

Statistical modeling for causal inference

Cornell STSCI / INFO / ILRST 3900

Fall 2024

causal3900.github.io

Oct 20 2026

Learning goals for today

At the end of class, you will be able to

- ▶ Estimate average causal effects via g-formula with a statistical model for the outcome and treatment
- ▶ Reason about the bias variance tradeoff when selecting a statistical model

Standardization

Under consistency and conditional exchangeability and consistency:

$$E(Y \mid A = 1, L = \ell) = E(Y^{a=1} \mid A = 1, L = \ell) = E(Y^{a=1} \mid L = \ell)$$

Standardization

Under consistency and conditional exchangeability and consistency:

$$E(Y \mid A = 1, L = \ell) = E(Y^{a=1} \mid A = 1, L = \ell) = E(Y^{a=1} \mid L = \ell)$$

The ATE can be identified via:

$$\begin{aligned} E(Y^{a=1}) - E(Y^{a=0}) &= \sum_{\ell} \underbrace{\left[E(Y^{a=1} \mid L = \ell) - E(Y^{a=0} \mid L = \ell) \right]}_{\text{Conditional ATE for } \ell} \underbrace{Pr(L = \ell)}_{\text{Prop of group } \ell} \\ &= \underbrace{\sum_{\ell} \left[E(Y^{a=1} \mid L = \ell) Pr(L = \ell) \right]}_{E(Y^{a=1})} - \underbrace{\sum_{\ell} \left[E(Y^{a=0} \mid L = \ell) Pr(L = \ell) \right]}_{E(Y^{a=0})} \end{aligned}$$

Standardization

Under consistency and conditional exchangeability and consistency:

$$E(Y \mid A = 1, L = \ell) = E(Y^{a=1} \mid A = 1, L = \ell) = E(Y^{a=1} \mid L = \ell)$$

The ATE can be identified via:

$$\begin{aligned} E(Y^{a=1}) - E(Y^{a=0}) \\ &= \sum_{\ell} \underbrace{\left[E(Y^{a=1} \mid L = \ell) - E(Y^{a=0} \mid L = \ell) \right]}_{\text{Conditional ATE for } \ell} \underbrace{Pr(L = \ell)}_{\text{Prop of group } \ell} \\ &= \underbrace{\sum_{\ell} \left[E(Y^{a=1} \mid L = \ell) Pr(L = \ell) \right]}_{E(Y^{a=1})} - \underbrace{\sum_{\ell} \left[E(Y^{a=0} \mid L = \ell) Pr(L = \ell) \right]}_{E(Y^{a=0})} \end{aligned}$$

Aggregate the average over sub-groups to get the overall average

$$E(Y^a) = \sum_{\ell} \underbrace{E(Y^a \mid L = \ell)}_{\text{Avg of sub-group}} \times \underbrace{Pr(L = \ell)}_{\text{Prob of sub-group}}$$

Standardization

We want to estimate:

$$E(Y^a) = \sum_{\ell} \underbrace{E(Y^a \mid L = \ell)}_{\text{Avg of sub-group}} \times \underbrace{Pr(L = \ell)}_{\text{Prob of sub-group}}$$

Standardization

We want to estimate:

$$E(Y^a) = \sum_{\ell} \underbrace{E(Y^a \mid L = \ell)}_{\text{Avg of sub-group}} \times \underbrace{Pr(L = \ell)}_{\text{Prob of sub-group}}$$

Given data, we can

- ▶ Estimate $Pr(L = \ell)$ use observed proportion of people with $L_i = \ell$

$$\widehat{Pr}(L = \ell)$$

- ▶ Estimate $E(Y^{a=1} \mid L = \ell)$, use observed average outcome for people with $A_i = 1, L_i = \ell$

$$\hat{E}(Y^{a=1} \mid L = \ell) = \hat{E}(Y \mid A = 1, L = \ell)$$

Standardization

We want to estimate:

$$E(Y^a) = \sum_{\ell} \underbrace{E(Y^a | L = \ell)}_{\text{Avg of sub-group}} \times \underbrace{Pr(L = \ell)}_{\text{Prob of sub-group}}$$

Given data, we can

- ▶ Estimate $Pr(L = \ell)$ use observed proportion of people with $L_i = \ell$

$$\widehat{Pr}(L = \ell)$$

- ▶ Estimate $E(Y^{a=1} | L = \ell)$, use observed average outcome for people with $A_i = 1, L_i = \ell$

$$\hat{E}(Y^{a=1} | L = \ell) = \hat{E}(Y | A = 1, L = \ell)$$

Plug-in estimates for each quantity

$$\hat{E}(Y^a) = \sum_{\ell} \underbrace{\hat{E}(Y^a | L = \ell)}_{\text{Observed avg of sub-group}} \times \underbrace{\widehat{Pr}(L = \ell)}_{\text{Observed prob of sub-group}}$$

Estimating from data

Two small tweaks

- ▶ Estimate $Pr(L = \ell)$ use observed proportion of people with $L_i = \ell$

$$\widehat{Pr}(L = \ell)$$

- ▶ Rewrite as an average across individuals (instead of groups)

Estimating from data

Two small tweaks

- ▶ Estimate $Pr(L = \ell)$ use observed proportion of people with $L_i = \ell$

$$\widehat{Pr}(L = \ell)$$

- ▶ Rewrite as an average across individuals (instead of groups)
- ▶ Estimate $E(Y^{a=1} \mid L = \ell)$, use observed average outcome for people with $A_i = 1, L_i = \ell$

$$\hat{E}(Y^{a=1} \mid L = \ell) = \hat{E}(Y \mid A = 1, L = \ell)$$

- ▶ Fit a model to estimate

$$\hat{E}(Y^{a=1} \mid L = \ell) = \hat{E}(Y \mid A = 1, L = \ell)$$

Standardization

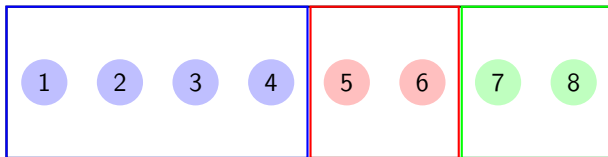
Aggregate the average over sub-groups to get the overall average

$$\hat{E}(Y^a) = \sum_{\ell} \underbrace{\hat{E}(Y^a \mid L = \ell)}_{\text{Avg of sub-group}} \times \underbrace{\hat{Pr}(L = \ell)}_{\text{Prob of sub-group}}$$

Standardization

Aggregate the average over sub-groups to get the overall average

$$\hat{E}(Y^a) = \sum_{\ell} \underbrace{\hat{E}(Y^a | L = \ell)}_{\text{Avg of sub-group}} \times \underbrace{\hat{Pr}(L = \ell)}_{\text{Prob of sub-group}}$$

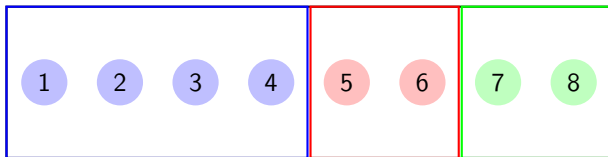


$$\hat{E}(Y^a) = \hat{E}(Y^a | L = \bullet) \times \frac{4}{8} + \hat{E}(Y^a | L = \bullet) \times \frac{2}{8} + \hat{E}(Y^a | L = \bullet) \times \frac{2}{8}$$

Standardization

Aggregate the average over sub-groups to get the overall average

$$\hat{E}(Y^a) = \sum_{\ell} \underbrace{\hat{E}(Y^a | L = \ell)}_{\text{Avg of sub-group}} \times \underbrace{\hat{P}_r(L = \ell)}_{\text{Prob of sub-group}}$$



$$\hat{E}(Y^a) = \hat{E}(Y^a | L = \bullet) \times \frac{4}{8} + \hat{E}(Y^a | L = \bullet) \times \frac{2}{8} + \hat{E}(Y^a | L = \bullet) \times \frac{2}{8}$$

Calculate the same quantity but sum over individuals

$$\begin{aligned} \hat{E}(Y^a) = & \left(\hat{E}(Y^a | L = \text{color}_1) + \hat{E}(Y^a | L = \text{color}_2) + \hat{E}(Y^a | L = \text{color}_3) \right. \\ & + \hat{E}(Y^a | L = \text{color}_4) + \hat{E}(Y^a | L = \text{color}_5) + \hat{E}(Y^a | L = \text{color}_6) \\ & \left. + \hat{E}(Y^a | L = \text{color}_7) + \hat{E}(Y^a | L = \text{color}_8) \right) / 8 \end{aligned}$$

Summing over individuals

Aggregate the average over sub-groups to get the overall average

$$\begin{aligned}\hat{E}(Y^a) &= \sum_{\ell} \underbrace{\hat{E}(Y^a \mid L = \ell)}_{\text{Avg of sub-group}} \times \underbrace{\hat{Pr}(L = \ell)}_{\text{Prob of sub-group}} \\ &= \frac{1}{n} \sum_i \underbrace{\hat{E}(Y^a \mid L = \ell_i)}_{\text{Avg of sub-group for unit } i}\end{aligned}$$

Estimating from data

Two small tweaks

- ▶ Estimate $Pr(L = \ell)$ use observed proportion of people with $L_i = \ell$

$$\widehat{Pr}(L = \ell)$$

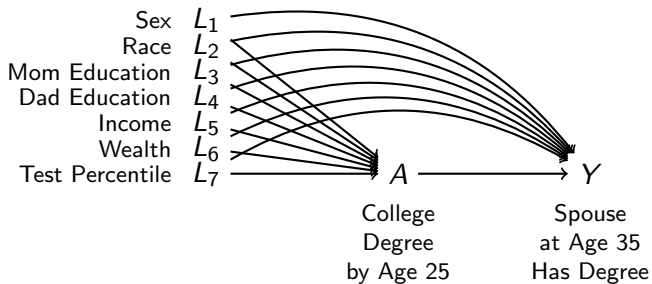
- ▶ Rewrite as an average across individuals (instead of groups)
- ▶ Estimate $E(Y^{a=1} \mid L = \ell)$, use observed average outcome for people with $A_i = 1, L_i = \ell$

$$\hat{E}(Y^{a=1} \mid L = \ell) = \hat{E}(Y \mid A = 1, L = \ell)$$

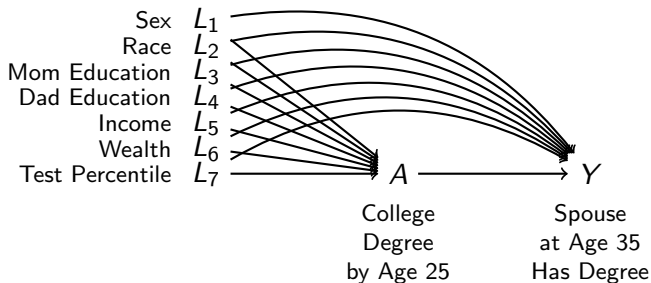
- ▶ Fit a model to estimate

$$\hat{E}(Y^{a=1} \mid L = \ell) = \hat{E}(Y \mid A = 1, L = \ell)$$

Nonparametric estimation breaks down



Nonparametric estimation breaks down



We need both:

$$\hat{E}(Y \mid A = 1, L = \ell_i)$$

$$\hat{E}(Y \mid A = 0, L = \ell_i)$$

Nonparametric estimation breaks down

Can't estimate $\hat{E}(Y \mid L = \ell_i, A = a)$ by taking an average for every sub-group

Hispanic		Non-Hispanic Black		Non-Hispanic Non-Black			
No dad	No college	No dad	No college	No dad	No college	No dad	No mom
< HS	No college	< HS	No college	< HS	No college	< HS	No mom
High school	No college	High school	No college	High school	No college	High school	No mom
Some college	No college	Some college	No college	Some college	No college	Some college	No mom
College	No college	College	No college	College	No college	College	No mom
No dad	< HS	No dad	< HS	No dad	< HS	No dad	< HS
< HS	< HS	< HS	< HS	< HS	< HS	< HS	< HS
High school	< HS	High school	< HS	High school	< HS	High school	< HS
Some college	< HS	Some college	< HS	Some college	< HS	Some college	< HS
College	< HS	College	< HS	College	< HS	College	< HS
No dad	High school	No dad	High school	No dad	High school	No dad	High school
< HS	High school	< HS	High school	< HS	High school	< HS	High school
High school	High school	High school	High school	High school	High school	High school	High school
Some college	High school	Some college	High school	Some college	High school	Some college	High school
College	High school	College	High school	College	High school	College	High school
No dad	Some college	No dad	Some college	No dad	Some college	No dad	Some college
< HS	Some college	< HS	Some college	< HS	Some college	< HS	Some college
High school	Some college	High school	Some college	High school	Some college	High school	Some college
Some college	Some college	Some college	Some college	Some college	Some college	Some college	Some college
College	Some college	College	Some college	College	Some college	College	Some college
No dad	College	No dad	College	No dad	College	No dad	College
< HS	College	< HS	College	< HS	College	< HS	College
High school	College	High school	College	High school	College	High school	College
Some college	College	Some college	College	Some college	College	Some college	College
College	College	College	College	College	College	College	College

Nonparametric estimation breaks down

Can't estimate $\hat{E}(Y \mid L = \ell_i, A = a)$ by taking an average for every sub-group

Hispanic		Non-Hispanic Black		Non-Hispanic Non-Black			
No dad	No dad	No dad	No dad	No dad	No dad	No dad	No mom
< HS	< HS	< HS	< HS	< HS	< HS	< HS	No mom
High school	High school	High school	High school	High school	High school	High school	No mom
Some college	Some college	Some college	Some college	Some college	Some college	Some college	No mom
College	College	College	College	College	College	College	No mom
No dad	No dad	No dad	No dad	No dad	No dad	No dad	< HS
< HS	< HS	< HS	< HS	< HS	< HS	< HS	< HS
High school	High school	High school	High school	High school	High school	High school	< HS
Some college	Some college	Some college	Some college	Some college	Some college	Some college	< HS
College	College	College	College	College	College	College	< HS
No dad	No dad	No dad	No dad	No dad	No dad	No dad	High school
< HS	< HS	< HS	< HS	< HS	< HS	< HS	High school
High school	High school	High school	High school	High school	High school	High school	High school
Some college	Some college	Some college	Some college	Some college	Some college	Some college	High school
College	College	College	College	College	College	College	High school
No dad	No dad	No dad	No dad	No dad	No dad	No dad	Some college
< HS	< HS	< HS	< HS	< HS	< HS	< HS	Some college
High school	High school	High school	High school	High school	High school	High school	Some college
Some college	Some college	Some college	Some college	Some college	Some college	Some college	Some college
College	College	College	College	College	College	College	Some college
No dad	No dad	No dad	No dad	No dad	No dad	No dad	College
< HS	< HS	< HS	< HS	< HS	< HS	< HS	College
High school	High school	High school	High school	High school	High school	High school	College
Some college	Some college	Some college	Some college	Some college	Some college	Some college	College
College	College	College	College	College	College	College	College

We instead use a statistical model to estimate

$$\hat{E}(Y \mid L = \ell_i, A = a)$$

Parametric estimation: Outcome model

Model the conditional expectation of Y given L and A

- Linear regression

$$\hat{E}(Y \mid L, A) = \hat{\alpha} + L'\hat{\gamma} + A\hat{\beta}$$

Parametric estimation: Outcome model

Model the conditional expectation of Y given L and A

- ▶ Linear regression

$$\hat{E}(Y \mid L, A) = \hat{\alpha} + L'\hat{\gamma} + A\hat{\beta}$$

$$\hat{E}(Y_i \mid \text{Test Score}_i, A_i) = .2 + .003 \times \text{Test Score}_i + .2 \times A_i$$

- ▶ If $\text{Test Score}_i = 80$ and $A_i = 1$ then

$$\hat{E}(Y \mid \text{Test Score}, A) = .2 + .003 \times (80) + .2(1) = .64$$

- ▶ If $\text{Test Score}_i = 62$ and $A_i = 0$ then

$$\hat{E}(Y \mid \text{Test Score}, A) = .2 + .003 \times (62) + .2(0) = 0.386$$

G-formula

Causal assumptions



Learn a model to predict Y given L and A

$$\hat{E}(Y \mid L, A) = \hat{\alpha} + L' \hat{\gamma} + A \hat{\beta}$$

G-formula

Causal assumptions



Learn a model to predict Y given L and A

$$\hat{E}(Y \mid L, A) = \hat{\alpha} + L' \hat{\gamma} + A \hat{\beta}$$

Standardization estimator

$$\hat{E}(Y^a) = \frac{1}{n} \sum_i \hat{E}(Y \mid L = \ell_i, A = a)$$

G-formula

Causal assumptions



Learn a model to predict Y given L and A

$$\hat{E}(Y \mid L, A) = \hat{\alpha} + L'\hat{\gamma} + A\hat{\beta}$$

Standardization estimator

$$\hat{E}(Y^a) = \frac{1}{n} \sum_i \hat{E}(Y \mid L = \ell_i, A = a)$$

For every unit i ,

- ▶ Set the treatment value to a
- ▶ Predict the outcome

Then average over all units

The parametric g-formula: Connection to $\hat{\beta}$

The parametric g-formula: Connection to $\hat{\beta}$

Estimator for the effect $E(Y^1) - E(Y^0)$:

The parametric g-formula: Connection to $\hat{\beta}$

Estimator for the effect $E(Y^1) - E(Y^0)$:

$$\begin{aligned}\hat{E}(Y^1) - \hat{E}(Y^0) &= \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\alpha} + \hat{\gamma} \ell_i + \hat{\beta} \times 1 \right) \right) \\ &\quad - \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\alpha} + \hat{\gamma} \ell_i + \hat{\beta} \times 0 \right) \right)\end{aligned}$$

The parametric g-formula: Connection to $\hat{\beta}$

Estimator for the effect $E(Y^1) - E(Y^0)$:

$$\begin{aligned}\hat{E}(Y^1) - \hat{E}(Y^0) &= \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\alpha} + \hat{\gamma} \ell_i + \hat{\beta} \times 1 \right) \right) \\ &\quad - \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\alpha} + \hat{\gamma} \ell_i + \hat{\beta} \times 0 \right) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \hat{\beta}\end{aligned}$$

The parametric g-formula: Connection to $\hat{\beta}$

Estimator for the effect $E(Y^1) - E(Y^0)$:

$$\begin{aligned}\hat{E}(Y^1) - \hat{E}(Y^0) &= \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\alpha} + \hat{\gamma} \ell_i + \hat{\beta} \times 1 \right) \right) \\ &\quad - \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\alpha} + \hat{\gamma} \ell_i + \hat{\beta} \times 0 \right) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \hat{\beta} \\ &= \hat{\beta}\end{aligned}$$

The parametric g-formula: Connection to $\hat{\beta}$

Estimator for the effect $E(Y^1) - E(Y^0)$:

$$\begin{aligned}\hat{E}(Y^1) - \hat{E}(Y^0) &= \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\alpha} + \hat{\gamma} \ell_i + \hat{\beta} \times 1 \right) \right) \\ &\quad - \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\alpha} + \hat{\gamma} \ell_i + \hat{\beta} \times 0 \right) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \hat{\beta} \\ &= \hat{\beta}\end{aligned}$$

With OLS, the parametric g-formula collapses on the coefficient.

The parametric g-formula allows for more complex models

As long as we have a model for predicting $E(Y \mid L, A)$, we can apply the g-formula

- ▶ Linear models with interaction terms
- ▶ Other types of regression: logistic regression, poisson regression, etc
- ▶ Machine learning models
 - ▶ Deep Neural Networks
 - ▶ Random forests
 - ▶ etc

Bias Variance trade-off

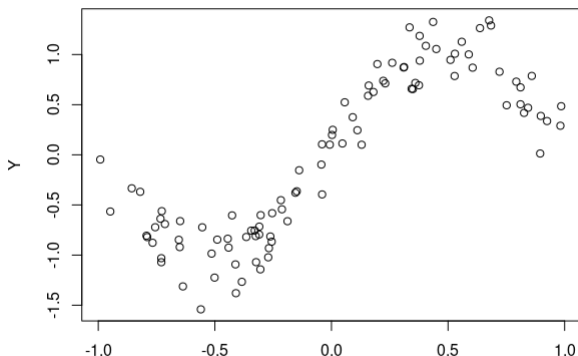
In statistics, the bias variance trade off is a fundamental constraint

- ▶ **Bias:** The functions we may estimate are not complex enough to capture the “true relationship”
- ▶ **Variance:** The model we are fitting is too complex so our estimated parameters change a lot from sample to sample

Bias Variance trade-off

In statistics, the bias variance trade off is a fundamental constraint

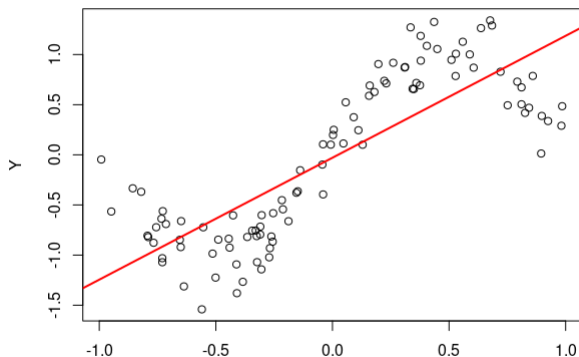
- **Bias:** The functions we may estimate are not complex enough to capture the “true relationship”
- **Variance:** The model we are fitting is too complex so our estimated parameters change a lot from sample to sample



Bias Variance trade-off

In statistics, the bias variance trade off is a fundamental constraint

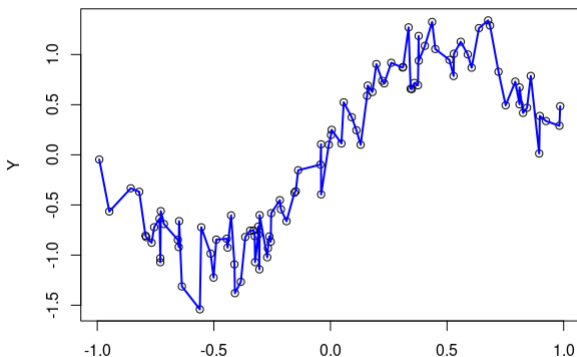
- **Bias:** The functions we may estimate are not complex enough to capture the “true relationship”
- **Variance:** The model we are fitting is too complex so our estimated parameters change a lot from sample to sample



Bias Variance trade-off

In statistics, the bias variance trade off is a fundamental constraint

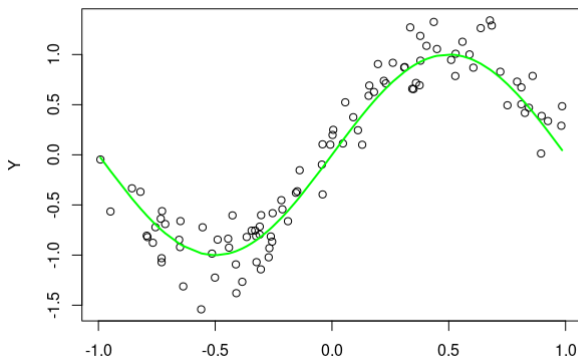
- **Bias:** The functions we may estimate are not complex enough to capture the “true relationship”
- **Variance:** The model we are fitting is too complex so our estimated parameters change a lot from sample to sample



Bias Variance trade-off

In statistics, the bias variance trade off is a fundamental constraint

- **Bias:** The functions we may estimate are not complex enough to capture the “true relationship”
- **Variance:** The model we are fitting is too complex so our estimated parameters change a lot from sample to sample



Bias Variance trade-off

Bias and variance in making cakes:



Figure: High Bias, low variance



Figure: Low bias, High variance

Bias and variance in choosing conditional expectation model

- ▶ Linear model: 1 parameter per covariate (probably high bias)
- ▶ Observed average: 2^p means to estimate for p binary covariates (probably high variance)
- ▶ Other methods are typically somewhere in between
- ▶ Larger sample allows for more complex models

Bias and variance in choosing causal model

- ▶ Is a a DAG ever “truly correct”?
- ▶ Can always add more confounders
- ▶ Would the bias from the confounders you could add substantially change your claim?
- ▶ Including additional confounders makes estimation more difficult

Parametric g-formula: Outcome model recap

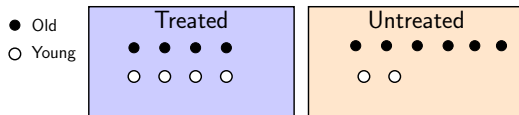


1. Estimate the outcome mean $E(Y \mid A, L)$ with some model
2. Change everyone's treatment to the value of interest
3. Predict for everyone
4. Take the average

$$\hat{E}(Y^a) = \frac{1}{n} \sum_{i=1}^n \hat{E}(Y \mid L = \ell_i, A = a)$$

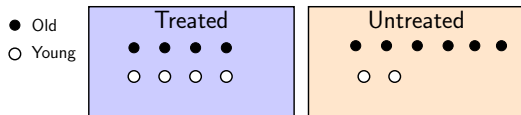
Inverse probability weighting

Observe: A conditionally randomized experiment:



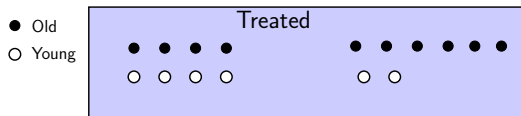
Inverse probability weighting

Observe: A conditionally randomized experiment:



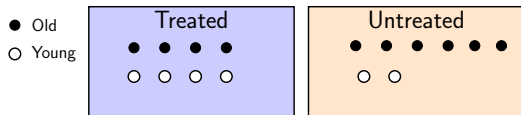
In order to get the $ATE = E(Y^{a=1}) - E(Y^{a=0})$

Want: a world where everyone is treated



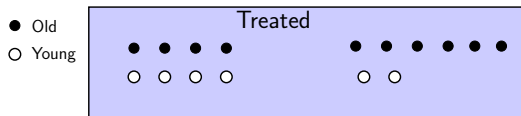
Inverse probability weighting

Observe: A conditionally randomized experiment:



In order to get the $ATE = E(Y^{a=1}) - E(Y^{a=0})$

Want: a world where everyone is treated



Want: a world where everyone is not treated



IPW: Everything together

General rule: Each individual should be reweighted by

$$\text{Weight} = \frac{1}{\text{Prob of their obs treatment status given their age}}$$

- ▶ Propensity Score: $\pi_i = \text{Pr}(A_i = 1 \mid L_i)$
- ▶ Let N be the number of observations in the hypothetical world (which is the same as the number of actual observations)
- ▶ Hypothetical world where everyone is treated

$$E(Y^{a=1}) = \frac{1}{N} \sum_{i:A_i=1} \frac{1}{\pi_i} \times Y_i$$

- ▶ Hypothetical world where everyone is not treated

$$E(Y^{a=0}) = \frac{1}{N} \sum_{i:A_i=0} \frac{1}{1 - \pi_i} \times Y_i$$

$$\text{ATE} = \underbrace{\frac{1}{N} \sum_{i:A_i=1} \frac{1}{\pi_i} \times Y_i}_{\text{Hypothetical world where everyone is treated}} - \underbrace{\frac{1}{N} \sum_{i:A_i=0} \frac{1}{1 - \pi_i} \times Y_i}_{\text{Hypothetical world where everyone is not treated}}$$

IPW: with data

Model the treatment assignment

$$\hat{\pi}_i = \hat{P}(A = 1 \mid L) = \text{logit}^{-1}(\hat{\alpha} + \hat{\gamma}L)$$

Estimate by inverse probability weighting

$$\hat{E}(Y^{a=1}) = \frac{1}{N} \sum_{i:A_i=1} \frac{Y_i}{\hat{\pi}_i}$$

$$\hat{E}(Y^{a=0}) = \frac{1}{N} \sum_{i:A_i=0} \frac{Y_i}{1 - \hat{\pi}_i}$$

Learning goals for today

At the end of class, you will be able to

- ▶ Estimate average causal effects via g-formula with a statistical model for the outcome and treatment
- ▶ Reason about the bias variance tradeoff when selecting a statistical model