Parametric Modeling: Propensity modeling

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

Oct 7, 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
- Use the Augmented IPW estimator to guard against model misspecification

After class:

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

Logistics

- Problem Set 4 due Oct 8
- Peer Review 3 due Oct 16
- Quiz 3 Oct 16
- Project Part 1 due Oct 20

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

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

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

Nonparametric estimation

Causal assumptions

Nonparametric estimation

Causal assumptions

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

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

Parametric estimation: Outcome model

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

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

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

Parametric estimation: Outcome model

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

- 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
- Predict the outcome

Then average over all units

Bias Variance trade-off

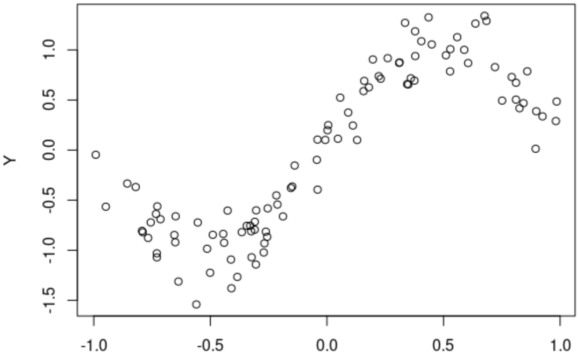
In statistics, the bias variance trade off is a fundamental constraint

- **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

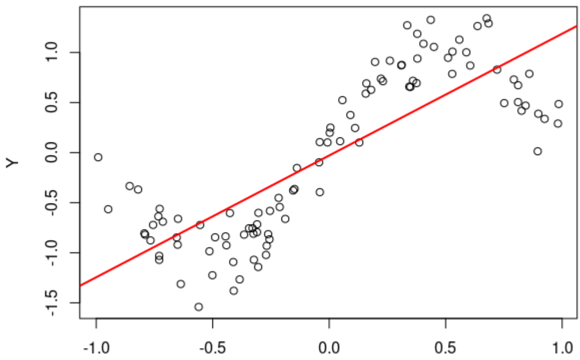
- **Bias**: The functions we may estimate are not complex enough to capture the "true relationship"
- **Variance**: The model we are fitting is too complex so our estimated parameters change a lot from sample to sample



Bias Variance trade-off

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

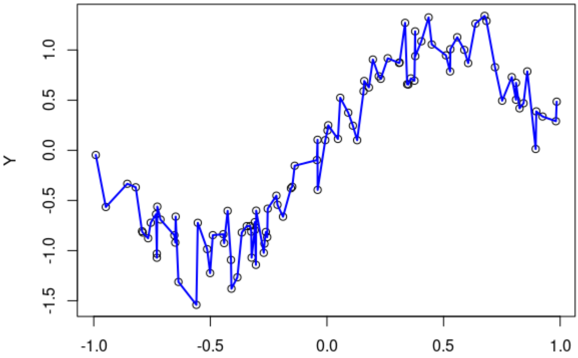
- **Bias**: The functions we may estimate are not complex enough to capture the "true relationship"
- **Variance**: The model we are fitting is too complex so our estimated parameters change a lot from sample to sample



Bias Variance trade-off

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

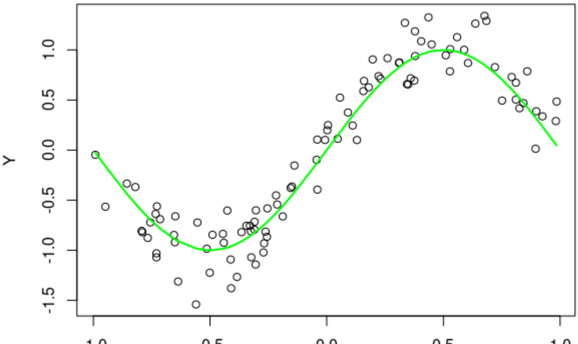
- **Bias**: The functions we may estimate are not complex enough to capture the "true relationship"
- **Variance**: The model we are fitting is too complex so our estimated parameters change a lot from sample to sample



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 (high bias, low variance)
- Non-parametric estimate: 2^p means to estimate for p binary covariates (low bias, 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

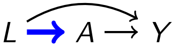
$$L \longrightarrow A \longrightarrow Y$$

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

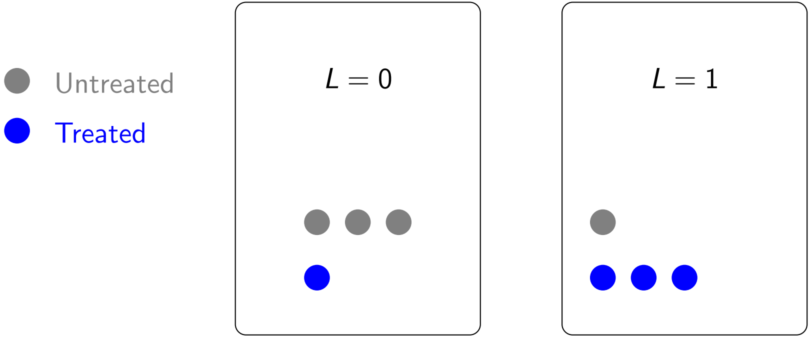
$$\hat{\text{E}}(Y^a) = \frac{1}{n} \sum_{i=1}^{n} \hat{\text{E}}(Y \mid L = \ell_i, A = a)$$

Outcome model

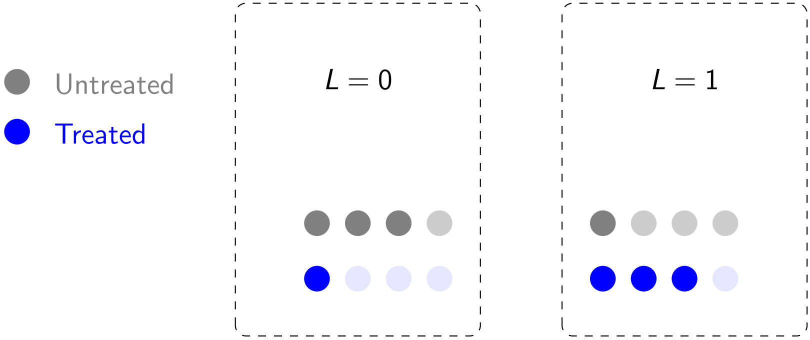
Propensity score model



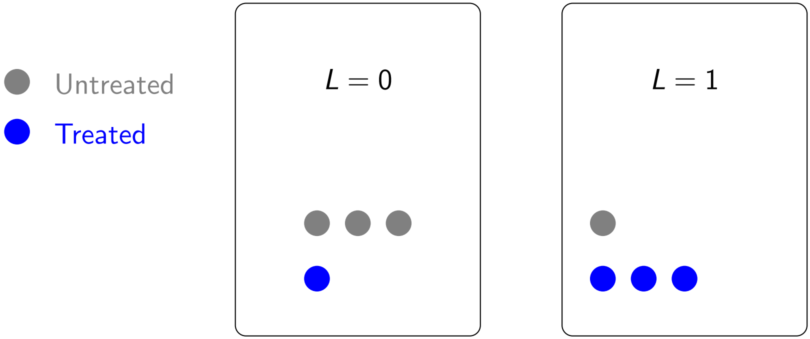
Inverse probability of treatment weighting



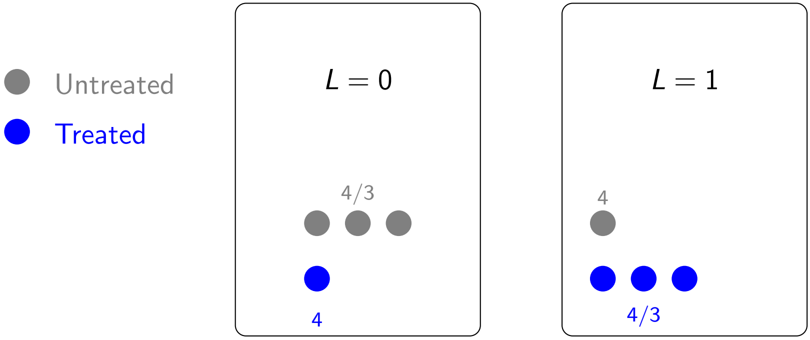
Inverse probability of treatment weighting



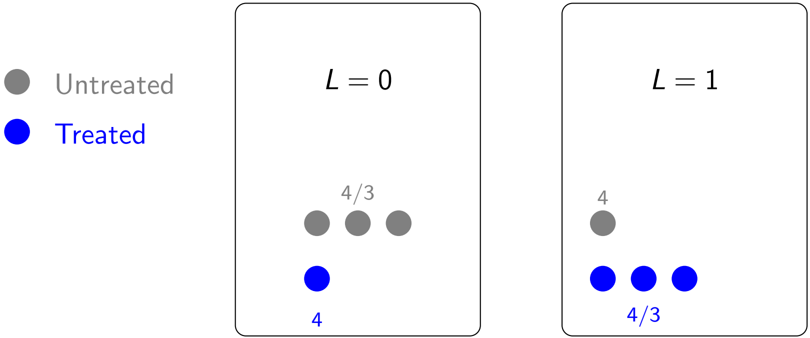
Inverse probability of treatment weighting



Inverse probability of treatment weighting



Inverse probability of treatment weighting



$\pi_i = 1/4$

$\frac{1}{\pi_i} = 4, \frac{1}{1-\pi_i} = 4/3$

$\pi_i = 3/4$

$\frac{1}{\pi_i} = 4/3, \frac{1}{1-\pi_i} = 4$

Propensity score: $\pi_i = \mathrm{P}(A = 1 \mid L = L_i)$

Inverse probability of treatment weighting

$$\begin{aligned}\hat{\mathrm{E}}(Y^1) &= \frac{1}{N}\sum_{i:A_i=1}\frac{Y_i}{\hat{\pi}_i} \\ &= \frac{1}{N}\left(\sum_{i:A_i=1}\frac{A_iY_i}{\hat{\pi}_i} + \sum_{i:A_i=0}\frac{A_iY_i}{\hat{\pi}_i}\right) \quad = \frac{1}{N}\sum_{i}\frac{A_iY_i}{\hat{\pi}_i}\end{aligned} \tag{1}$$

Inverse probability of treatment weighting

$$\hat{\mathrm{E}}(Y^1) = \frac{1}{N}\sum_{i:A_i=1}\frac{Y_i}{\hat{\pi}_i} = \frac{1}{N}\left(\sum_{i:A_i=1}\frac{A_iY_i}{\hat{\pi}_i} + \sum_{i:A_i=0}\frac{A_iY_i}{\hat{\pi}_i}\right) = \frac{1}{N}\sum_i\frac{A_iY_i}{\hat{\pi}_i} \tag{1}$$

$$\hat{\mathrm{E}}(Y^0) = \frac{1}{N}\sum_{i:A_i=0}\frac{Y_i}{1-\hat{\pi}_i} = \frac{1}{N}\left(\sum_{i:A_i=1}\frac{(1-A_i)Y_i}{1-\hat{\pi}_i} + \sum_{i:A_i=0}\frac{(1-A_i)Y_i}{1-\hat{\pi}_i}\right) = \frac{1}{N}\sum_i\frac{(1-A_i)Y_i}{1-\hat{\pi}_i} \tag{2}$$

Parametric model: propensity model

Model the treatment assignment

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

Estimate by inverse probability weighting (IPW)

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

Outcome modeling vs Propensity score modeling

- If our model captures the true relationship, either will work
- Outcome modeling is used more because it typically has lower variance

Outcome modeling vs Propensity score modeling

- If our model captures the true relationship, either will work
- Outcome modeling is used more because it typically has lower variance
- What if our models are wrong?

Outcome modeling vs Propensity score modeling

- If our model captures the true relationship, either will work
- Outcome modeling is used more because it typically has lower variance
- What if our models are wrong?
- We can use flexible machine learning methods with low bias when sample size is very large
- What if we don't include the right covariates?

Augmented IPW

We can use both outcome modeling and IPW together

$$\hat{E}(Y^1) = \frac{1}{N}\left(\sum_i \frac{A_i Y_i}{\hat{\pi}_i} - \frac{(A_i - \hat{\pi}_i)\hat{E}(Y \mid A = 1, L = \ell_i)}{\hat{\pi}_i}\right) \quad (3)$$

$$\hat{E}(Y^0) = \frac{1}{N}\left(\sum_i \frac{(1 - A_i)Y_i}{1 - \hat{\pi}_i} - \frac{([1 - A_i] - [1 - \hat{\pi}_i])\hat{E}(Y \mid A = 0, L = \ell_i)}{1 - \hat{\pi}_i}\right) \quad (4)$$

Augmented IPW

Why is this a good idea?

$$
\begin{aligned}
&\mathrm{E}\left(\frac{A_i Y_i}{\hat{\pi}_i} - \frac{(A_i - \hat{\pi}_i)\hat{E}(Y \mid A = 1, L = \ell_i)}{\hat{\pi}_i}\right) \\
&= \mathrm{E}\left(Y_i^1 - Y_i^1 \frac{\hat{\pi}_i}{\hat{\pi}_i} + \frac{A_i Y_i}{\hat{\pi}_i} - \frac{(A_i - \hat{\pi}_i)\hat{E}(Y \mid A = 1, L = \ell_i)}{\hat{\pi}_i}\right) \\
&= \mathrm{E}\left(Y_i^1 + \frac{(A_i - \pi_i)Y_i^1}{\hat{\pi}_i} - \frac{(A_i - \hat{\pi}_i)\hat{E}(Y \mid A = 1, L = \ell_i)}{\hat{\pi}_i}\right) \\
&= \mathrm{E}\left(Y_i^1\right) + \mathrm{E}\left(\frac{(A_i - \hat{\pi}_i)}{\hat{\pi}_i}\left[Y_i^1 - \hat{E}(Y \mid A = 1, L = \ell_i)\right]\right) \\
&= \mathrm{E}\left(Y_i^1\right) \\
&\quad + \mathrm{E}\left[\mathrm{E}\left(\frac{(A_i - \hat{\pi}_i)}{\hat{\pi}_i} \mid L = \ell_i\right)\mathrm{E}\left(Y_i^1 - \hat{\mathrm{E}}[Y \mid A = 1, L = \ell_i] \mid L = \ell_i\right)\right]
\end{aligned} \tag{5}
$$

Augmented IPW

Why is this a good idea?

$$
\begin{aligned}
\mathrm{E}(\hat{\mathrm{E}}(Y_i^1)) &= \mathrm{E}\left(Y_i^1\right) \\
&\quad + \mathrm{E}\left[\mathrm{E}\left(\frac{(A_i - \hat{\pi}_i)}{\hat{\pi}_i} \mid L = \ell_i\right)\mathrm{E}\left(Y_i^1 - \hat{\mathrm{E}}[Y \mid A = 1, L = \ell_i] \mid L = \ell_i\right)\right]
\end{aligned} \tag{6}
$$

Augmented IPW

Why is this a good idea?

$$\mathrm{E}(\hat{\mathrm{E}}(Y_i^1)) = \mathrm{E}\left(Y_i^1\right) + \mathrm{E}\left[\mathrm{E}\left(\frac{(A_i - \hat{\pi}_i)}{\hat{\pi}_i} \mid L = \ell_i\right)\mathrm{E}\left(Y_i^1 - \hat{\mathrm{E}}[Y \mid A = 1, L = \ell_i] \mid L = \ell_i\right)\right] \tag{6}$$

$$\mathrm{E}\left(\frac{(A_i - \hat{\pi}_i)}{\hat{\pi}_i} \mid L = \ell_i\right) = \frac{\pi_i - \hat{\pi}_i}{\hat{\pi}_i} \tag{7}$$

has expectation zero when $\hat{\pi}_i$ is correctly specified and non-zero

Augmented IPW

Why is this a good idea?

$$\mathrm{E}(\hat{\mathrm{E}}(Y_i^1)) = \mathrm{E}\left(Y_i^1\right) + \mathrm{E}\left[\mathrm{E}\left(\frac{(A_i - \hat{\pi}_i)}{\hat{\pi}_i} \mid L = \ell_i\right) \mathrm{E}\left(Y_i^1 - \hat{\mathrm{E}}[Y \mid A = 1, L = \ell_i] \mid L = \ell_i\right)\right] \tag{6}$$

$$\mathrm{E}\left(Y_i^1 - \hat{\mathrm{E}}[Y \mid A = 1, L = \ell_i] \mid L = \ell_i\right) \tag{7}$$

has expectation zero when the outcome model is correctly specified and non-zero

Augmented IPW

- Estimator of ATE is "doubly robust"
  - Second term has expectation 0 if
    - propensity score model is well specified, or
    - the outcome model is well specified
  - Robust against misspecification of either (but not both)
- If the outcome model is well specified, using standardization with just the outcome model often has less variance
- If the outcome model is not well specified, using standardization with just the outcome model will not be consistent
- 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 parametric model for the outcome $E(Y \mid A, L)$ and treatment
- Reason about the bias variance tradeoff
- Use the Augmented IPW estimator to guard against model misspecification

After class:

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