# Statistical Modeling: Propensity modeling

Cornell STSCI / INFO / ILRST 3900
causal3900.github.io

Oct 22, 2026

# Learning goals for today

At the end of class, you will be able to

- Estimate average causal effects with a model for the treatment
- Reason about the bias variance tradeoff
- Use the Augmented IPW estimator to guard against model misspecification

After class:

- Hernán and Robins 2020 Chapter 12.1–12.2, Technical Points 13.2 and 13.3

# G-formula: Outcome model

Standardization estimator

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

# G-formula: Outcome model

Standardization estimator

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

Learn a model to predict $Y$ given $L$ and $A$

- Linear models potentially with interaction terms
- Other types of regression: logistic regression, poisson regression, etc
- Machine learning models

# G-formula: Outcome model

Standardization estimator

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

Learn a model to predict $Y$ given $L$ and $A$

- Linear models potentially with interaction terms
- Other types of regression: logistic regression, poisson regression, etc
- Machine learning models

For every unit $i$,

- Set the treatment value to $A = 1$
- Predict the outcome

Then average over all units

# Bias Variance trade-off

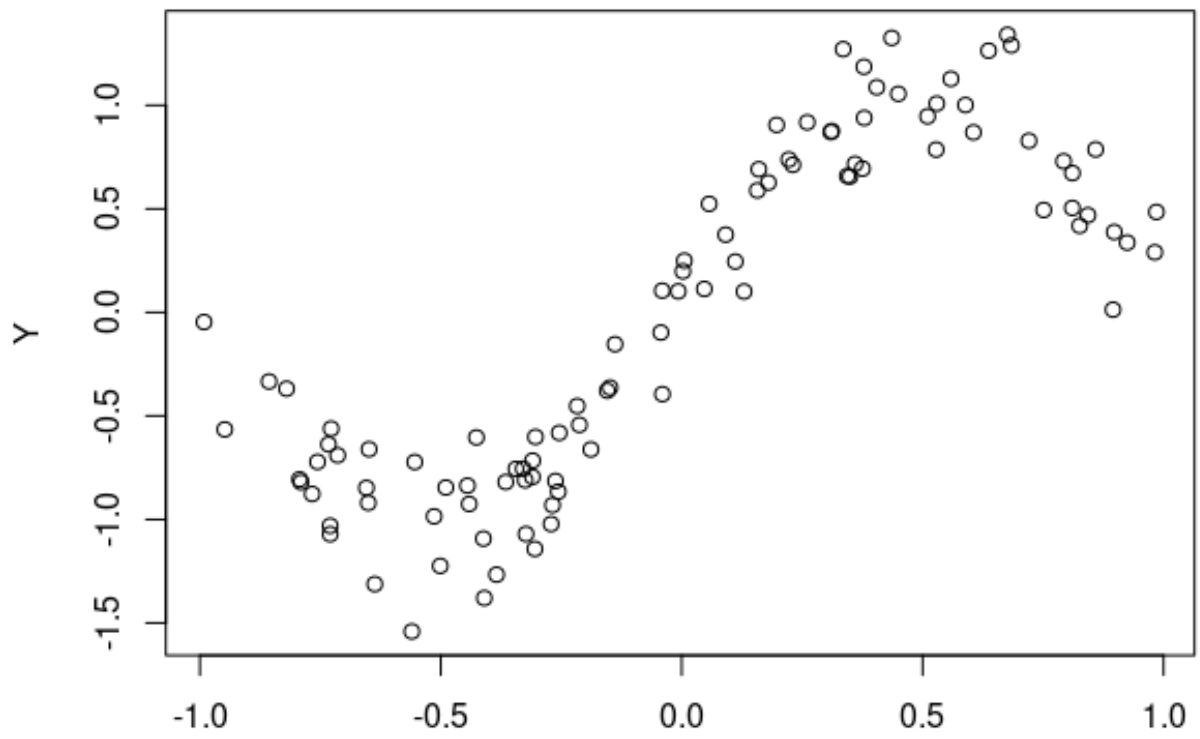
Bias and variance in making cakes:

Figure: High Bias, low variance

Figure: Low bias, High variance

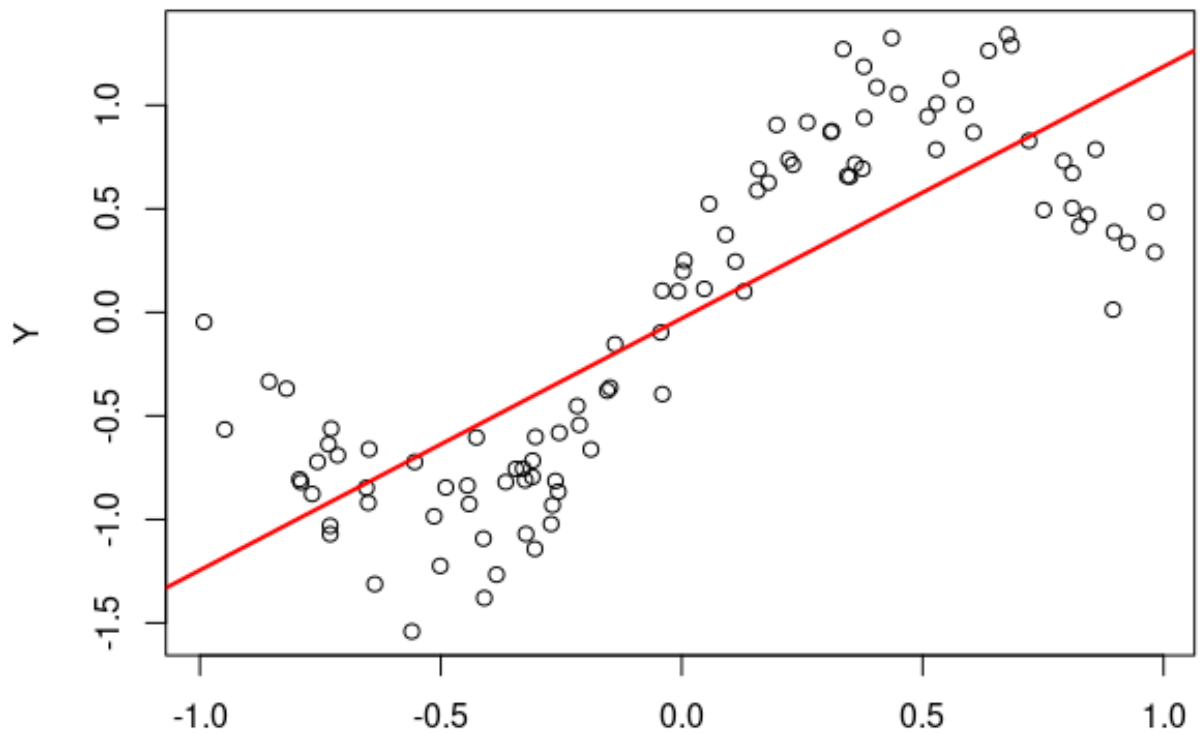
# Bias and variance in choosing causal model

- More complex models require a larger sample size to fit well
- Roughly speaking, more complex means that the model is more difficult to describe
- More complex can also mean we allow for non-linear relationships or also that we include more covariates



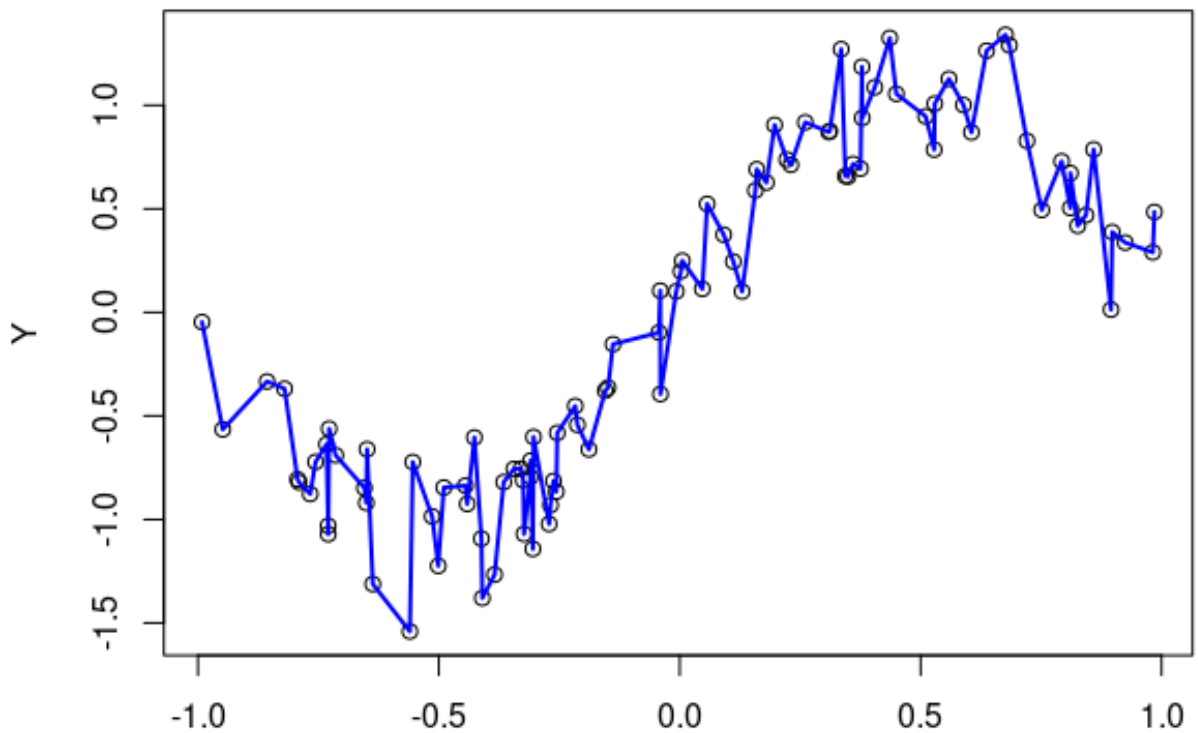
# Bias and variance in choosing causal model

- More complex models require a larger sample size to fit well
- Roughly speaking, more complex means that the model is more difficult to describe
- More complex can also mean we allow for non-linear relationships or also that we include more covariates



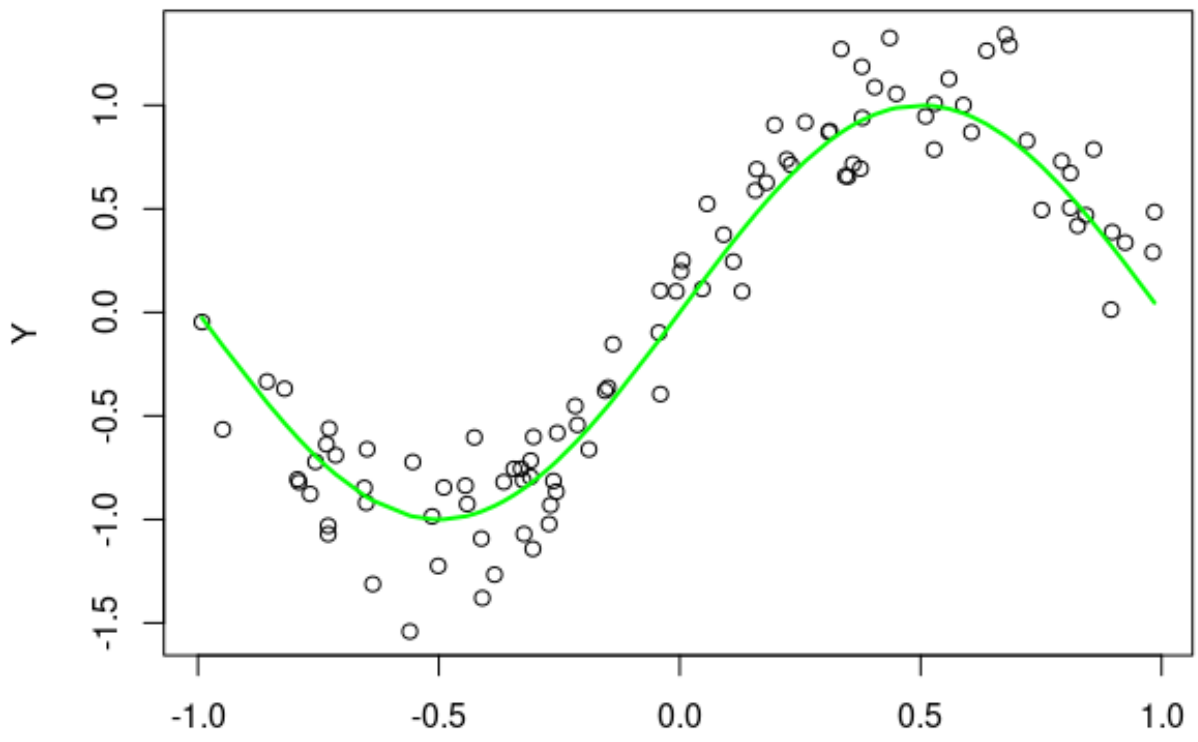
# Bias and variance in choosing causal model

- More complex models require a larger sample size to fit well
- Roughly speaking, more complex means that the model is more difficult to describe
- More complex can also mean we allow for non-linear relationships or also that we include more covariates



# Bias and variance in choosing causal model

- More complex models require a larger sample size to fit well
- Roughly speaking, more complex means that the model is more difficult to describe
- More complex can also mean we allow for non-linear relationships or also that we include more covariates

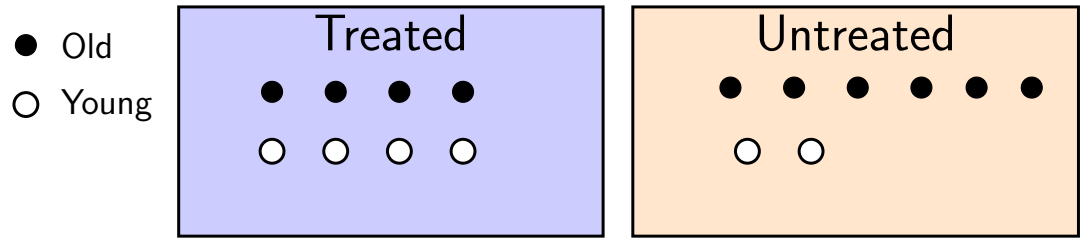


# Bias and variance in choosing causal model

- 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, requires a larger sample to fit well

# Inverse probability weighting

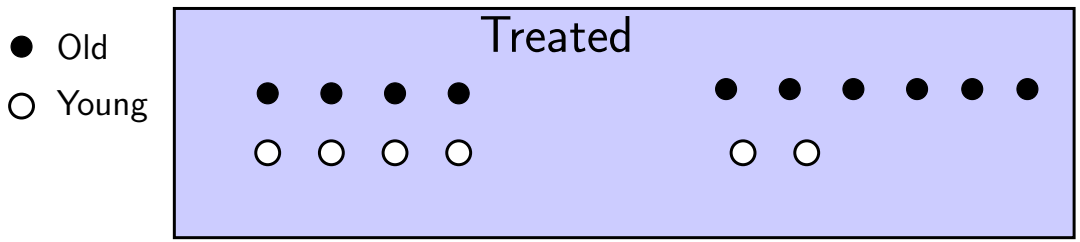
# Inverse probability weighting



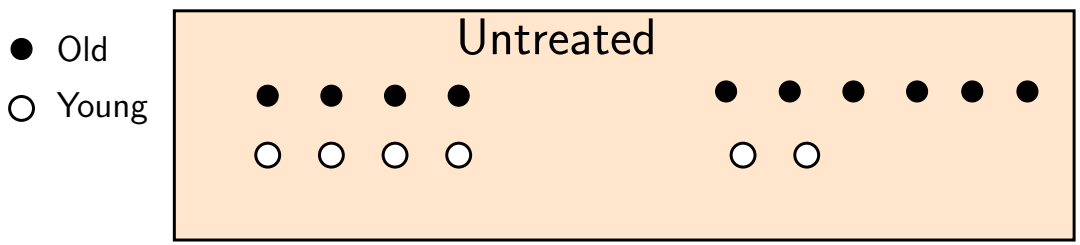

**Observe**:

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: $\pi_i = Pr(A_i = 1 \mid L_i)$
- Let $N$ be the number of total observations
- Hypothetical world where everyone is treated

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

- Hypothetical world where everyone is not treated

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

- $$ATE = \underbrace{\frac{1}{N} \sum_{i:A_i=1} \frac{1}{\pi_i} \times Y_i}_{\text{Weighted avg of treated obs}} - \underbrace{\frac{1}{N} \sum_{i:A_i=0} \frac{1}{1-\pi_i} \times Y_i}_{\text{Weighted Avg of untreated obs}}$$

# IPW: with data

**Model** the treatment assignment

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

**Estimate** by inverse probability weighting

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

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

# Inverse probability of treatment weighting

$$
\begin{aligned}
\hat{\mathrm{E}}(Y^{a=1}) &= \frac{1}{N} \sum_{i:A_i=1} \frac{Y_i}{\hat{\pi}_i} \\
&= \frac{1}{N} \sum_{i:A_i=1} \frac{A_i Y_i}{\hat{\pi}_i} && \text{(people with } A_i = 1\text{)} \\
&= \frac{1}{N} \left( \sum_{i:A_i=1} \frac{A_i Y_i}{\hat{\pi}_i} + \underbrace{\sum_{i:A_i=0} \frac{A_i Y_i}{\hat{\pi}_i}}_{=0} \right) && \text{(people with } A_i = 0\text{)} \\
&= \frac{1}{N} \sum_{i} \frac{A_i Y_i}{\hat{\pi}_i} && \text{(everyone)}
\end{aligned}
$$

(1)

# Inverse probability of treatment weighting

$$
\begin{aligned}
\hat{\mathrm{E}}(Y^{a=0}) &= \frac{1}{N} \sum_{i:A_i=0} \frac{Y_i}{1-\hat{\pi}_i} \\
&= \frac{1}{N} \sum_{i:A_i=0} \frac{(1-A_i)Y_i}{\hat{\pi}_i} && \text{(people with } A_i = 0\text{)} \\
&= \frac{1}{N} \left( \sum_{i:A_i=0} \frac{(1-A_i)Y_i}{1-\hat{\pi}_i} + \underbrace{\sum_{i:A_i=1} \frac{(1-A_i)Y_i}{\hat{\pi}_i}}_{=0} \right) && \text{(people with } A_i = 1\text{)} \\
&= \frac{1}{N} \sum_{i} \frac{(1-A_i)Y_i}{1-\hat{\pi}_i} && \text{(everyone)}
\end{aligned}
\tag{2}
$$

# Parametric model: propensity model

**Model** the treatment assignment

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

**Estimate** by inverse probability weighting (IPW)

$$\hat{\mathrm{E}}(Y^1) - \hat{\mathrm{E}}(Y^0) = \frac{1}{N}\left(\sum_i \frac{A_i Y_i}{\hat{\pi}_i} - \sum_i \frac{(1 - A_i) Y_i}{1 - \hat{\pi}_i}\right)$$

# Standardization vs Propensity score modeling

- When our predictions are exactly right, the two are equivalent
  - $\hat{\mathrm{E}}(Y \mid A, L) = \mathrm{E}(Y \mid A, L)$
  - $\hat{P}(A = a \mid L) = P(A = a \mid L)$
- Standardization (g-formula) is probably more popular, but it's not inherently better
- If outcomes easier to predict than treatment status, use standardization
- If treatment is easier to predict than outcomes, use IPW

# Standardization vs Propensity score modeling

- When our predictions are exactly right, the two are equivalent
  - $\hat{\mathrm{E}}(Y \mid A, L) = \mathrm{E}(Y \mid A, L)$
  - $\hat{P}(A = a \mid L) = P(A = a \mid L)$
- Standardization (g-formula) is probably more popular, but it's not inherently better
- If outcomes easier to predict than treatment status, use standardization
- If treatment is easier to predict than outcomes, use IPW
- What if our statistical models are wrong?
  - The statistical model is not complex enough to accurately capture the true relationship?
  - What if we don't include the right covariates?

# Augmented IPW

- Augmented IPW uses both estimates for outcome and propensity scores together
- Has a "double robustness" property

$$\hat{E}(Y^1) = \frac{1}{N}\left(\underbrace{\sum_i \hat{E}(Y \mid A=1, L=\ell_i)}_{\text{g-computation}} + \underbrace{\sum_i \frac{A_i Y_i}{\hat{\pi}_i}}_{\text{IPW}} - \underbrace{\sum_i \frac{A_i}{\hat{\pi}_i}\hat{E}(Y \mid A=1, L=\ell_i)}_{\text{new term}}\right)$$

# Augmented IPW

Why is this a good idea?

- Standardization with the true conditional expectation would recover $\mathrm{E}(Y^{a=1})$ when the sample size is large

$$\frac{1}{N}\sum_i \mathrm{E}(Y \mid A = 1, L = \ell_i)$$

- AIPW estimates $\mathrm{E}(Y^{a=1})$ by

$$\hat{\mathrm{E}}(Y^{a=1}) = \frac{1}{N}\sum_i \left(\hat{\mathrm{E}}(Y \mid A = 1, L = \ell_i) + \frac{A_i}{\hat{\pi}_i} Y_i - \frac{A_i}{\hat{\pi}_i}\hat{E}(Y \mid A = 1, L = \ell_i)\right)$$

# Augmented IPW

Why is this a good idea?

- Standardization with the true conditional expectation would recover $\mathrm{E}(Y^{a=1})$ when the sample size is large

$$\frac{1}{N}\sum_i \mathrm{E}(Y \mid A = 1, L = \ell_i)$$

- AIPW estimates $\mathrm{E}(Y^{a=1})$ by

$$\hat{\mathrm{E}}(Y^{a=1}) = \frac{1}{N}\sum_i \left(\hat{\mathrm{E}}(Y \mid A = 1, L = \ell_i) + \frac{A_i}{\hat{\pi}_i} Y_i - \frac{A_i}{\hat{\pi}_i}\hat{E}(Y \mid A = 1, L = \ell_i)\right)$$

- If the two terms inside are close to each other, then the AIPW will recover $\mathrm{E}(Y^{a=1})$ when the sample size is large

# Augmented IPW

- We want the expectation of the AIPW term to be close to the true conditional expectation

# Augmented IPW

- We want the expectation of the AIPW term to be close to the true conditional expectation
- Note that $\mathrm{E}(A_i) = Pr(A_i = 1 \mid L_i) = \pi_i$
- Fixing the predictions $\hat{\mathrm{E}}(\cdot)$ and $\hat{\pi}$, take the expectation of AIPW term

$$\mathrm{E}\left(\hat{E}(Y \mid A = 1, L = \ell_i) + \frac{A_i}{\hat{\pi}_i} Y_i - \frac{A_i}{\hat{\pi}_i} \hat{E}(Y \mid A = 1, L = \ell_i)\right)$$

$$= \hat{\mathrm{E}}(Y \mid A = 1, L = \ell_i) + \frac{\pi_i}{\hat{\pi}_i} \mathrm{E}(Y \mid A = 1, L = \ell_i) - \frac{\pi_i}{\hat{\pi}_i} \hat{E}(Y \mid A = 1, L = \ell_i)$$

# Augmented IPW

- We want the expectation of the AIPW term to be close to the true conditional expectation
- Note that $\mathrm{E}(A_i) = Pr(A_i = 1 \mid L_i) = \pi_i$
- Fixing the predictions $\hat{\mathrm{E}}(\cdot)$ and $\hat{\pi}$, take the expectation of AIPW term

$$\mathrm{E}\left(\hat{E}(Y \mid A = 1, L = \ell_i) + \frac{A_i}{\hat{\pi}_i} Y_i - \frac{A_i}{\hat{\pi}_i} \hat{E}(Y \mid A = 1, L = \ell_i)\right)$$

$$= \hat{\mathrm{E}}(Y \mid A = 1, L = \ell_i) + \frac{\pi_i}{\hat{\pi}_i} \mathrm{E}(Y \mid A = 1, L = \ell_i) - \frac{\pi_i}{\hat{\pi}_i} \hat{E}(Y \mid A = 1, L = \ell_i)$$

$$= \hat{\mathrm{E}}(Y \mid A = 1, L = \ell_i) + \frac{\pi_i}{\hat{\pi}_i} \left[\mathrm{E}(Y \mid A = 1, L = \ell_i) - \hat{E}(Y \mid A = 1, L = \ell_i)\right]$$

# Augmented IPW

If the predicted outcome is correct:
$\hat{E}(Y \mid A = 1, L = \ell_i) = E(Y \mid A = 1, L = \ell_i)$;

AIPW term

$$
\begin{aligned}
&= \hat{\mathrm{E}}(Y \mid A = 1, L = \ell_i) + \frac{\pi_i}{\hat{\pi}_i}\left[\mathrm{E}(Y \mid A = 1, L = \ell_i) - \hat{E}(Y \mid A = 1, L = \ell_i)\right] \\
&= \mathrm{E}(Y \mid A = 1, L = \ell_i) + \frac{\pi_i}{\hat{\pi}_i}\underbrace{[\mathrm{E}(Y \mid A = 1, L = \ell_i) - E(Y \mid A = 1, L = \ell_i)]}_{=0} \\
&= \mathrm{E}(Y \mid A = 1, L = \ell_i)
\end{aligned}
\tag{3}
$$

# Augmented IPW

If the predicted outcome is correct:
$\hat{E}(Y \mid A = 1, L = \ell_i) = E(Y \mid A = 1, L = \ell_i)$;

AIPW term

$$
\begin{aligned}
&= \hat{\mathrm{E}}(Y \mid A = 1, L = \ell_i) + \frac{\pi_i}{\hat{\pi}_i}\left[\mathrm{E}(Y \mid A = 1, L = \ell_i) - \hat{E}(Y \mid A = 1, L = \ell_i)\right] \\
&= \mathrm{E}(Y \mid A = 1, L = \ell_i) + \frac{\pi_i}{\hat{\pi}_i}\underbrace{[\mathrm{E}(Y \mid A = 1, L = \ell_i) - E(Y \mid A = 1, L = \ell_i)]}_{=0} \\
&= \mathrm{E}(Y \mid A = 1, L = \ell_i)
\end{aligned}
\tag{3}
$$

Regardless of how different $\hat{\pi}$ and $\pi$ are!

# Augmented IPW

If the predicted propensity score is correct $\hat{\pi} = \pi$:

AIPW term

$$
\begin{aligned}
&= \hat{\mathrm{E}}(Y \mid A = 1, L = \ell_i) + \frac{\pi_i}{\hat{\pi}_i} \Big[ \mathrm{E}(Y \mid A = 1, L = \ell_i) - \hat{E}(Y \mid A = 1, L = \ell_i) \Big] \\
&= \hat{\mathrm{E}}(Y \mid A = 1, L = \ell_i) + \underbrace{\frac{\pi_i}{\pi_i}}_{=1} \Big[ \mathrm{E}(Y \mid A = 1, L = \ell_i) - \hat{E}(Y \mid A = 1, L = \ell_i) \Big] \\
&= \hat{\mathrm{E}}(Y \mid A = 1, L = \ell_i) + \Big[ \mathrm{E}(Y \mid A = 1, L = \ell_i) - \hat{E}(Y \mid A = 1, L = \ell_i) \Big] \\
&= E(Y \mid A = 1, L = \ell_i)
\end{aligned}
\tag{4}
$$

# Augmented IPW

If the predicted propensity score is correct $\hat{\pi} = \pi$:

AIPW term

$$
\begin{aligned}
&= \hat{\mathrm{E}}(Y \mid A = 1, L = \ell_i) + \frac{\pi_i}{\hat{\pi}_i}\Big[\mathrm{E}(Y \mid A = 1, L = \ell_i) - \hat{E}(Y \mid A = 1, L = \ell_i)\Big] \\
&= \hat{\mathrm{E}}(Y \mid A = 1, L = \ell_i) + \underbrace{\frac{\pi_i}{\pi_i}}_{=1}\Big[\mathrm{E}(Y \mid A = 1, L = \ell_i) - \hat{E}(Y \mid A = 1, L = \ell_i)\Big] \\
&= \hat{\mathrm{E}}(Y \mid A = 1, L = \ell_i) + \Big[\mathrm{E}(Y \mid A = 1, L = \ell_i) - \hat{E}(Y \mid A = 1, L = \ell_i)\Big] \\
&= E(Y \mid A = 1, L = \ell_i)
\end{aligned}
\tag{4}
$$

Regardless of how different $\hat{E}(Y \mid A = 1, L = \ell_i)$ and $E(Y \mid A = 1, L = \ell_i)$ are!

# Augmented IPW

- If you only use g-formula then outcome model must be correct
- If you only use IPW then propensity score model must be correct

# Augmented IPW

- If you only use g-formula then outcome model must be correct
- If you only use IPW then propensity score model must be correct
- AIPW recovers $\mathrm{E}(Y^1)$ if **either** the outcome model or the propensity score model is correct
- You get shots to be correct!
- One can be completely wrong as long as the other is correct

# Augmented IPW

Double robustness allows a single "failure"

[1]

[1] Daniel Craig in Knives Out

# Augmented IPW

Augmented IPW uses both estimates for outcome and propensity scores together

$$\hat{E}(Y^{a=1}) = \frac{1}{N}\left(\underbrace{\sum_i \hat{E}(Y \mid A=1, L=\ell_i)}_{\text{g-computation}} + \underbrace{\sum_i \frac{A_i Y_i}{\hat{\pi}_i}}_{\text{IPW}} - \underbrace{\sum_i \frac{A_i}{\hat{\pi}_i}\hat{E}(Y \mid A=1, L=\ell_i)}_{\text{new term}}\right)$$

$$\hat{E}(Y^{a=0}) = \frac{1}{N}\left(\underbrace{\sum_i \hat{E}(Y \mid A=0, L=\ell_i)}_{\text{g-computation}} + \underbrace{\sum_i \frac{(1-A_i) Y_i}{1-\hat{\pi}_i}}_{\text{IPW}} - \underbrace{\sum_i \frac{1-A_i}{1-\hat{\pi}_i}\hat{E}(Y \mid A=0, L=\ell_i)}_{\text{new term}}\right)$$

# Augmented IPW

Augmented IPW uses both estimates for outcome and propensity scores together

$$\hat{E}(Y^{a=1}) = \frac{1}{N}\left(\underbrace{\sum_i \hat{E}(Y \mid A=1, L=\ell_i)}_{\text{g-computation}} + \underbrace{\sum_i \frac{A_i Y_i}{\hat{\pi}_i}}_{\text{IPW}} - \underbrace{\sum_i \frac{A_i}{\hat{\pi}_i}\hat{E}(Y \mid A=1, L=\ell_i)}_{\text{new term}}\right)$$

$$\hat{E}(Y^{a=0}) = \frac{1}{N}\left(\underbrace{\sum_i \hat{E}(Y \mid A=0, L=\ell_i)}_{\text{g-computation}} + \underbrace{\sum_i \frac{(1-A_i) Y_i}{1-\hat{\pi}_i}}_{\text{IPW}} - \underbrace{\sum_i \frac{1-A_i}{1-\hat{\pi}_i}\hat{E}(Y \mid A=0, L=\ell_i)}_{\text{new term}}\right)$$

Estimate ATE

$$\hat{E}(Y^{a=1}) - \hat{E}(Y^{a=0})$$

# Augmented IPW

- Double robustness: robust against misspecification of either (but not both)
- If the outcome model is well specified, using standardization with just the outcome model often has smaller variance
- If the outcome model is not well specified, using standardization with just the outcome model may not be correct even as the sample size grows very large
- Using AIPW provides insurance against misspecification

# Learning goals for today

At the end of class, you will be able to

- Estimate average causal effects with a model for the treatment
- Reason about the bias variance tradeoff
- Use the Augmented IPW estimator to guard against model misspecification

After class:

- Hernán and Robins 2020 Chapter 12.1–12.2, Technical Points 13.2 and 13.3