Parametric g-Formula

Cornell STSCI / INFO / ILRST 3900
Fall 2024
causal3900.github.io

Oct 2 2025

Learning goals for today

At the end of class, you will be able to

- estimate average causal effects with a parametric model for the outcome $\mathrm{E}(Y \mid A, L)$ and treatment
- Reason about the bias variance tradeoff

After class:

- Hernán and Robins 2020 Chapter 12.1–12.5, 13, 15.1

Sample vs population

- Conditional Mean: Average outcome for individuals with specific characteristics

Descriptive	Causal
$E(Y \mid A = a, L = \ell)$	$E(Y^a \mid A = a, L = \ell)$

Sample vs population

- Conditional Mean: Average outcome for individuals with specific characteristics

Descriptive	Causal
$E(Y \mid A = a, L = \ell)$	$E(Y^a \mid A = a, L = \ell)$

- Population quantities: average outcome for **all units in the population** with specific characteristics

$$E(Y \mid A = a, L = \ell)$$

- Sample conditional mean: average outcome for **units in our sample** with specific characteristics

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

- Population quantities can be descriptive or causal
- Sample quantities can be descriptive or causal

Standardization

We use sample quantities to estimate population quantities

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

Standardization

We use sample quantities to estimate population quantities

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

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

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

$$\mathsf{E}(Y^{a=1}) - \mathsf{E}(Y^{a=0}) = \sum_{\ell} \left[\mathsf{E}(Y^{a=1} \mid L = \ell) - E(Y^{a=0} \mid L = \ell)\right] Pr(L = \ell)$$

Standardization

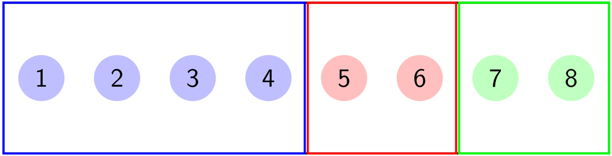
Aggregate the average over sub-groups to get the overall average

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

Standardization

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

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

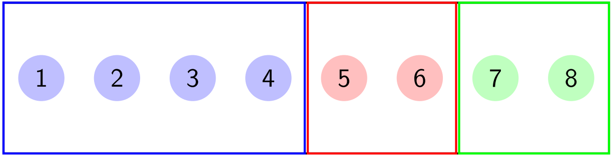


$$\hat{\mathrm{E}}(Y^a) = \hat{\mathrm{E}}(Y^a \mid L = {\color{blue}\bullet}) \times \frac{4}{8} + \hat{\mathrm{E}}(Y^a \mid L = {\color{red}\bullet}) \times \frac{2}{8} + \hat{\mathrm{E}}(Y^a \mid L = {\color{green}\bullet}) \times \frac{2}{8}$$

Standardization

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

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



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

Calculate the same quantity but sum over individuals

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

Standardization

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

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

Nonparametric estimation

Causal assumptions

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

Nonparametric estimation

Causal assumptions

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

Estimate population quantity with sample quantity

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

Nonparametric estimation

Causal assumptions

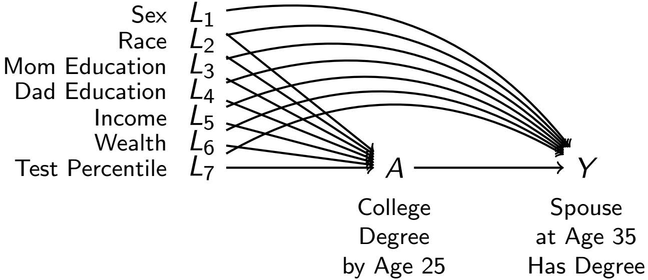
$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

Estimate population quantity with sample quantity

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

To estimate $\hat{E}(Y^{a=1}) - \hat{\mathrm{E}}(Y^{a=0})$ we need observations with both $A = 1$ and $A = 0$ for every observed ℓ_i

Nonparametric estimation breaks down



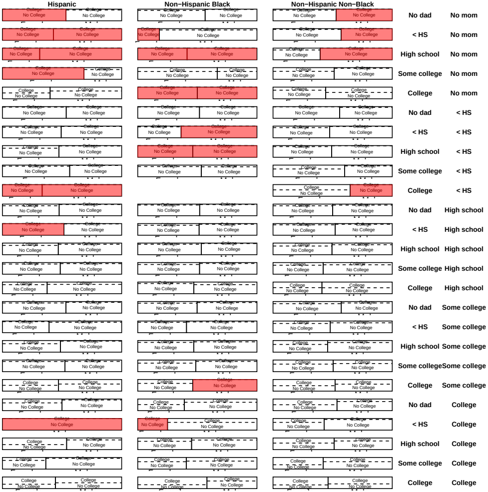
Sex L_1
Race L_2
Mom Education L_3
Dad Education L_4
Income L_5
Wealth L_6
Test Percentile L_7

A
College
Degree
by Age 25

Y
Spouse
at Age 35
Has Degree

Nonparametric estimation breaks down

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



Parametric estimation: Outcome model

Model the conditional expectation of Y given L and A

- Linear regression

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

Parametric estimation: Outcome model

Model the conditional expectation of Y given L and A

- Linear regression

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

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

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

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

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

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

Parametric estimation: Outcome model

Causal assumptions

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

Standardization estimator

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

Parametric estimation: Outcome model

Causal assumptions

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

Standardization estimator

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

Learn a parametric model to predict Y given L and A

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

Parametric estimation: Outcome model

Causal assumptions

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

Standardization estimator

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

Learn a parametric model to predict Y given L and A

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

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 $\mathrm{E}(Y^1) - \mathrm{E}(Y^0)$:

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

Estimator for the effect $\mathrm{E}(Y^1) - \mathrm{E}(Y^0)$:

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

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

Estimator for the effect $\mathrm{E}(Y^1) - \mathrm{E}(Y^0)$:

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

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

Estimator for the effect $\mathsf{E}(Y^1) - \mathsf{E}(Y^0)$:

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

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

Estimator for the effect $\text{E}(Y^1) - \text{E}(Y^0)$:

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

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

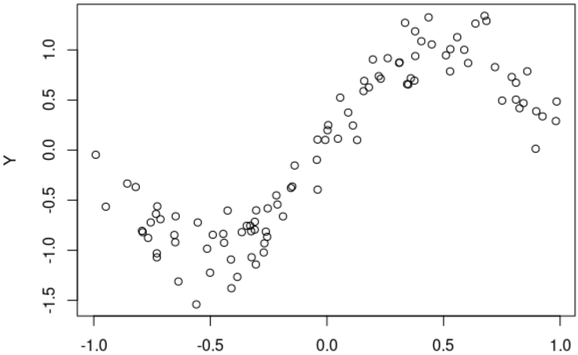
In statistics, the bias variance trade off is a fundamental constraint

- **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

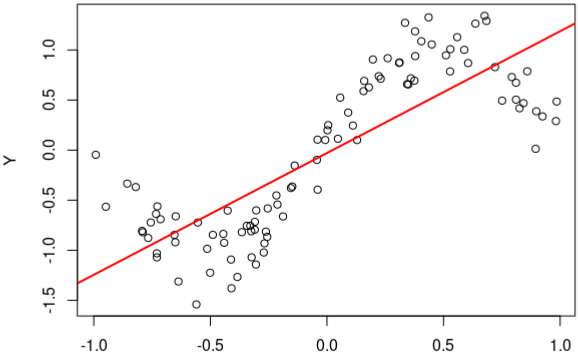
- **Bias**: The functions we may estimate are not complex enough to capture the "true relationship"
- **Variance**: The model we are fitting is too complex so our estimated parameters change a lot from sample to sample



Bias Variance trade-off

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

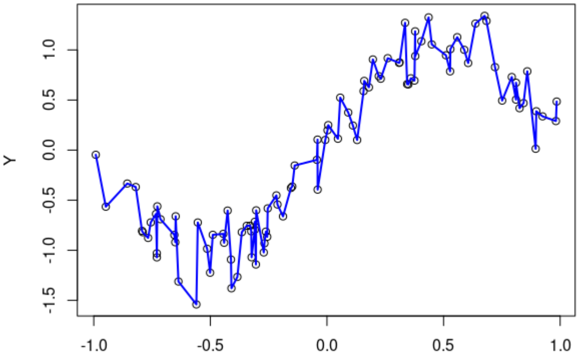
- **Bias**: The functions we may estimate are not complex enough to capture the "true relationship"
- **Variance**: The model we are fitting is too complex so our estimated parameters change a lot from sample to sample



Bias Variance trade-off

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

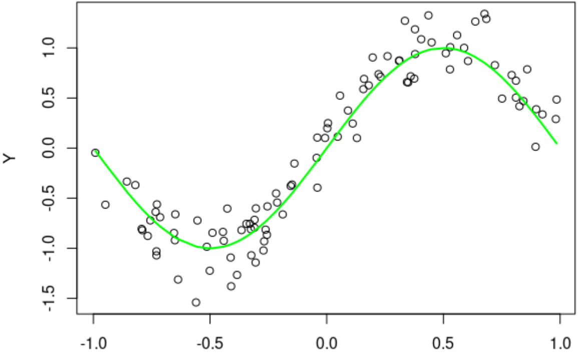
- **Bias**: The functions we may estimate are not complex enough to capture the "true relationship"
- **Variance**: The model we are fitting is too complex so our estimated parameters change a lot from sample to sample



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)
- Non-parametric estimate: 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

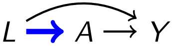
- Is a a DAG ever "truly correct"?
- Can always add more confounders
- Would the bias from the confounders you could add substantially change your claim?
- Including additional confounders makes estimation more difficult

Parametric g-formula: Outcome model recap

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

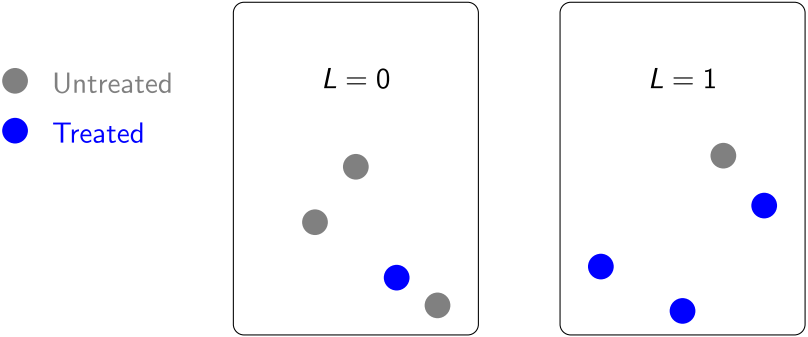
$$\hat{\mathrm{E}}(Y^a) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathrm{E}}(Y \mid L = \ell_i, A = a)$$

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$


L
A
Y


L
A
Y

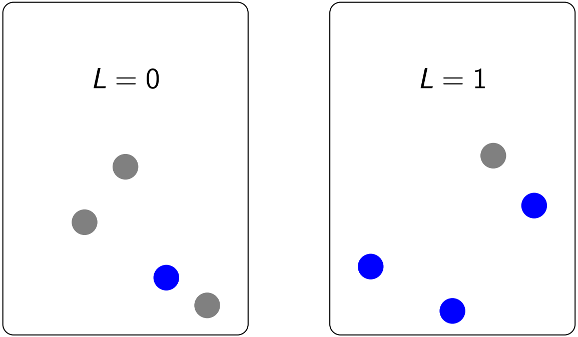
Inverse probability of treatment weighting



Inverse probability of treatment weighting

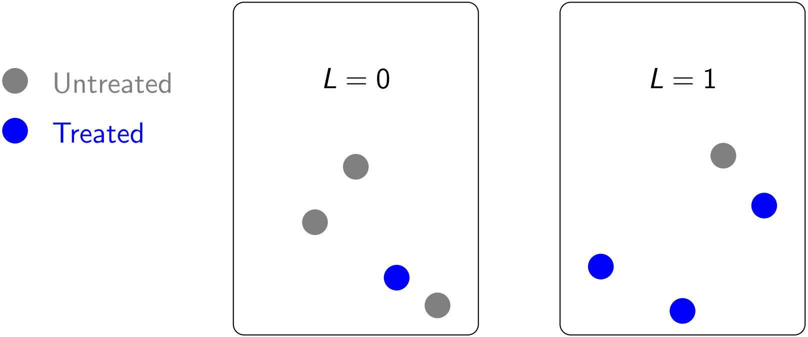
● Untreated

● Treated



Propensity score: $\pi_i = \mathrm{P}(A = A_i \mid L = L_i)$

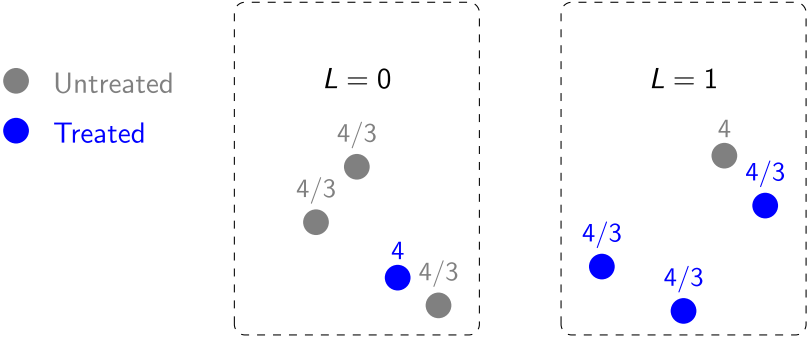
Inverse probability of treatment weighting



Propensity score: $\pi_i = \mathrm{P}(A = A_i \mid L = L_i)$

Inverse probability weight: $w_i = \frac{1}{\pi_i}$

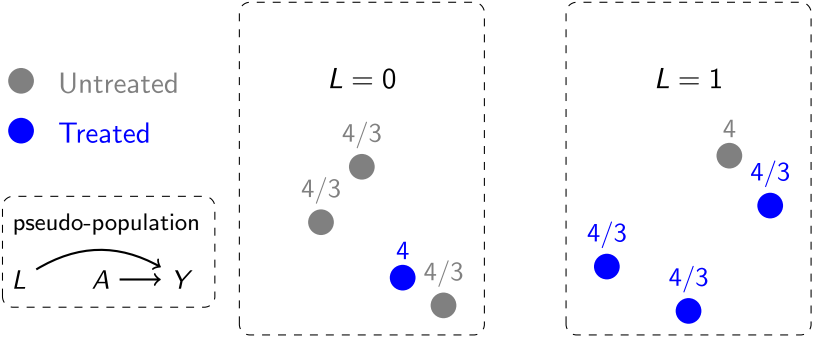
Inverse probability of treatment weighting



Propensity score: $\pi_i = \mathrm{P}(A = A_i \mid L = L_i)$

Inverse probability weight: $w_i = \frac{1}{\pi_i}$

Inverse probability of treatment weighting



Propensity score: $\pi_i = \mathrm{P}(A = A_i \mid L = L_i)$

Inverse probability weight: $w_i = \frac{1}{\pi_i}$

Model the treatment assignment

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

Predict the propensity score for each unit

$$\hat{\pi}_i = \begin{cases} \text{logit}^{-1}\left(\hat{\alpha} + \hat{\gamma}L\right) & \text{if } A_i = 1 \\ 1 - \text{logit}^{-1}\left(\hat{\alpha} + \hat{\gamma}L\right) & \text{if } A_i = 0 \end{cases}$$

Estimate by inverse probability weighting

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

Learning goals for today

At the end of class, you will be able to

- estimate average causal effects with a parametric model for the outcome $\mathrm{E}(Y \mid A, L)$ and treatment
- Reason about the bias variance tradeoff

After class:

- Hernán and Robins 2020 Chapter 12.1–12.5, 13, 15.1