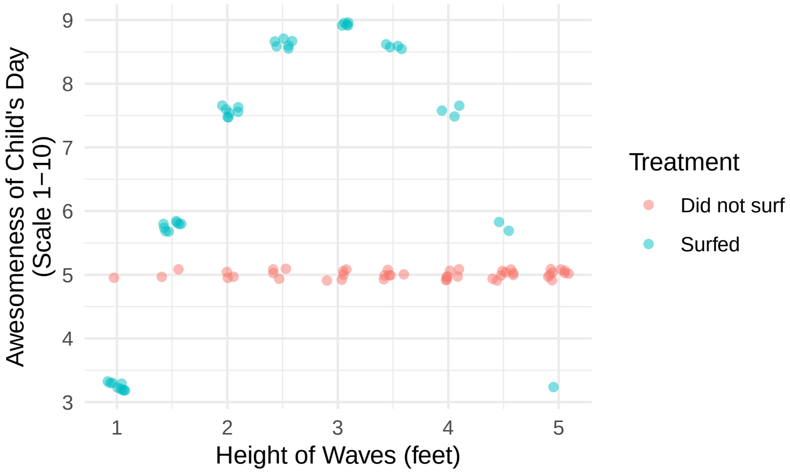
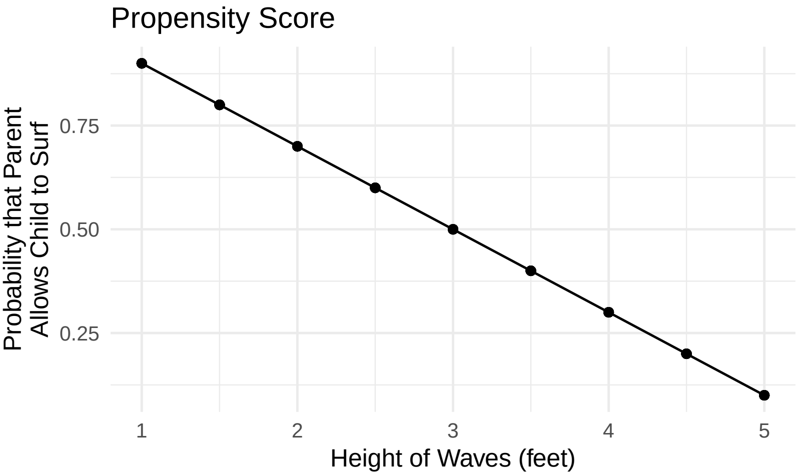
Doubly-Robust Estimation[1]

Ian Lundberg
Soc 212b
ilundberg.github.io/soc212b

Winter 2025

[1]Especially today, slides are a high-level overview and we will rely on the website for some technical things.

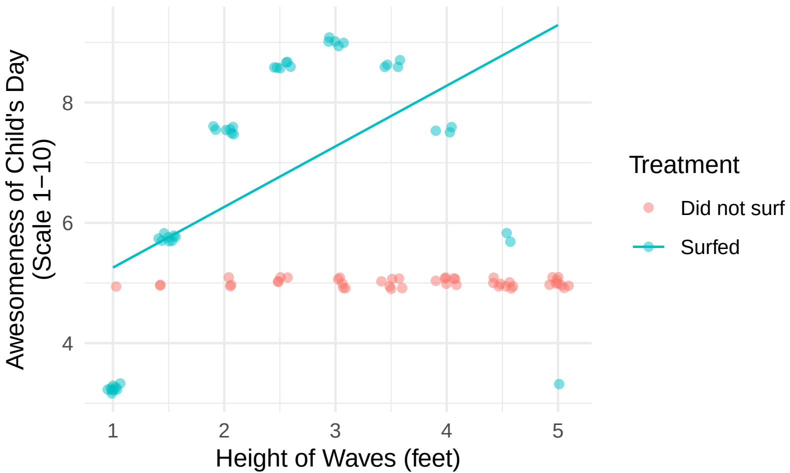

Awesomeness of Child's Day
(Scale 1–10)
9
8
7
6
5
4
3
1
2
3
4
5
Height of Waves (feet)
Treatment
Did not surf
Surfed

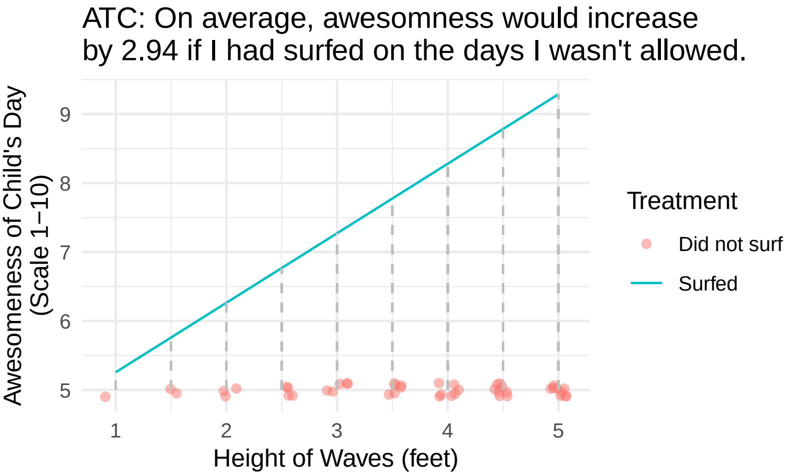


Propensity Score
Probability that Parent Allows Child to Surf
0.75
0.50
0.25
1
2
3
4
5
Height of Waves (feet)

Child:
How much more awesome would my day have been if I had surfed on the days when my parents didn't let me?

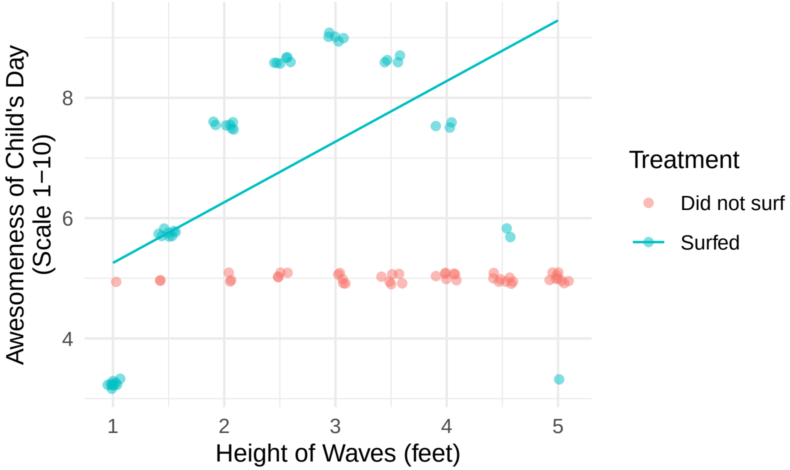
$$\mathrm{ATC} = \frac{1}{n_0} \sum_{i:A_i=0} \left(Y_i^1 - Y_i^0\right)$$

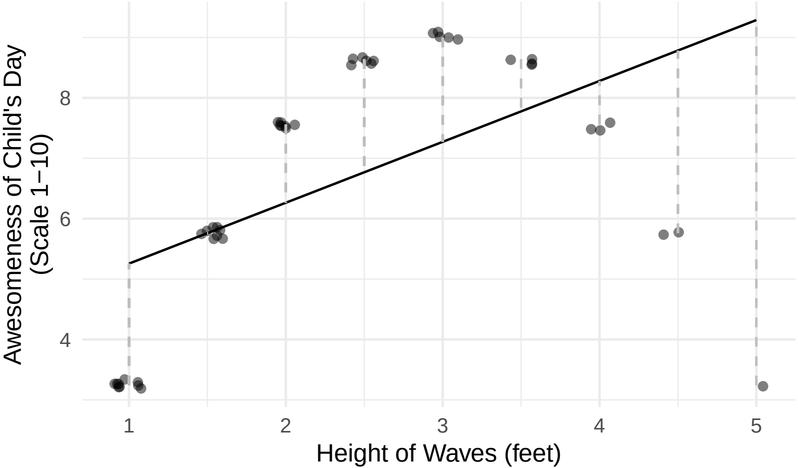

Awesomeness of Child's Day
(Scale 1–10)
8
6
4
1
2
3
4
5
Height of Waves (feet)
Treatment
Did not surf
Surfed


ATC: On average, awesomness would increase
by 2.94 if I had surfed on the days I wasn't allowed.
Awesomeness of Child's Day
(Scale 1–10)
9
8
7
6
5
1
2
3
4
5
Height of Waves (feet)
Treatment
Did not surf
Surfed

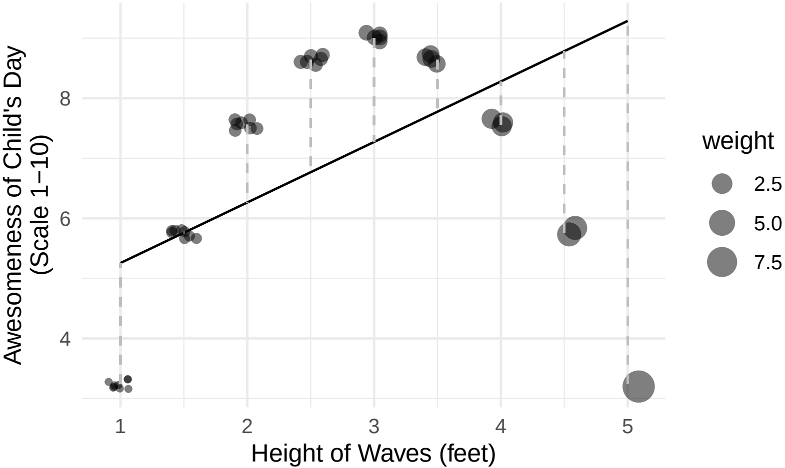
To discuss:

- In what sense is this line best-fit to the wrong goal?
- How important is the error at each x-value?



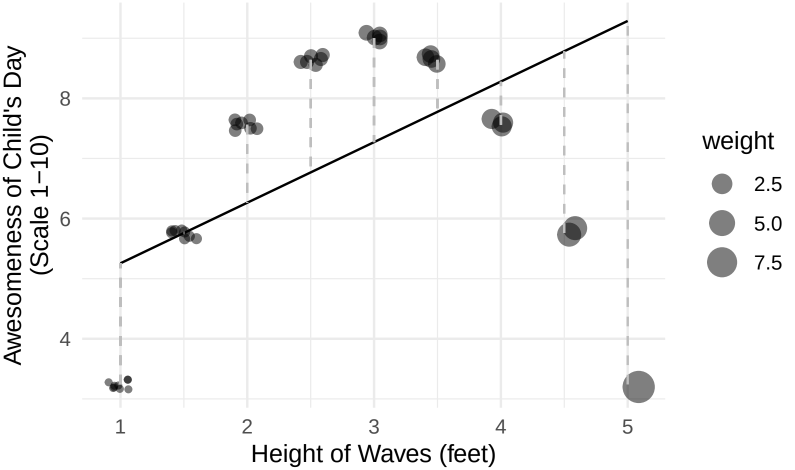


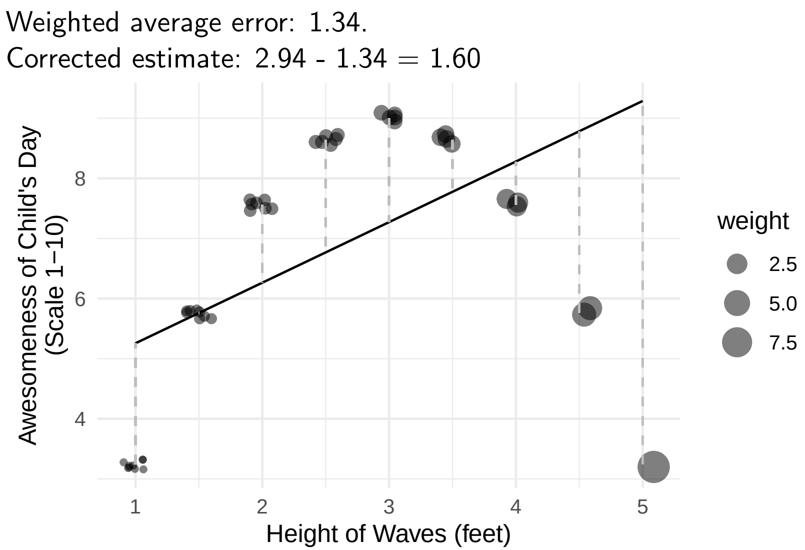
Awesomeness of Child's Day
(Scale 1–10)
8
6
4
1
2
3
4
5
Height of Waves (feet)



Awesomeness of Child's Day
(Scale 1–10)
4
6
8
1
2
3
4
5
Height of Waves (feet)
weight
2.5
5.0
7.5

Weighted average error: 1.34.





Weighted average error: 1.34.
Corrected estimate: 2.94 - 1.34 = 1.60
Awesomeness of Child's Day
(Scale 1–10)
4
6
8
1
2
3
4
5
Height of Waves (feet)
weight
2.5
5.0
7.5

Doubly-robust estimation: Summary

For the ATC:

- Predict $\hat{Y}^1$
- Among treated cases,
  - Weight by $\frac{\hat{P}(A=1)}{\hat{\mathrm{P}}(A=0)}$
  - Take weighted average error: $\hat{Y}^1 - Y$
  - This is a bias correction:
    model was fit at x-values of treated cases,
    target to predict is x-values of untreated cases
- Among untreated cases, take average $\hat{Y}^1$
- Then subtract the bias correction

Three estimators of $\hat{\mathrm{E}}(Y^a)$

What is right when $\hat{g}(a, \vec{x}) \to \mathrm{E}(Y \mid A = a, \vec{X} = \vec{x})$?

What is right when $\hat{m}(a, \vec{x}) \to \mathrm{P}(A = a \mid \vec{X} = \vec{x})$?

$$\hat{\tau}_{\text{Outcome}}(a) = \frac{1}{n} \sum_i \hat{g}(a, \vec{X}_i)$$

$$\hat{\tau}_{\text{Treatment}}(a) = \frac{1}{\sum_{i:A_i=a} \frac{1}{\hat{m}(A_i, \vec{X}_i)}} \sum_{i:A_i=a} \frac{Y_i}{\hat{m}(A_i, \vec{X}_i)}$$

$$\hat{\tau}_{\text{AIPW}}(a) = \frac{1}{n} \sum_i \hat{g}(a, \vec{X}_i) - \frac{1}{\sum_{i:A_i=a} \frac{1}{\hat{m}(A_i, \vec{X}_i)}} \sum_{i:A_i=a} \frac{\hat{g}(A_i, X_i) - Y_i}{\hat{m}(A_i, \vec{X}_i)}$$

Double robustness: When is each estimator correct?

With $\hat{g}$ as the outcome model and $\hat{m}$ as the treatment model:

	$\hat{\tau}_{\text{Outcome}}(a)$	$\hat{\tau}_{\text{Treatment}}(a)$	$\hat{\tau}_{\text{AIPW}}(a)$
when $\hat{g}$ and $\hat{m}$ are correct			

Double robustness: When is each estimator correct?

With $\hat{g}$ as the outcome model and $\hat{m}$ as the treatment model:

	$\hat{\tau}_{\text{Outcome}}(a)$	$\hat{\tau}_{\text{Treatment}}(a)$	$\hat{\tau}_{\text{AIPW}}(a)$
when $\hat{g}$ and $\hat{m}$ are correct	✓		

Double robustness: When is each estimator correct?

With $\hat{g}$ as the outcome model and $\hat{m}$ as the treatment model:

	$\hat{\tau}_{\text{Outcome}}(a)$	$\hat{\tau}_{\text{Treatment}}(a)$	$\hat{\tau}_{\text{AIPW}}(a)$
when $\hat{g}$ and $\hat{m}$ are correct	✓	✓	

Double robustness: When is each estimator correct?

With $\hat{g}$ as the outcome model and $\hat{m}$ as the treatment model:

	$\hat{\tau}_{\text{Outcome}}(a)$	$\hat{\tau}_{\text{Treatment}}(a)$	$\hat{\tau}_{\text{AIPW}}(a)$
when $\hat{g}$ and $\hat{m}$ are correct	✓	✓	✓

Double robustness: When is each estimator correct?

With $\hat{g}$ as the outcome model and $\hat{m}$ as the treatment model:

	$\hat{\tau}_{\text{Outcome}}(a)$	$\hat{\tau}_{\text{Treatment}}(a)$	$\hat{\tau}_{\text{AIPW}}(a)$
when $\hat{g}$ and $\hat{m}$ are correct	✓	✓	✓
when only $\hat{g}$ is correct			

Double robustness: When is each estimator correct?

With $\hat{g}$ as the outcome model and $\hat{m}$ as the treatment model:

	$\hat{\tau}_{\text{Outcome}}(a)$	$\hat{\tau}_{\text{Treatment}}(a)$	$\hat{\tau}_{\text{AIPW}}(a)$
when $\hat{g}$ and $\hat{m}$ are correct	✓	✓	✓
when only $\hat{g}$ is correct	✓		

Double robustness: When is each estimator correct?

With $\hat{g}$ as the outcome model and $\hat{m}$ as the treatment model:

	$\hat{\tau}_{\text{Outcome}}(a)$	$\hat{\tau}_{\text{Treatment}}(a)$	$\hat{\tau}_{\text{AIPW}}(a)$
when $\hat{g}$ and $\hat{m}$ are correct	✓	✓	✓
when only $\hat{g}$ is correct	✓	×	

Double robustness: When is each estimator correct?

With $\hat{g}$ as the outcome model and $\hat{m}$ as the treatment model:

	$\hat{\tau}_{\text{Outcome}}(a)$	$\hat{\tau}_{\text{Treatment}}(a)$	$\hat{\tau}_{\text{AIPW}}(a)$
when $\hat{g}$ and $\hat{m}$ are correct	✓	✓	✓
when only $\hat{g}$ is correct	✓	$\times$	✓

Double robustness: When is each estimator correct?

With $\hat{g}$ as the outcome model and $\hat{m}$ as the treatment model:

	$\hat{\tau}_{\text{Outcome}}(a)$	$\hat{\tau}_{\text{Treatment}}(a)$	$\hat{\tau}_{\text{AIPW}}(a)$
when $\hat{g}$ and $\hat{m}$ are correct	✓	✓	✓
when only $\hat{g}$ is correct	✓	$\times$	✓
when only $\hat{m}$ is correct			

Double robustness: When is each estimator correct?

With $\hat{g}$ as the outcome model and $\hat{m}$ as the treatment model:

	$\hat{\tau}_{\text{Outcome}}(a)$	$\hat{\tau}_{\text{Treatment}}(a)$	$\hat{\tau}_{\text{AIPW}}(a)$
when $\hat{g}$ and $\hat{m}$ are correct	✓	✓	✓
when only $\hat{g}$ is correct	✓	$\times$	✓
when only $\hat{m}$ is correct	$\times$		

Double robustness: When is each estimator correct?

With $\hat{g}$ as the outcome model and $\hat{m}$ as the treatment model:

	$\hat{\tau}_{\text{Outcome}}(a)$	$\hat{\tau}_{\text{Treatment}}(a)$	$\hat{\tau}_{\text{AIPW}}(a)$
when $\hat{g}$ and $\hat{m}$ are correct	✓	✓	✓
when only $\hat{g}$ is correct	✓	$\times$	✓
when only $\hat{m}$ is correct	$\times$	✓	

Double robustness: When is each estimator correct?

With $\hat{g}$ as the outcome model and $\hat{m}$ as the treatment model:

	$\hat{\tau}_{\text{Outcome}}(a)$	$\hat{\tau}_{\text{Treatment}}(a)$	$\hat{\tau}_{\text{AIPW}}(a)$
when $\hat{g}$ and $\hat{m}$ are correct	✓	✓	✓
when only $\hat{g}$ is correct	✓	$\times$	✓
when only $\hat{m}$ is correct	$\times$	✓	✓

The problem of overfitting

Suppose $\hat{g}$ is very complicated

- e.g. regress Y on $p = 100$ predictors in a sample of $n = 150$

Debiasing relies on errors: $\hat{g}(A, \vec{X}) - Y$

- What is wrong with these errors?
- How to fix it?

Sample splitting for AIPW

Sample splitting for AIPW

1. Split data into sample $\mathcal{S}_1$ and $\mathcal{S}_2$

Sample splitting for AIPW

1. Split data into sample $\mathcal{S}_1$ and $\mathcal{S}_2$
2. Using $\mathcal{S}_1$, estimate $\hat{g}$ and $\hat{m}$

Sample splitting for AIPW

1. Split data into sample $\mathcal{S}_1$ and $\mathcal{S}_2$
2. Using $\mathcal{S}_1$, estimate $\hat{g}$ and $\hat{m}$
3. Using $\mathcal{S}_2$, calculate the AIPW estimator
   - so that errors are on out-of-sample cases

Sample splitting for AIPW

1. Split data into sample $\mathcal{S}_1$ and $\mathcal{S}_2$
2. Using $\mathcal{S}_1$, estimate $\hat{g}$ and $\hat{m}$
3. Using $\mathcal{S}_2$, calculate the AIPW estimator
   - so that errors are on out-of-sample cases

Popularized as double machine learning (Chernozhukov et al. 2018)

Sample splitting for AIPW

1. Split data into sample $\mathcal{S}_1$ and $\mathcal{S}_2$
2. Using $\mathcal{S}_1$, estimate $\hat{g}$ and $\hat{m}$
3. Using $\mathcal{S}_2$, calculate the AIPW estimator
    - so that errors are on out-of-sample cases

Popularized as double machine learning (Chernozhukov et al. 2018)

Concern: Loss of sample size due to splitting.

Sample splitting for AIPW

1. Split data into sample $\mathcal{S}_1$ and $\mathcal{S}_2$
2. Using $\mathcal{S}_1$, estimate $\hat{g}$ and $\hat{m}$
3. Using $\mathcal{S}_2$, calculate the AIPW estimator
   - so that errors are on out-of-sample cases

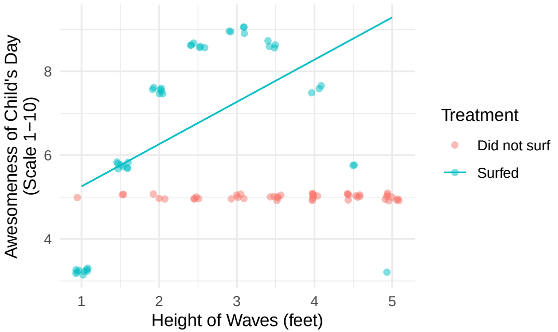
Popularized as double machine learning (Chernozhukov et al. 2018)

Concern: Loss of sample size due to splitting.
Answer: Cross fitting. Swap $\mathcal{S}_1$ and $\mathcal{S}_2$. Average result.

Targeted learning

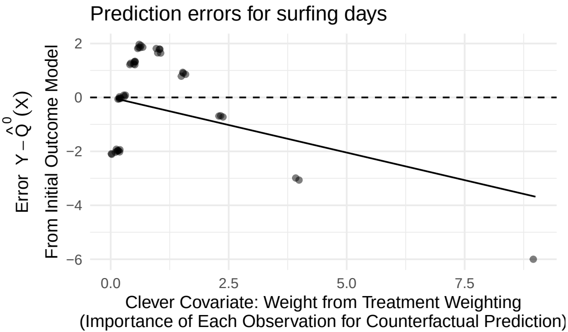
Initial outcome model

$$\underbrace{\hat{Q}^0(\vec{x})}_{\text{The 0 superscript indicates an untargeted initial estimate}} = \hat{\mathrm{E}}(Y \mid A = 1, \vec{X}) = \hat{\alpha} + \hat{\beta}\vec{x}$$



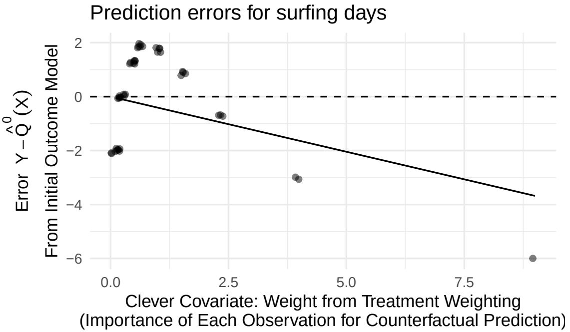
Clever covariate

$$H(x) = \frac{\text{P}(A = \text{Not Surfed} \mid X = x)}{\text{P}(A = \text{Surfed} \mid X = x)}$$



Targeted outcome model

$$\hat{Q}^1(x) = \hat{Q}^0(x) + \hat{\gamma} \underbrace{\left(\frac{\text{P}(A = \text{Not surfed} \mid X = x)}{\text{P}(A = \text{Surfed} \mid X = x)} \right)}_{\text{Clever covariate } h(x)}$$



Initial and targeted estimates

$$\text{Estimand:} \quad \tau = \mathrm{E}(Y^{\text{Surfed}} - Y^{\text{Not Surfed}} \mid A = \text{Not Surfed})$$

$$\text{Initial estimate:} \quad \hat{\tau}^0 = \frac{1}{n_{\text{NotSurfed}}} \sum_{i:A_i=\text{NotSurfed}} \left(\hat{Q}^0(x_i) - y_i \right)$$

$$\text{Targeted estimate:} \quad \hat{\tau}^1 = \frac{1}{n_{\text{NotSurfed}}} \sum_{i:A_i=\text{NotSurfed}} \left(\hat{Q}^1(x_i) - y_i \right)$$

Why targeted learning?

Why targeted learning?

- Doubly robust
- Intuition: Targeting the outcome model
- Generalizes to GLM outcome models