10. The generality of the g-formula: Using any estimator

Ian Lundberg

Cornell Info 6751: Causal Inference in Observational Settings

Fall 2022

22 Sep 2022

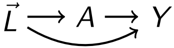
Learning goals for today

At the end of class, you will be able to:

1. Use machine learning methods to estimate causal effects
2. Select an estimator using predictive performance

You are now well-versed in the g-formula

1. Assume a DAG



2. By consistency, exchangeability, and positivity,

$$\underbrace{\mathrm{E}(Y^a \mid \vec{L} = \vec{\ell})}_{\text{Causal}} = \underbrace{\mathrm{E}(Y \mid A = a, \vec{L} = \vec{\ell})}_{\text{Statistical}}$$

3. Using regression, estimate $\hat{\mathrm{E}}(Y \mid A, \vec{L})$
4. Predict unknown potential outcomes and average

$$\hat{\mathrm{E}}(Y^a) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathrm{E}}\left(Y \mid A = a, \vec{L} = \vec{\ell}_i\right)$$

Big idea: Why constrain ourselves to regression for $\hat{\mathrm{E}}(Y \mid A, \vec{L})$?

Hill, Jennifer L. 2011.
"Bayesian nonparametric modeling for causal inference."
Journal of Computational and Graphical Statistics 20.1:217-240.

- Binary treatment (simulated)
- Continuous confounder X (simulated)

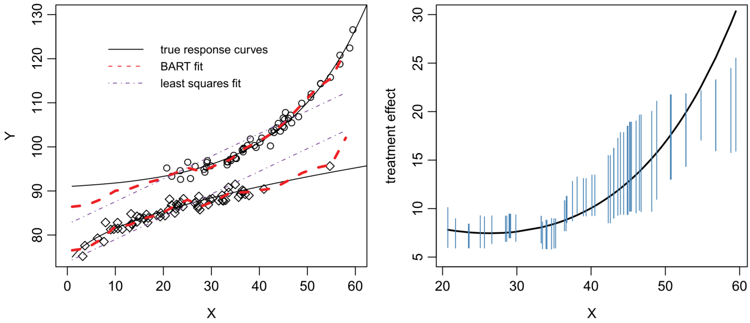


Figure 1. Left panel: simulated data with linear regression and BART fits. Right panel: BART inference for treatment effect on the treated. A color version of this figure is available in the electronic version of this article.

How did she do that?[1]

1) Learn an automated partitioning of the data (aka a "tree")

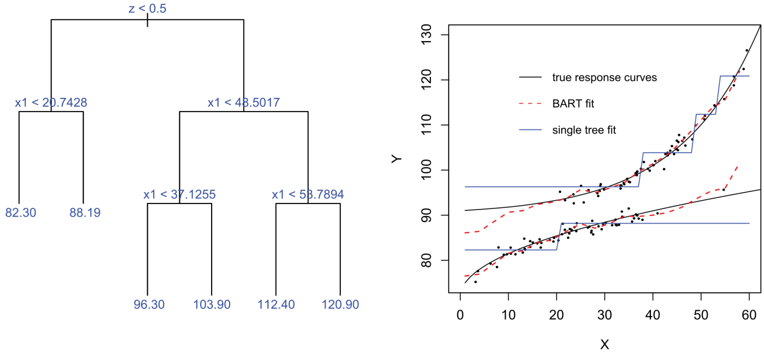


Figure 2. Left panel: the binary tree fit to the data from Figure 1. Right panel: single-tree fits (solid lines) and BART fits (dashed lines). A color version of this figure is available in the electronic version of this article.

[1] Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. "BART: Bayesian additive regression trees." The Annals of Applied Statistics 4.1 (2010): 266-298.

How did she do that?

2) Repeat many times. Take the average.

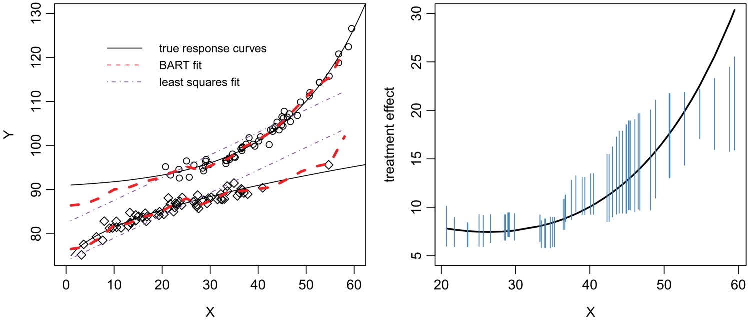


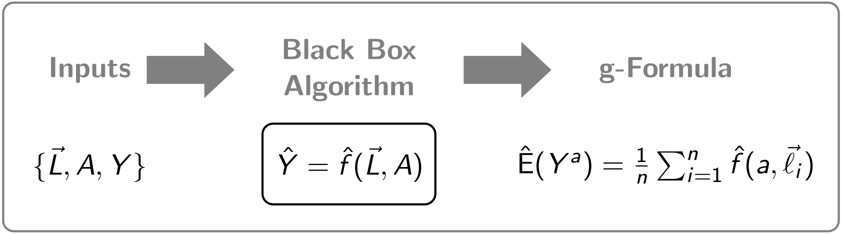
Figure 1. Left panel: simulated data with linear regression and BART fits. Right panel: BART inference for treatment effect on the treated. A color version of this figure is available in the electronic version of this article.

Core idea: Causal assumptions unlock machine learning[2]

Once you make this assumption

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

you get to do this



Inputs

Black Box Algorithm

g-Formula

$\{\vec{L}, A, Y\}$

$\boxed{\hat{Y} = \hat{f}(\vec{L}, A)}$

$\hat{\mathrm{E}}(Y^a) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(a, \vec{\ell}_i)$

[2] Caveat: There are ways to do even better. This is just a start.
See Van der Laan, M. J., & Rose, S. (2018). Targeted learning in data science. Springer International Publishing.

There are **so many** algorithms you might use!

Statistical Science
2019, Vol. 34, No. 1, 43–68
https://doi.org/10.1214/18-STS667

Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition[1]

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott and Dan Cervone

Abstract. Statisticians have made great progress in creating methods that reduce our reliance on parametric assumptions. However, this explosion in research has resulted in a breadth of inferential strategies that both create opportunities for more reliable inference as well as complicate the choices that an applied researcher has to make and defend. Relatedly, researchers advocating for new methods typically compare their method to at best 2 or 3 other causal inference strategies and test using simulations that may or may not be designed to equally tease out flaws in all the competing methods. The causal inference data analysis challenge, "Is Your SATT Where It's At?", launched as part of the 2016 Atlantic Causal Inference Conference, sought to make progress with respect to both of these issues. The researchers creating the data testing grounds were distinct from the researchers submitting methods whose efficacy would be evaluated. Results from 30 competitors across the two versions of the competition (black-box algorithms and do-it-yourself analyses) are presented along with post-hoc analyses that reveal information about the characteristics of causal inference strategies and settings that affect performance. The most consistent conclusion was that methods that flexibly model the response surface perform better overall than methods that fail to do so. Finally new methods are proposed that combine features of several of the top-performing submitted methods.

Key words and phrases: Causal inference, competition, machine learning, automated algorithms, evaluation.

1. INTRODUCTION

In the absence of a controlled randomized or natural experiment,[2] inferring causal effects involves the difficult task of constructing fair comparisons between ob-

Vincent Dorie is Associate Research Scientist, Data Science Institute, Columbia University, 475 Riverside Drive, Room 320L, New York, New York 10115, USA (e-mail: vdorie@gmail.com). Jennifer Hill is Professor of Applied Statistics and Data Science, Department of Applied Statistics, Social Science, and Humanities, New York University, 246 Greene Street, 3rd Floor, New York, New York 10003, USA (e-mail: jennifer.hill@nyu.edu). Uri Shalit is Assistant Professor, Faculty of Industrial Engineering and Management, Technion, Technion—Israel Institute of Technology, Technion City, Haifa 3200003, Israel (e-mail: urishalit@technion.ac.il). Marc Scott is Professor of Applied Statistics, Department of Applied Statistics, Social Science, and Humanities, New York University, 246 Greene Street, 3rd Floor, New York, New York 10003, USA (e-mail: marc.scott@nyu.edu). Dan Cervone is Director of Quantitative Research, Los Angeles Dodgers, Dodger Stadium, 1000 Vin Scully Ave., Los Angeles, California 90012, USA (e-mail: dcervone@gmail.com).

[1] Discussed in 10.1214/18-STS684; 10.1214/18-STS680; 10.1214/18-STS690; 10.1214/18-STS689; 10.1214/18-STS679; 10.1214/18-STS682; 10.1214/18-STS688

[2] We use natural experiment to include (1) studies where the causal variable is randomized not for the purposes of a study (for instance, a school lottery), (2) studies where a variable is randomized but the causal variable of interest is downstream of this (e.g., plays the role of an instrumental variable), and (3) regression discontinuity designs.

Dorie et al. 2019[3]: Is Your SATT Where It's At?

[3] Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition." Statistical Science, 34(1), 43-68. See also `https://jenniferhill7.wixsite.com/acic-2016/competition`

Dorie et al. 2019[3]: Is Your SATT Where It's At?

- Goal: The Sample Average Treatment Effect on the Treated

$$\text{SATT} = \frac{1}{n_{\text{Treated}}} \sum_{i:A_i=1} \left(Y_i^1 - Y_i^0\right)$$

[3] Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition." Statistical Science, 34(1), 43-68. See also `https://jenniferhill7.wixsite.com/acic-2016/competition`

Dorie et al. 2019[3]: Is Your SATT Where It's At?

- Goal: The Sample Average Treatment Effect on the Treated

$$\text{SATT} = \frac{1}{n_{\text{Treated}}} \sum_{i:A_i=1} \left(Y_i^1 - Y_i^0\right)$$

- Simulated data. SATT was known to organizers

[3] Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition." Statistical Science, 34(1), 43-68. See also `https://jenniferhill7.wixsite.com/acic-2016/competition`

Dorie et al. 2019[3]: Is Your SATT Where It's At?

- Goal: The Sample Average Treatment Effect on the Treated

$$\text{SATT} = \frac{1}{n_{\text{Treated}}} \sum_{i:A_i=1} \left(Y_i^1 - Y_i^0\right)$$

- Simulated data. SATT was known to organizers
- Confounders were defined by the organizers

[3] Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition." Statistical Science, 34(1), 43-68. See also https://jenniferhill7.wixsite.com/acic-2016/competition

Dorie et al. 2019[3]: Is Your SATT Where It's At?

- Goal: The Sample Average Treatment Effect on the Treated

$$\text{SATT} = \frac{1}{n_{\text{Treated}}} \sum_{i:A_i=1} \left(Y_i^1 - Y_i^0\right)$$

- Simulated data. SATT was known to organizers
- Confounders were defined by the organizers
- Participants could use any algorithm to estimate SATT

[3] Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition." Statistical Science, 34(1), 43-68. See also https://jenniferhill7.wixsite.com/acic-2016/competition

Dorie et al. 2019[3]: Is Your SATT Where It's At?

- Goal: The Sample Average Treatment Effect on the Treated

$$\text{SATT} = \frac{1}{n_{\text{Treated}}} \sum_{i:A_i=1} \left(Y_i^1 - Y_i^0\right)$$

- Simulated data. SATT was known to organizers
- Confounders were defined by the organizers
- Participants could use any algorithm to estimate SATT
- 30 teams attempted the task

[3] Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition." Statistical Science, 34(1), 43-68. See also `https://jenniferhill7.wixsite.com/acic-2016/competition`

Dorie et al. 2019[3]: Is Your SATT Where It's At?

- Goal: The Sample Average Treatment Effect on the Treated

$$\text{SATT} = \frac{1}{n_{\text{Treated}}} \sum_{i:A_i=1} \left(Y_i^1 - Y_i^0\right)$$

- Simulated data. SATT was known to organizers
- Confounders were defined by the organizers
- Participants could use any algorithm to estimate SATT
- 30 teams attempted the task
- Today you will attempt it!

[3] Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition." Statistical Science, 34(1), 43-68. See also https://jenniferhill7.wixsite.com/acic-2016/competition

A few things to help you succeed

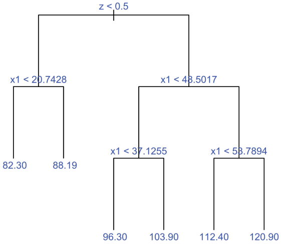
1. A few algorithms you might consider
   - Ridge, LASSO, elastic net (`glmnet`)
   - Random forest (`ranger`)
   - Bayesian additive regression trees (`BART`)
   - Super Learner (`SuperLearner`)
2. How do I choose a black-box algorithm?
3. Overview of the data structure

A few algorithms you might consider: `glmnet`

- Idea: With many coefficients, OLS can by high-variance.
- `glmnet` **penalizes** coefficients to reduce sample variance
  - Pushes coefficeints toward 0
  - When we are uncertain about $\hat{\beta}$, better to keep $\hat{\beta}$ small
- Three ways to penalize
  - Ridge penalty: Minimize $\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 + \lambda\sum_j \beta_j^2$
    - Coefficients are pulled toward 0, but not exactly to 0
  - Lasso penalty: Minimize $\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 + \lambda\sum_j |\beta_j|$
    - Some coefficients pushed exactly to 0 (dropped out entirely)
  - Elastic net: Penalize both β_j^2 and $|\beta_j|$

A few algorithms you might consider: `ranger`[4]

- Random forest: A frequentist sum-of-trees model



(tree visualization from Hill 2011)

- Good at learning interactions among covariates
- Ranger is fast

[4] Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software, 77(i01).

A few algorithms you might consider: `BART`[5]

- Bayesian version of random forest
  - A prior regularizes estimates
  - Bonus: Free posterior variance estimates!
- Warning: A bit slower than ranger

[5] Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. "BART: Bayesian additive regression trees." The Annals of Applied Statistics 4.1 (2010): 266-298.

A few algorithms you might consider: `SuperLearner`[6]

Why pick just one algorithm?

1. Fit many candidate learners $f_1(), f_2(), \dots$
2. Predict out-of-sample (using cross-validation)
3. Learn a set of weights to take a weighted average

$$\hat{f}(\vec{\ell}, a) = \hat{\beta}_1 \underbrace{\hat{f}_1(\vec{\ell}, a)}_{\text{e.g. OLS}} + \beta_2 \underbrace{\hat{f}_2(\vec{\ell}, a)}_{\text{e.g. glmnet}} + \beta_3 \underbrace{\hat{f}_3(\vec{\ell}, a)}_{\text{e.g. ranger}}$$

[6] Original: Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. Statistical Applications in Genetics and Molecular Biology, 6(1). Good intro paper: Naimi, A. I., & Balzer, L. B. (2018). Stacked generalization: an introduction to super learning. European Journal of Epidemiology, 33(5), 459-464.

A few algorithms you might consider

- You could use any of these
- You could use something else
- You could do the entire exercise with OLS
  - try various functional forms
- Choice is yours!

How do I choose the black-box algorithm?[7]

Could choose by an empirical metric of **predictive performance**

Task:	Predict Y given $\{A, \vec{L}\}$
Performance metric:	Minimize out-of-sample mean squared error (MSE)

Why MSE?

- The lowest-MSE predictor would be the true mean $\text{E}(Y \mid \vec{L}, A)$
- Therefore, MSE is a principled metric when selecting an approximation

[7] One can improve on this metric; we will come back to this later in the semester. See Van der Laan, M. J., & Rose, S. (2018). Targeted learning in data science. Springer International Publishing.

Selecting an algorithm: The role of a train-test split

Case 1	$\{\vec{L}_1, A_1\}$	Y_1
Case 2	$\{\vec{L}_2, A_2\}$	Y_2
Case 3	$\{\vec{L}_3, A_3\}$	Y_3
Case 4	$\{\vec{L}_4, A_4\}$	Y_4
Case 5	$\{\vec{L}_5, A_5\}$	Y_5
Case 6	$\{\vec{L}_6, A_6\}$	Y_6
Case 7	$\{\vec{L}_7, A_7\}$	Y_7
Case 8	$\{\vec{L}_8, A_8\}$	Y_8
Case 9	$\{\vec{L}_9, A_9\}$	Y_9

Selecting an algorithm: The role of a train-test split

1) Randomly assign cases to a `train` and `test` set

`train`	Case 1	$\{\vec{L}_1, A_1\}$	Y_1
`train`	Case 2	$\{\vec{L}_2, A_2\}$	Y_2
`test`	Case 3	$\{\vec{L}_3, A_3\}$	Y_3
`train`	Case 4	$\{\vec{L}_4, A_4\}$	Y_4
`test`	Case 5	$\{\vec{L}_5, A_5\}$	Y_5
`test`	Case 6	$\{\vec{L}_6, A_6\}$	Y_6
`test`	Case 7	$\{\vec{L}_7, A_7\}$	Y_7
`train`	Case 8	$\{\vec{L}_8, A_8\}$	Y_8
`train`	Case 9	$\{\vec{L}_9, A_9\}$	Y_9

Selecting an algorithm: The role of a train-test split

2) First, use only the `train` set.

`train`	Case 1	$\{\vec{L}_1, A_1\}$	Y_1
`train`	Case 2	$\{\vec{L}_2, A_2\}$	Y_2
`test`	Case 3	$\{\vec{L}_3, A_3\}$	Y_3
`train`	Case 4	$\{\vec{L}_4, A_4\}$	Y_4
`test`	Case 5	$\{\vec{L}_5, A_5\}$	Y_5
`test`	Case 6	$\{\vec{L}_6, A_6\}$	Y_6
`test`	Case 7	$\{\vec{L}_7, A_7\}$	Y_7
`train`	Case 8	$\{\vec{L}_8, A_8\}$	Y_8
`train`	Case 9	$\{\vec{L}_9, A_9\}$	Y_9

Selecting an algorithm: The role of a train-test split

2) First, use only the `train` set. Learn a prediction function.

$$f() : \{\vec{L}, A\} \rightarrow Y$$

`train`	Case 1	$\{\vec{L}_1, A_1\}$	$\longrightarrow$	Y_1
`train`	Case 2	$\{\vec{L}_2, A_2\}$	$\longrightarrow$	Y_2
`test`	Case 3	$\{\vec{L}_3, A_3\}$		Y_3
`train`	Case 4	$\{\vec{L}_4, A_4\}$	$\longrightarrow$	Y_4
`test`	Case 5	$\{\vec{L}_5, A_5\}$		Y_5
`test`	Case 6	$\{\vec{L}_6, A_6\}$		Y_6
`test`	Case 7	$\{\vec{L}_7, A_7\}$		Y_7
`train`	Case 8	$\{\vec{L}_8, A_8\}$	$\longrightarrow$	Y_8
`train`	Case 9	$\{\vec{L}_9, A_9\}$	$\longrightarrow$	Y_9

Selecting an algorithm: The role of a train-test split

3) Open the `test` set.

`train`	Case 1	$\{\vec{L}_1, A_1\}$	$\longrightarrow$	Y_1
`train`	Case 2	$\{\vec{L}_2, A_2\}$	$\longrightarrow$	Y_2
`test`	Case 3	$\{\vec{L}_3, A_3\}$		Y_3
`train`	Case 4	$\{\vec{L}_4, A_4\}$	$\longrightarrow$	Y_4
`test`	Case 5	$\{\vec{L}_5, A_5\}$		Y_5
`test`	Case 6	$\{\vec{L}_6, A_6\}$		Y_6
`test`	Case 7	$\{\vec{L}_7, A_7\}$		Y_7
`train`	Case 8	$\{\vec{L}_8, A_8\}$	$\longrightarrow$	Y_8
`train`	Case 9	$\{\vec{L}_9, A_9\}$	$\longrightarrow$	Y_9

Selecting an algorithm: The role of a train-test split

3) Open the `test` set. Predict.

`train`	Case 1	$\{\vec{L}_1, A_1\}$	$\longrightarrow$		Y_1
`train`	Case 2	$\{\vec{L}_2, A_2\}$	$\longrightarrow$		Y_2
`test`	Case 3	$\{\vec{L}_3, A_3\}$	$\longrightarrow$	$\hat{Y}_3$	Y_3
`train`	Case 4	$\{\vec{L}_4, A_4\}$	$\longrightarrow$		Y_4
`test`	Case 5	$\{\vec{L}_5, A_5\}$	$\longrightarrow$	$\hat{Y}_5$	Y_5
`test`	Case 6	$\{\vec{L}_6, A_6\}$	$\longrightarrow$	$\hat{Y}_6$	Y_6
`test`	Case 7	$\{\vec{L}_7, A_7\}$	$\longrightarrow$	$\hat{Y}_7$	Y_7
`train`	Case 8	$\{\vec{L}_8, A_8\}$	$\longrightarrow$		Y_8
`train`	Case 9	$\{\vec{L}_9, A_9\}$	$\longrightarrow$		Y_9

Selecting an algorithm: The role of a train-test split

3) Open the `test` set. Predict. Evaluate squared error.

`train`	Case 1	$\{\vec{L}_1, A_1\}$	$\longrightarrow$	Y_1
`train`	Case 2	$\{\vec{L}_2, A_2\}$	$\longrightarrow$	Y_2
`test`	Case 3	$\{\vec{L}_3, A_3\}$	$\longrightarrow$	$(\hat{Y}_3 - Y_3)^2$
`train`	Case 4	$\{\vec{L}_4, A_4\}$	$\longrightarrow$	Y_4
`test`	Case 5	$\{\vec{L}_5, A_5\}$	$\longrightarrow$	$(\hat{Y}_5 - Y_5)^2$
`test`	Case 6	$\{\vec{L}_6, A_6\}$	$\longrightarrow$	$(\hat{Y}_6 - Y_6)^2$
`test`	Case 7	$\{\vec{L}_7, A_7\}$	$\longrightarrow$	$(\hat{Y}_7 - Y_7)^2$
`train`	Case 8	$\{\vec{L}_8, A_8\}$	$\longrightarrow$	Y_8
`train`	Case 9	$\{\vec{L}_9, A_9\}$	$\longrightarrow$	Y_9

Selecting an algorithm: The role of a train-test split

3) Open the `test` set. Predict. Evaluate squared error. Average.

`train`	Case 1	$\{\vec{L}_1, A_1\}$	$\longrightarrow$	Y_1
`train`	Case 2	$\{\vec{L}_2, A_2\}$	$\longrightarrow$	Y_2
`test`	Case 3	$\{\vec{L}_3, A_3\}$	$\longrightarrow$	$(\hat{Y}_3 - Y_3)^2$
`train`	Case 4	$\{\vec{L}_4, A_4\}$	$\longrightarrow$	Y_4
`test`	Case 5	$\{\vec{L}_5, A_5\}$	$\longrightarrow$	$(\hat{Y}_5 - Y_5)^2$
`test`	Case 6	$\{\vec{L}_6, A_6\}$	$\longrightarrow$	$(\hat{Y}_6 - Y_6)^2$
`test`	Case 7	$\{\vec{L}_7, A_7\}$	$\longrightarrow$	$(\hat{Y}_7 - Y_7)^2$
`train`	Case 8	$\{\vec{L}_8, A_8\}$	$\longrightarrow$	Y_8
`train`	Case 9	$\{\vec{L}_9, A_9\}$	$\longrightarrow$	Y_9

$$\widehat{\text{MSE}} = \frac{1}{n_{\texttt{test}}} \sum_{i \in \texttt{test}} (\hat{Y}_i - Y_i)^2$$

Is your SATT where it's at? **Data structure**

- Simplified dataset on the course site: `lecture_10.csv`
- This is one simulation from the many in Dorie et al.
- Variables include
  - `y`: outcome (numeric)
  - `z`: treatment (binary)
  - `x_*`: confounders
  - `set`: I created this, coded `train` or `test`

Code in `lecture_10_example_code.R` can help you get started.

At the end of class, you will produce on $\widehat{\text{SATT}}$.
Report your estimate here: tinyurl.com/SATTEstimate
I have the truth. We will see who is closest!

Learning goals for today

At the end of class, you will be able to:

1. Use machine learning methods to estimate causal effects
2. Select an estimator using predictive performance

Let me know what you are thinking

tinyurl.com/CausalQuestions

Office hours TTh 11am-12pm and at
calendly.com/ianlundberg/office-hours
Come say hi!