \vec{x}) p(\vec{x}).$$

Inverse Probability Weighting (Importance Sampling)

- Recall definition of $E[Y]$:

$$\sum_y y \cdot p(Y = y) = \sum_{y,\vec{x},a} y \cdot p(y \mid a, \vec{x}) p(a \mid \vec{x}) p(\vec{x}).$$

- and definition of $E[Y(a)]$:

$$\sum_y y \cdot p(Y(a) = y) = \sum_{y,\vec{x}} y \cdot p(y \mid a, \vec{x}) p(\vec{x}).$$

Inverse Probability Weighting (Importance Sampling)

- Recall definition of $E[Y]$:

$$\sum_y y \cdot p(Y = y) = \sum_{y,\vec{x},a} y \cdot p(y \mid a, \vec{x}) p(a \mid \vec{x}) p(\vec{x}).$$

- and definition of $E[Y(a)]$:

$$\sum_y y \cdot p(Y(a) = y) = \sum_{y,\vec{x}} y \cdot p(y \mid a, \vec{x}) p(\vec{x}).$$

- Only difference is $p(a \mid \vec{x})$ term.
- Implies (skipping formal proof) that

$$E[Y(a)] = \frac{1}{N} \sum_i Y_i \frac{\mathbb{I}(A_i = 1)}{p(A_i = 1 \mid \vec{x}_i; \hat{\alpha})}$$

- is consistent if we have correct model for $p(A_i = 1 \mid \vec{x}_i)$.

Inverse Probability Weighting (Importance Sampling)

- Recall definition of $E[Y]$:

$$\sum_y y \cdot p(Y = y) = \sum_{y,\vec{x},a} y \cdot p(y \mid a, \vec{x}) p(a \mid \vec{x}) p(\vec{x}).$$

- and definition of $E[Y(a)]$:

$$\sum_y y \cdot p(Y(a) = y) = \sum_{y,\vec{x}} y \cdot p(y \mid a, \vec{x}) p(\vec{x}).$$

- Only difference is $p(a \mid \vec{x})$ term.
- Implies (skipping formal proof) that

$$E[Y(a)] = \frac{1}{N} \sum_i Y_i \frac{\mathbb{I}(A_i = 1)}{p(A_i = 1 \mid \vec{x}_i; \hat{\alpha})}$$

- is consistent if we have correct model for $p(A_i = 1 \mid \vec{x}_i)$.
- Often know this model (treatment assignment, design, etc).
- $p(A_i = 1 \mid \vec{x}_i)$ is called the propensity score.

Inverse Probability Weighting (Step By Step Guide)

Given n data points on $A, Y, \vec{X}$, and assuming conditional ignorablity and consistency:

1. Posit statistical model for $p[A \mid \vec{X}; \alpha]$.
2. Fit model by MLE, yielding $\hat{\alpha}$.
3. Estimate ACE by:

$$\frac{1}{n}\left(\sum_i Y_i \frac{\mathbb{I}(A_i = 1)}{p(a = 1 \mid \vec{x}_i; \hat{\alpha})}\right) - \frac{1}{n}\left(\sum_i Y_i \frac{\mathbb{I}(A_i = 0)}{p(a = 0 \mid \vec{x}_i; \hat{\alpha})}\right)$$

4. Report confidence intervals using bootstrap.

Augmented IPW

- Previous had an estimator that uses $E[Y \mid A, \vec{X}]$, and another that uses $p(A \mid \vec{X})$.
- In fact, is possible to combine them as follows:

$$0 = \frac{I(A = a)}{p(A \mid \vec{X})}\{Y - E[Y \mid A, \vec{X}]\} + E[Y \mid A = a, \vec{X}] - E[E[Y \mid A = a, \vec{X}]].$$

- Can solve for $E[Y(a)] = E[E[Y \mid A = a, \vec{X}]]$.
- Known as the augmented IPW estimator (adding a term to regular IPW).
- Is RAL for $E[Y(a)]$, obtained from the influence function for $E[E[Y \mid A = a, \vec{X}]]$.

Augmented IPW (Step By Step Guide)

Given n data points on $A, Y, \vec{X}$, and assuming conditional ignorablity and consistency:

1. Posit statistical model for $p[A \mid \vec{X}; \alpha_1], E[Y \mid A, \vec{X}; \alpha_2]$.
2. Fit models by MLE, yielding $\hat{\alpha}_1, \hat{\alpha}_2$.
3. Estimate ACE by:

$$\frac{1}{n}\sum_i \{Y_i - E[Y \mid a_i = 1, \vec{x}_i; \hat{\alpha}_2]\}\frac{\mathbb{I}(a_i = 1)}{p(a = 1 \mid \vec{x}_i; \hat{\alpha}_1)} + E[Y \mid a_i = 1, \vec{x}_i; \hat{\alpha}_2] -$$
$$\frac{1}{n}\sum_i \{Y_i - E[Y \mid a_i = 0, \vec{x}_i; \hat{\alpha}_2]\}\frac{\mathbb{I}(A_i = 0)}{p(a = 0 \mid \vec{x}_i; \hat{\alpha}_1)} + E[Y \mid a_i = 0, \vec{x}_i; \hat{\alpha}_2]$$

4. Report confidence intervals using bootstrap.

Pros and Cons Of IPW

- Pros:
  - Simpler model $p(A \mid \vec{X}; \alpha)$, as A is often binary.
  - Needed model often known by design.
  - Simple to fit and implement.
  - Directly produces a dataset where A is set to a. Can run any analysis on top, not just estimating $E[Y(a)]$.
- Cons:
  - Not a likelihood method (statistically inefficient).
  - Dividing by small numbers can lead to problems.

Pros and Cons Of Parametric g-formula

- Pros:
  - Most efficient method if we know the model $E[Y \mid A, \vec{X}]$ (standard results on MLE).
  - In practice, has been found to perform well with flexible models and enough data.
- Cons:
  - Need a more complex model, easy to misspecify.
  - If we need response surface rather than mean, have to use density estimation.
  - Leads to the null paradox for certain queries (will discuss later).

Pros and Cons Of Augmented IPW

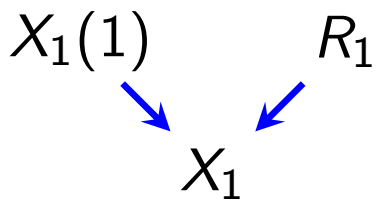
- Pros:
  - Better behaved than IPW.
  - Doubly robust: consistent if either $p(A \mid \vec{X})$ or $E[Y \mid A, \vec{X}]$ is specified correctly.
- Cons:
  - Less efficient than the parametric g-formula if $E[Y \mid A, \vec{X}]$ is correct.
  - May inherit issues of ordinary IPW (division by small numbers).

Introducing: Missing Data Graphs

- Causal models may be represented by two graphs (representing the observed situations, and the counterfactual situation after we "split nodes.").
- Missing data models are represented by a single graph, with $X_i(1)$, X_i and R_i, where $X_i(1)$ and R_i are constrained to cause X_i.

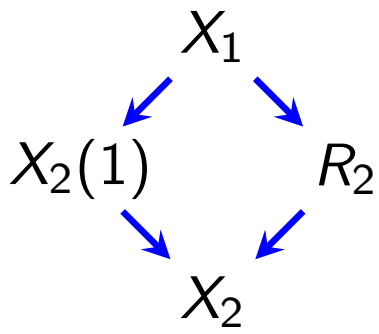
Missing Completely at Random

- MCAR: "events that lead to missingness occur independently of observed and unobserved data."
- Our translation: $X_1(1) \perp\!\!\!\perp R_1$. Graphically:



Missing at Random

- MAR: "events that lead to missingness occur independently of unobserved data given observed data."
- Our translation: $X_2(1) \perp\!\!\!\perp R_2 \mid X_1$ (X_1 fully observed).
- Graphically:

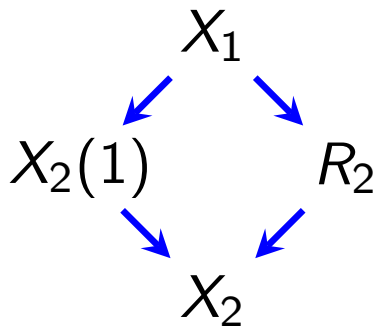


$$p(X_1, X_2(1)) = p(X_2(1) \mid X_1)p(X_1) = p(X_2(1) \mid X_1, R_2 = 1)p(X_1)$$
$$= p(X_2 \mid R_2 = 1, X_1)p(X_1).$$

- MAR is very similar to the conditionally ignorable model.

Estimation Under MAR

- Two estimators for $E[X_2(1)]$ under MAR:



$$E[X_2(1)] = \frac{1}{n}\sum_i E[X_2 \mid R_2 = 1, X_1^i] \text{ (likelihood based inference)}$$

$$E[X_2(1)] = \frac{1}{n}\sum_i \frac{\mathbb{I}(R_2 = 1)}{p(R_2 = 1 \mid X_1^i)} X_2^i \quad \left(\begin{array}{c} \text{propensity based inference:} \\ \text{Horvitz-Thompson} \end{array} \right)$$

- These should look very familiar – just parametric g-formula and IPW.
- What about imputation?

Multiple Imputation

- Parametric g-formula and IPW both deal with observed rows in a particular way.
- In general, have to make tricky modifications to inferences about parameters if missingness is present.
- From the point of view of a data analyst who doesn't know a lot about missing data, it would be nice to "abstract away" tricky issues and just deal with complete data.
- This is where multiple imputation comes in.

Multiple Imputation Overview

- Say we are interested in $\hat{\theta}(\vec{X})$, but there is missingness in $\vec{X}$.
  - Assume MAR, and a model $p(X_2 \mid R_2, X_1; \alpha)$.
  - Fit α (ML or Bayesian methods).
  - Sample all missing values of X_2 using model, and generate m completed datasets $\vec{X}^1, \ldots, \vec{X}^m$.
  - Estimate $\hat{\theta} \equiv \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}(\vec{X}^i)$.
  - Variance of $\hat{\theta}$ has two parts: variance due to $\vec{X}$, and variance due to the imputation (since we are sampling!)

$$Var(\hat{\theta}) = \left(\frac{1}{m} \sum_{i=1}^{m} Var(\hat{\theta}(\vec{X}^i)) \right) + \frac{m+1}{m} \left(\frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}(\vec{X}^i) - \hat{\theta})^2 \right)$$

Summary

- Introduced the conditionally ignorable model, and the missing at random model, where randomization is conditional on an observed set of variables (confounders).
- Can express target of inference ($E[Y(a)]$ or $E[Y(1)]$) as a function of observed data via the adjustment formula) if confounders are all observed.
- Two estimation strategies: parametric g-formula, and IPW.
- Many others exist:
  - Example: match "similar" people based on $p(A \mid \vec{X})$ (propensity score matching).
- We used simple parametric regression models, but can use any probabilistic model.
- General principle in this class: ML and stats teach you fancy models, this class teaches you how to combine them to predict results of hypothetical experiments.

Next time: What If Confounders Are Unobserved?