Data Science for Causal Inference

Ryan T. Moore

American University

The Lab @ DC

2024-07-15

Table of contents I

Introductions

About Me

- Associate Prof of Government
  (American University)
- Associate Director, Center for Data Science
  (American University)
- Senior Social Scientist
  (The Lab @ DC)
- Fellow in Methodology
  (US Office of Evaluation Sciences: "OES")

About Me

- Associate Prof of Government
  (American University)
- Associate Director, Center for Data Science
  (American University)
- Senior Social Scientist
  (The Lab @ DC)
- Fellow in Methodology
  (US Office of Evaluation Sciences: "OES")
- Research agenda: political methodology, causal inference, experimental design, experiments in public policy

About You!

- Name?

About You!

- Name?
- Role?

About You!

- Name?
- Role?
- Interests?

About You!

- Name?
- Role?
- Interests?
- Olympic sport you look forward to?

Plan

- Data Science in Causal Inference

Plan

- Data Science in Causal Inference
  - Models

Plan

- Data Science in Causal Inference
  - Models
  - Heterogeneous treatment effects

Plan

- Data Science in Causal Inference
  - Models
  - Heterogeneous treatment effects
  - Variable selection

Plan

- Data Science in Causal Inference
  - Models
  - Heterogeneous treatment effects
  - Variable selection
- Sensitivity

Plan

- Data Science in Causal Inference
  - Models
  - Heterogeneous treatment effects
  - Variable selection
- Sensitivity
  - Model specification

Plan

- Data Science in Causal Inference
  - Models
  - Heterogeneous treatment effects
  - Variable selection
- Sensitivity
  - Model specification
  - Unobservable parameter

Plan

- Data Science in Causal Inference
  - Models
  - Heterogeneous treatment effects
  - Variable selection
- Sensitivity
  - Model specification
  - Unobservable parameter
  - Unobserved confounders

Plan

- Data Science in Causal Inference
  - Models
  - Heterogeneous treatment effects
  - Variable selection
- Sensitivity
  - Model specification
  - Unobservable parameter
  - Unobserved confounders
- Modern difference-in-difference designs

Plan

- Data Science in Causal Inference
  - Models
  - Heterogeneous treatment effects
  - Variable selection
- Sensitivity
  - Model specification
  - Unobservable parameter
  - Unobserved confounders
- Modern difference-in-difference designs
  - Canonical DiD

Plan

- Data Science in Causal Inference
  - Models
  - Heterogeneous treatment effects
  - Variable selection
- Sensitivity
  - Model specification
  - Unobservable parameter
  - Unobserved confounders
- Modern difference-in-difference designs
  - Canonical DiD
  - Multiple time periods

Plan

- Data Science in Causal Inference
  - Models
  - Heterogeneous treatment effects
  - Variable selection
- Sensitivity
  - Model specification
  - Unobservable parameter
  - Unobserved confounders
- Modern difference-in-difference designs
  - Canonical DiD
  - Multiple time periods
  - Staggered adoption

Plan

- Data Science in Causal Inference
  - Models
  - Heterogeneous treatment effects
  - Variable selection
- Sensitivity
  - Model specification
  - Unobservable parameter
  - Unobserved confounders
- Modern difference-in-difference designs
  - Canonical DiD
  - Multiple time periods
  - Staggered adoption
  - Calloway-Sant'Anna approach

Data Science in Causal Inference

Causal Inference Approaches

The "potential outcomes" framework:

Causal Inference Approaches

The "potential outcomes" framework:

Citizen	Canvass?	Would Enroll if Canvass?	Would Enroll if No Canvass?	Enroll
1	Yes	Yes		Yes
2	Yes			Yes
3	No			No
4	No			No

Causal Inference Approaches

The "potential outcomes" framework:

Citizen	Canvass?	Would Enroll if Canvass?	Would Enroll if No Canvass?	Enroll
1	Yes	Yes		Yes
2	Yes	Yes		Yes
3	No			No
4	No			No

Causal Inference Approaches

The "potential outcomes" framework:

Citizen	Canvass?	Would Enroll if Canvass?	Would Enroll if No Canvass?	Enroll
1	Yes	Yes		Yes
2	Yes	Yes		Yes
3	No		No	No
4	No			No

Causal Inference Approaches

The "potential outcomes" framework:

Citizen	Canvass?	Would Enroll if Canvass?	Would Enroll if No Canvass?	Enroll
1	Yes	Yes		Yes
2	Yes	Yes		Yes
3	No		No	No
4	No		No	No

Causal Inference Approaches

The "potential outcomes" framework:

Citizen	Canvass?	Would Enroll if Canvass?	Would Enroll if No Canvass?	Enroll
1	Yes	Yes	(Yes)	Yes
2	Yes	Yes	(No)	Yes
3	No	(Yes)	No	No
4	No	(No)	No	No

Causal Inference Approaches

The "potential outcomes" framework, more abstractly:

Unit i	Treatment T	$Y(1)$	$Y(0)$	Y^{obs}	True τ $Y(1) - Y(0)$
1	1	10		10	
2	1	20		20	
3	0		15	15	
4	0		5	5	

Causal Inference Approaches

The "potential outcomes" framework, more abstractly:

Unit i	Treatment T	$Y(1)$	$Y(0)$	Y^{obs}	True τ $Y(1) - Y(0)$
1	1	10	(10)	10	0
2	1	20	(10)	20	10
3	0	(40)	15	15	25
4	0	(20)	5	5	15

Causal Inference Approaches

The "potential outcomes" framework, more abstractly:

Unit i	Treatment T	$Y(1)$	$Y(0)$	Y^{obs}	True τ $Y(1) - Y(0)$
1	1	10	(10)	10	0
2	1	20	(10)	20	10
3	0	(40)	15	15	25
4	0	(20)	5	5	15
				ATE $= \bar{\tau} =$	$\frac{50}{4} = 12.5$

Causal Inference Approaches

The "potential outcomes" framework, more abstractly:

Unit i	Treatment T	$Y(1)$	$Y(0)$	Y^{obs}	True τ $Y(1) - Y(0)$
1	1	10	(10)	10	0
2	1	20	(10)	20	10
3	0	(40)	15	15	25
4	0	(20)	5	5	15
				$\text{ATE} = \bar{\tau} =$	$\frac{50}{4} = 12.5$
				$\widehat{ATE} = \hat{\bar{\tau}} =$	$15 - 10 = 5$

Causal Inference Approaches

The "potential outcomes" framework, notation:

- Units indexed by i
- Treatment T_i or D_i or Z_i
- Outcome if treated $Y_i(1)$
- Outcome if control $Y_i(0)$
- True treatment effect $\tau_i = Y_i(1) - Y_i(0)$
- True average treatment effect $\bar{\tau} = \frac{1}{n} \sum_{i=1}^{n} (Y_i(1) - Y_i(0))$
- Pre-treatment covariates $\mathbf{X}$

Causal Inference Approaches

The "potential outcomes" framework, notation:

- Units indexed by i
- Treatment T_i or D_i or Z_i
- Outcome if treated $Y_i(1)$
- Outcome if control $Y_i(0)$
- True treatment effect $\tau_i = Y_i(1) - Y_i(0)$
- True average treatment effect $\bar{\tau} = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$
- Pre-treatment covariates $\mathbf{X}$

(and we'll draw some DAG's, too)

Data Science Approaches

Three tasks of data science:

- Description

Data Science Approaches

Three tasks of data science:

- Description
- Prediction

Data Science Approaches

Three tasks of data science:

- Description
- Prediction
- Causal Inference

Data Science Approaches

Three tasks of data science:

- Description
- Prediction
- Causal Inference

Data Science Approaches

Three tasks of data science:

- Description
- Prediction
- Causal Inference

Models/algorithms central to all three.

Data Science Approaches

Three tasks of data science:

- Description
- Prediction
- Causal Inference

Models/algorithms central to all three.

Hernán, Hsu, and Healy (2019)

Data Science Approaches

Description

- Identifying patterns, etc.

Data Science Approaches

Description

- Identifying patterns, etc.
- E.g., clustering to discover groups

Data Science Approaches

Prediction

- Components

Data Science Approaches

Prediction

- Components
  - Inputs/outputs (predictors/outcomes, features/responses, ...)

Data Science Approaches

Prediction

- Components
  - Inputs/outputs (predictors/outcomes, features/responses, ...)
  - Mapping from inputs to outputs (linear model, decision tree, ...)

Data Science Approaches

Prediction

- Components
  - Inputs/outputs (predictors/outcomes, features/responses, ...)
  - Mapping from inputs to outputs (linear model, decision tree, ...)
  - Metric for evaluating mapping

Data Science Approaches

Prediction

- Components
  - Inputs/outputs (predictors/outcomes, features/responses, ...)
  - Mapping from inputs to outputs (linear model, decision tree, ...)
  - Metric for evaluating mapping
- With these, model machine learning does the work

Data Science Approaches

Prediction

- Components
  - Inputs/outputs (predictors/outcomes, features/responses, ...)
  - Mapping from inputs to outputs (linear model, decision tree, ...)
  - Metric for evaluating mapping
- With these, model machine learning does the work
- E.g., regression, random forests, neural networks, ...

Data Science Approaches

Causal Inference

- Potential outcomes/counterfactual/interventionist perspective

Data Science Approaches

Causal Inference

- Potential outcomes/counterfactual/interventionist perspective
- Requires *expertise* different to description/prediction

Data Science Approaches

Causal Inference

- Potential outcomes/counterfactual/interventionist perspective
- Requires *expertise* different to description/prediction
- Requires more than summary statistics, metrics, etc.

Data Science Approaches

Causal Inference

- Potential outcomes/counterfactual/interventionist perspective
- Requires *expertise* different to description/prediction
- Requires more than summary statistics, metrics, etc.
- Requires some knowledge of causal structure

Data Science Approaches

Causal Inference

- Potential outcomes/counterfactual/interventionist perspective
- Requires *expertise* different to description/prediction
- Requires more than summary statistics, metrics, etc.
- Requires some knowledge of causal structure
  - Not all inputs treated same

Data Science Approaches

Causal Inference

- Potential outcomes/counterfactual/interventionist perspective
- Requires *expertise* different to description/prediction
- Requires more than summary statistics, metrics, etc.
- Requires some knowledge of causal structure
  - Not all inputs treated same
  - T v. $\mathbf{X}$ – very different!

Data Science Approaches

Causal Inference

- Potential outcomes/counterfactual/interventionist perspective
- Requires *expertise* different to description/prediction
- Requires more than summary statistics, metrics, etc.
- Requires some knowledge of causal structure
  - Not all inputs treated same
  - T v. $\mathbf{X}$ – very different!
  - (the more knowledge, the better!)

Data Science Approaches

Causal Inference

- Potential outcomes/counterfactual/interventionist perspective
- Requires *expertise* different to description/prediction
- Requires more than summary statistics, metrics, etc.
- Requires some knowledge of causal structure
  - Not all inputs treated same
  - T v. $\mathbf{X}$ – very different!
  - (the more knowledge, the better!)
  - (alternative: solve fundamental problem of causal inference!)

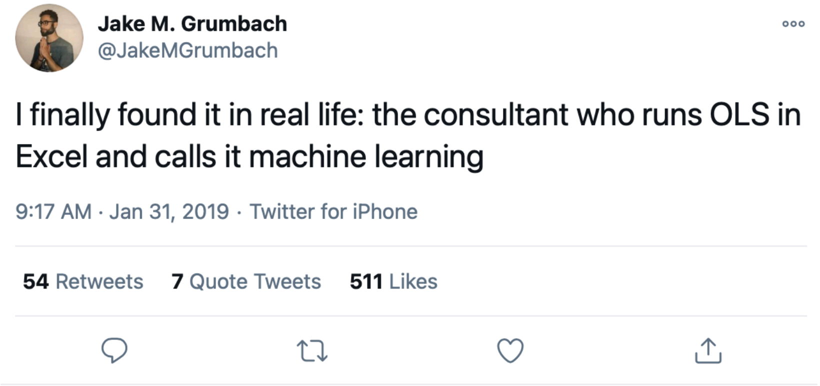
Data Science Approaches

Causal Inference

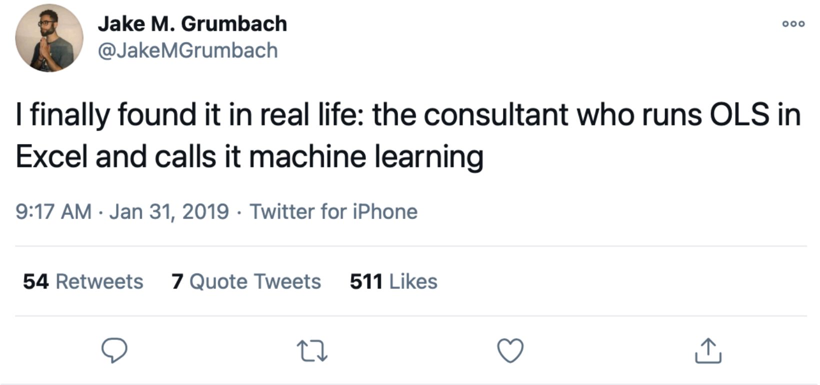
- Potential outcomes/counterfactual/interventionist perspective
- Requires *expertise* different to description/prediction
- Requires more than summary statistics, metrics, etc.
- Requires some knowledge of causal structure
  - Not all inputs treated same
  - T v. $\mathbf{X}$ – very different!
  - (the more knowledge, the better!)
  - (alternative: solve fundamental problem of causal inference!)
- E.g., experiments, observational causal designs, ...

Causal Inference with Machine Learning

Causal Inference with Machine Learning



Causal Inference with Machine Learning


Jake M. Grumbach
@JakeMGrumbach

I finally found it in real life: the consultant who runs OLS in Excel and calls it machine learning

9:17 AM · Jan 31, 2019 · Twitter for iPhone

54 Retweets 7 Quote Tweets 511 Likes

(OK, not "machine learning", perhaps, but *models* at least ...)

Causal Inference with Models

Loaded two datasets:

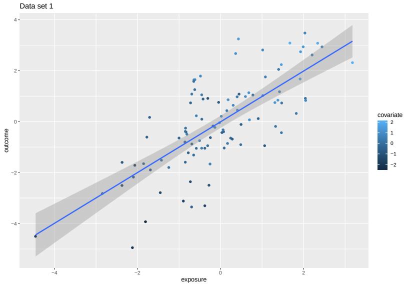
```
str(df1)
```

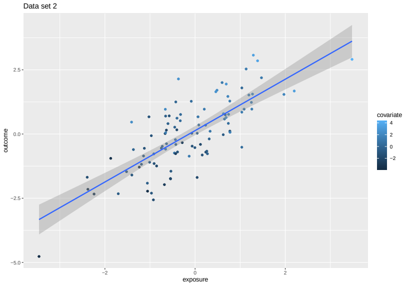
```
tibble [100 x 3] (S3: tbl_df/tbl/data.frame)
 $ covariate: num [1:100] -0.622 1.137 -0.238 1.529 -0.154
 $ exposure : num [1:100] 0.0332 0.3627 0.2422 1.4633 0.779
 $ outcome  : num [1:100] -0.429 2.675 -0.647 2.238 1.044
```

```
str(df2)
```

```
tibble [100 x 3] (S3: tbl_df/tbl/data.frame)
 $ exposure : num [1:100] 0.4862 0.0653 -1.4021 -0.546 -0.4
 $ outcome  : num [1:100] 1.706 0.669 -1.597 -1.733 0.617
 $ covariate: num [1:100] 2.24 0.924 -0.999 -2.343 0.207 ..
```

Causal Inference with Models





Causal Inference with Models

Model each

```r
lm_df1 <- lm(outcome ~ exposure, data = df1)
lm_df2 <- lm(outcome ~ exposure, data = df2)
```

```
# A tibble: 4 x 4
  data  term        estimate std.error
  <chr> <chr>          <dbl>     <dbl>
1 df1   (Intercept) -0.00671    0.120
2 df1   exposure     0.996      0.0927
3 df2   (Intercept)  0.133      0.0890
4 df2   exposure     1.00       0.0841
```

Causal Inference with Models

Model each

```r
lm_df1 <- lm(outcome ~ exposure, data = df1)
lm_df2 <- lm(outcome ~ exposure, data = df2)
```

```
# A tibble: 4 x 4
  data  term        estimate std.error
  <chr> <chr>          <dbl>     <dbl>
1 df1   (Intercept) -0.00671    0.120
2 df1   exposure     0.996      0.0927
3 df2   (Intercept)  0.133      0.0890
4 df2   exposure     1.00       0.0841
```

- Both cases: effect of exposure ≈ 1.

Causal Inference with Models

Model each

```
lm_df1 <- lm(outcome ~ exposure, data = df1)
lm_df2 <- lm(outcome ~ exposure, data = df2)
```

```
# A tibble: 4 x 4
  data  term        estimate std.error
  <chr> <chr>          <dbl>     <dbl>
1 df1   (Intercept) -0.00671    0.120
2 df1   exposure     0.996      0.0927
3 df2   (Intercept)  0.133      0.0890
4 df2   exposure     1.00       0.0841
```

- Both cases: effect of exposure ≈ 1.
- Is this good?

Causal Inference with Models

Model each

```r
lm_df1 <- lm(outcome ~ exposure, data = df1)
lm_df2 <- lm(outcome ~ exposure, data = df2)
```

```
# A tibble: 4 x 4
  data  term        estimate std.error
  <chr> <chr>          <dbl>     <dbl>
1 df1   (Intercept) -0.00671    0.120
2 df1   exposure     0.996      0.0927
3 df2   (Intercept)  0.133      0.0890
4 df2   exposure     1.00       0.0841
```

- Both cases: effect of exposure ≈ 1.
- Is this good?
- What if we adjust for covariate?

Causal Inference with Models

```
lm_df1_adj <- lm(outcome ~ exposure + covariate, data = df1
lm_df2_adj <- lm(outcome ~ exposure + covariate, data = df2
```

```
# A tibble: 4 x 4
  data  term       estimate std.error
  <chr> <chr>         <dbl>     <dbl>
1 df1   exposure      0.501    0.108
2 df1   covariate     0.970    0.147
3 df2   exposure      0.554    0.0990
4 df2   covariate     0.385    0.0598
```

- Both cases: effect of exposure ≈ 0.5.

Causal Inference with Models

```
lm_df1_adj <- lm(outcome ~ exposure + covariate, data = df1
lm_df2_adj <- lm(outcome ~ exposure + covariate, data = df2
```

```
# A tibble: 4 x 4
  data  term      estimate std.error
  <chr> <chr>        <dbl>     <dbl>
1 df1   exposure     0.501    0.108
2 df1   covariate    0.970    0.147
3 df2   exposure     0.554    0.0990
4 df2   covariate    0.385    0.0598
```

- Both cases: effect of exposure ≈ 0.5.
- Is this good?

Causal Inference with Models

```
lm_df1_adj <- lm(outcome ~ exposure + covariate, data = df
lm_df2_adj <- lm(outcome ~ exposure + covariate, data = df
```

```
# A tibble: 4 x 4
  data  term      estimate std.error
  <chr> <chr>        <dbl>     <dbl>
1 df1   exposure     0.501    0.108
2 df1   covariate    0.970    0.147
3 df2   exposure     0.554    0.0990
4 df2   covariate    0.385    0.0598
```

- Both cases: effect of exposure ≈ 0.5.
- Is this good?
- Which is correct? $\beta = 1$? $\beta = 0.5$?

Causal Inference with Models

```r
lm_df1_adj <- lm(outcome ~ exposure + covariate, data = df1
lm_df2_adj <- lm(outcome ~ exposure + covariate, data = df2
```

```
# A tibble: 4 x 4
  data  term      estimate std.error
  <chr> <chr>        <dbl>     <dbl>
1 df1   exposure     0.501    0.108
2 df1   covariate    0.970    0.147
3 df2   exposure     0.554    0.0990
4 df2   covariate    0.385    0.0598
```

- Both cases: effect of exposure ≈ 0.5.
- Is this good?
- Which is correct? $\beta = 1$? $\beta = 0.5$?
- *Should* we adjust for covariate?

Causal Inference with Models

There is nothing in the data that tells us.

Causal Inference with Models

There is nothing in the data that tells us. ☹

Causal Inference with Models

There is nothing in the data that tells us. ☹

Here are the true structures:

Causal Inference with Models

When know structures, adjustment sets for unbiasedness differ:

- `df1`: confounding ⇒ **adjust for** X
- `df2`: collider ⇒ **do not adjust for** X

```
g_conf <- dagitty("dag{ x -> y ; x <- c -> y }")
g_coll <- dagitty("dag{ x -> y ; x -> c <- y }")

adjustmentSets(g_conf, "x", "y")
```

```
{ c }
```

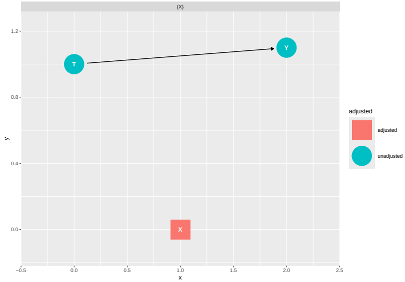
```
adjustmentSets(g_coll, "x", "y")
```

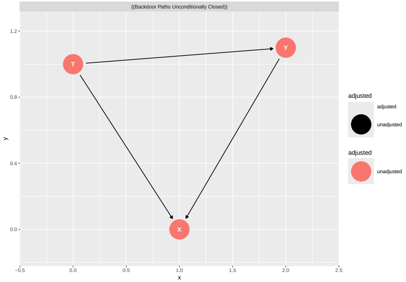
```
{}
```

Causal Inference with Models

When know structures, adjustment sets for unbiasedness differ:

- `df1`: confounding $\Rightarrow$ **adjust for** X
- `df2`: collider $\Rightarrow$ **do not adjust for** X

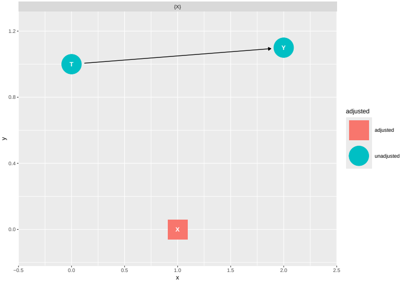


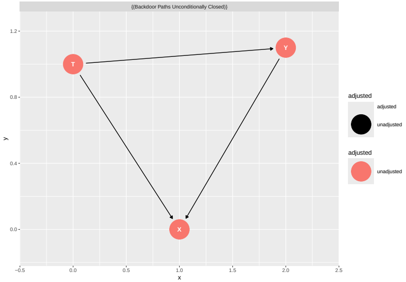


Causal Inference with Models

When know structures, adjustment sets for unbiasedness differ:

- `df1`: confounding ⇒ **adjust for** X
- `df2`: collider ⇒ **do not adjust for** X





(Data from D'Agostino McGowan (2023))

Causal Inference with Models

- Importance of identifying "pre-treatment covariates", "proper covariates"; doing "design before analysis"

Causal Inference with Models

- Importance of identifying “pre-treatment covariates”, “proper covariates”; doing “design before analysis”
- Importance of experiments: strong knowledge about (part of) causal structure

Causal Inference with Models

- Importance of identifying "pre-treatment covariates", "proper covariates"; doing "design before analysis"
- Importance of experiments: strong knowledge about (part of) causal structure
- Causal inference is critical to scientific questions, and separate from prediction

Causal Inference with Models

- Importance of identifying "pre-treatment covariates", "proper covariates"; doing "design before analysis"
- Importance of experiments: strong knowledge about (part of) causal structure
- Causal inference is critical to scientific questions, and separate from prediction
- Though, methods from prediction can aid causal inference

Causal Inference with Models

- Importance of identifying “pre-treatment covariates”, “proper covariates”; doing “design before analysis”
- Importance of experiments: strong knowledge about (part of) causal structure
- Causal inference is critical to scientific questions, and separate from prediction
- Though, methods from prediction can aid causal inference
- (A perspective on “causal euphimisms”: Hernán (2018))

Approaches of Prediction and Causal Inference

Two Cultures, (Breiman 2001)

- *Data Models*: our "social science modeling"
- *Algorithmic Models*: our "data science algorithms"

Methods for Prediction and Causal Inference

- Cross-validation
- Regression/Decision trees
- Random forests

James et al. (2021)

Cross-validation

k-fold cross-validation

- Randomly partition data into k groups
- Apply method to $k-1$ groups
- Use result to predict for left-out group
- Calculate $\mathrm{MSE}_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Calculate test error as average of the k MSE's:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

- Select model that minimises $CV_{(k)}$

CV for Linear Model

```r
library(tidyverse)

## Make data

mk_data <- function(n = 90, n_folds = 10){

  df <- tibble(
    x1 = rnorm(n),
    x2 = rnorm(n),
    x3 = rnorm(n),
    y = 0.1 * x1 + 0.2 * x2 + 0.5 * x3 + rnorm(n),
    cv_fold = sample(rep(1:n_folds, (n / n_folds)))
  )

}

df <- mk_data()
```

CV for Linear Model

```r
head(df)
```

```
# A tibble: 6 x 5
      x1     x2     x3      y cv_fold
   <dbl>  <dbl>  <dbl>  <dbl>   <int>
1  0.964  0.984  0.259  2.55        2
2  0.156 -2.27  -1.92  -1.16        8
3 -0.226  0.476  0.217 -0.936       9
4 -0.713 -1.22   1.97   0.721       6
5  0.903  0.732 -1.08  -0.346       5
6  0.379 -1.53  -1.59  -0.675       6
```

CV for Linear Model

```
head(df)
```

```
# A tibble: 6 x 5
      x1     x2     x3      y cv_fold
   <dbl>  <dbl>  <dbl>  <dbl>   <int>
1  0.964  0.984  0.259  2.55        2
2  0.156 -2.27  -1.92  -1.16        8
3 -0.226  0.476  0.217 -0.936       9
4 -0.713 -1.22   1.97   0.721       6
5  0.903  0.732 -1.08  -0.346       5
6  0.379 -1.53  -1.59  -0.675       6
```

```
table(df$cv_fold)
```

```
 1  2  3  4  5  6  7  8  9 10
 9  9  9  9  9  9  9  9  9  9
```

CV for Linear Model

```
cv_lm <- function(data, fmla){

  n_folds <- max(data$cv_fold)
  store_mses <- vector("numeric", length = n_folds)

  for(idx in 1:n_folds){

    df_train <- data |> filter(cv_fold != idx)
    df_test <- data |> filter(cv_fold == idx)

    lm_out <- lm(fmla, data = df_train)

    predictions <- predict(lm_out, newdata = df_test)

    store_mses[idx] <- mean((df_test$y - predictions)^2)}

  test_error_cv_k <- mean(store_mses)
  return(test_error_cv_k)
```

CV for Linear Model

```r
cv_lm(data = df, fmla = y ~ x1 + x2)
```

```
[1] 1.245331
```

CV for Linear Model

```r
cv_lm(data = df, fmla = y ~ x1 + x2)
```

```
[1] 1.245331
```

```r
df <- mk_data()
cv_lm(df, y ~ x1 + x2)
```

```
[1] 1.430696
```

CV for Linear Model

```
cv_lm(data = df, fmla = y ~ x1 + x2)
```

```
[1] 1.245331
```

```
df <- mk_data()
cv_lm(df, y ~ x1 + x2)
```

```
[1] 1.430696
```

```
df <- mk_data()
cv_lm(df, y ~ x1 + x2 + x3)
```

```
[1] 0.8193824
```

CV for Linear Model

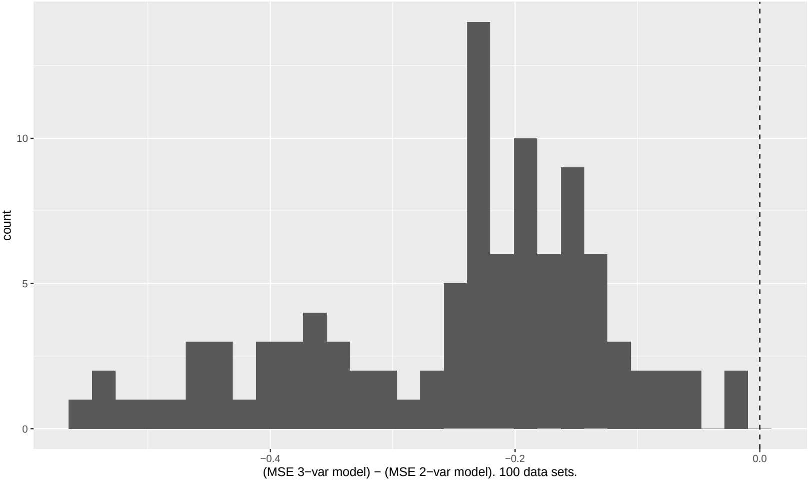


Figure 1: MSE always less (better) for 3-variable model.

Regression Trees

- Partition predictor space into regions $R_1, R_2, \ldots, R_J$.
- If unit falls in region R_j, use average outcome in R_j as predicted value: $\widehat{y}_{R_j}$
- (For *decision* on discrete outcome, count votes in R_j)
- Goal: minimise residual sum of squares (RSS), just like LS regression:

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \widehat{y}_{R_j} \right)$$

Regression Trees

How to define regions R_j?

Regression Trees

How to define regions R_j?

- Top-down, greedy recursive binary split
- At each step, find predictor and cut-point that minimise

$$\sum_{i:x \in R_1(j,s)} \left(y_i - \hat{y}_{R_1(j,s)}\right)^2 + \sum_{i:x \in R_2(j,s)} \left(y_i - \hat{y}_{R_2(j,s)}\right)^2$$

Regression Trees

- Overfitting is a potential problem

Regression Trees

- Overfitting is a potential problem
- Can we increase predictive quality by only using *part* of a tree?

Regression Trees

- Overfitting is a potential problem
- Can we increase predictive quality by only using *part* of a tree?
- "Pruning"

Regression Trees

Pruning

- Build a large tree

Regression Trees

Pruning

- Build a large tree
- Select the subtree that gives least prediction error (via cross-validation)

Regression Trees

Pruning

- Build a large tree
- Select the subtree that gives least prediction error (via cross-validation)
- But, many possible subtrees, so penalise larger trees via α

Regression Trees

Pruning

- Build a large tree
- Select the subtree that gives least prediction error (via cross-validation)
- But, many possible subtrees, so penalise larger trees via α
- α: penalty parameter

Regression Trees

Pruning

- Build a large tree
- Select the subtree that gives least prediction error (via cross-validation)
- But, many possible subtrees, so penalise larger trees via α
- α: penalty parameter
- $|T|$: count of terminal nodes of T

Regression Trees

Pruning

- Build a large tree
- Select the subtree that gives least prediction error (via cross-validation)
- But, many possible subtrees, so penalise larger trees via α
- α: penalty parameter
- $|T|$: count of terminal nodes of T
- m: terminal node index

Regression Trees

Pruning

- Build a large tree
- Select the subtree that gives least prediction error (via cross-validation)
- But, many possible subtrees, so penalise larger trees via α
- α: penalty parameter
- $|T|$: count of terminal nodes of T
- m: terminal node index
- Find subtree that minimises

Regression Trees

Pruning

- Build a large tree
- Select the subtree that gives least prediction error (via cross-validation)
- But, many possible subtrees, so penalise larger trees via α
- α: penalty parameter
- $|T|$: count of terminal nodes of T
- m: terminal node index
- Find subtree that minimises

Regression Trees

Pruning

- Build a large tree
- Select the subtree that gives least prediction error (via cross-validation)
- But, many possible subtrees, so penalise larger trees via α
- α: penalty parameter
- $|T|$: count of terminal nodes of T
- m: terminal node index
- Find subtree that minimises

$$\sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2 + \alpha |T|$$

Regression Trees

Pruning

- Build a large tree
- Select the subtree that gives least prediction error (via cross-validation)
- But, many possible subtrees, so penalise larger trees via α
- α: penalty parameter
- $|T|$: count of terminal nodes of T
- m: terminal node index
- Find subtree that minimises

$$\sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2 + \alpha |T|$$

Sum squared pred. error (plus penalty that grows with tree size) across units in region, then regions.

Regression Trees

But, how to choose α? (Use cross-validation.)

1. Build big tree on training data (with some minimum terminal node size)

Regression Trees

But, how to choose α? (Use cross-validation.)

1. Build big tree on training data (with some minimum terminal node size)
2. For several values of α, find best subtree.

Regression Trees

But, how to choose α? (Use cross-validation.)

1. Build big tree on training data (with some minimum terminal node size)
2. For several values of α, find best subtree.
3. Find best value of α via CV. Create K folds. Then

Regression Trees

But, how to choose α? (Use cross-validation.)

1. Build big tree on training data (with some minimum terminal node size)
2. For several values of α, find best subtree.
3. Find best value of α via CV. Create K folds. Then
   3a. Do 1 and 2 on all but kth fold

Regression Trees

But, how to choose α? (Use cross-validation.)

1. Build big tree on training data (with some minimum terminal node size)
2. For several values of α, find best subtree.
3. Find best value of α via CV. Create K folds. Then
   3a. Do 1 and 2 on all but kth fold
   3b. Predict for kth fold, calculate MSE for several values of α

Regression Trees

But, how to choose α? (Use cross-validation.)

1. Build big tree on training data (with some minimum terminal node size)
2. For several values of α, find best subtree.
3. Find best value of α via CV. Create K folds. Then
   3a. Do 1 and 2 on all but kth fold
   3b. Predict for kth fold, calculate MSE for several values of α
   3c. Get avg MSE for each α

Regression Trees

But, how to choose α? (Use cross-validation.)

1. Build big tree on training data (with some minimum terminal node size)
2. For several values of α, find best subtree.
3. Find best value of α via CV. Create K folds. Then
   3a. Do 1 and 2 on all but kth fold
   3b. Predict for kth fold, calculate MSE for several values of α
   3c. Get avg MSE for each α
   3d. Pick α to minimise MSE

Regression Trees

But, how to choose α? (Use cross-validation.)

1. Build big tree on training data (with some minimum terminal node size)
2. For several values of α, find best subtree.
3. Find best value of α via CV. Create K folds. Then
   3a. Do 1 and 2 on all but kth fold
   3b. Predict for kth fold, calculate MSE for several values of α
   3c. Get avg MSE for each α
   3d. Pick α to minimise MSE

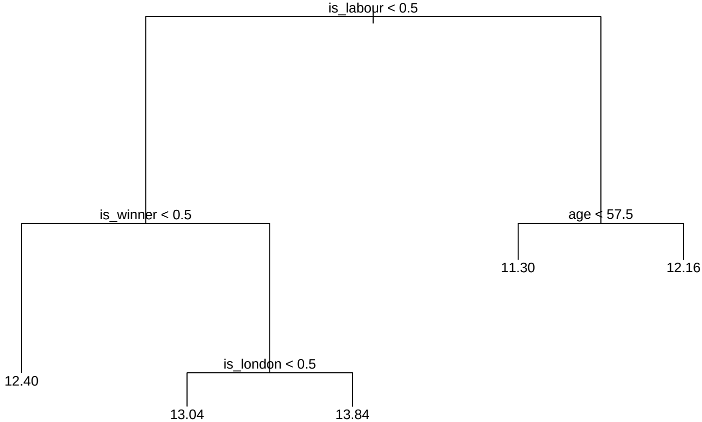
4. Using that α, select best subtree from Step 2

Example: Regression Tree

```
library(qss)
library(rsample)
library(tree)

data("MPs")
mps <- MPs |> mutate(age = yod - yob,
                     is_labour = if_else(party == "labour",
                     is_london = if_else(region == "Greater
                     is_winner = if_else(margin > 0, 1, 0))
  select(ln.net, age, is_labour, is_london, is_winner) |>
  na.omit()

set.seed(765076184)

mp_split <- initial_split(mps, prop = 0.7)

mp_train <- training(mp_split)
mp_test <- testing(mp_split)
```

Example: Regression Tree

```r
plot(tree_mp)
text(tree_mp)
```



Example: Regression Tree

Would pruning help?

```
cv_mps <- cv.tree(tree_mp, K = 10)

plot(cv_mps$size, cv_mps$dev, type = "b")
```

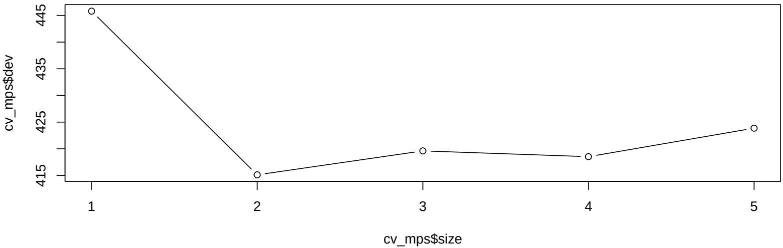


Figure 3: Subtree size 2 minimises SSR

Example: Regression Tree

Would pruning help?

```r
cv_mps <- cv.tree(tree_mp, K = 10)

plot(cv_mps$size, cv_mps$dev, type = "b")
```

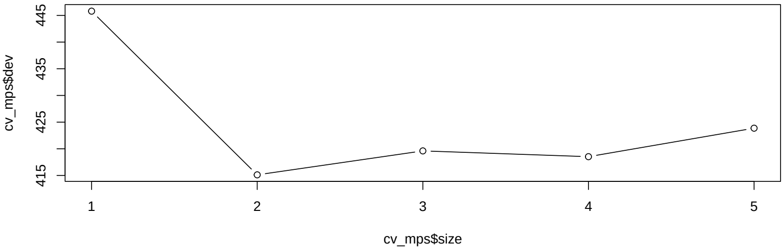


Figure 3: Subtree size 2 minimises SSR

Example: Regression Tree

```
prune_mps <- prune.tree(tree_mp, best = 2)

plot(prune_mps)
text(prune_mps)
```

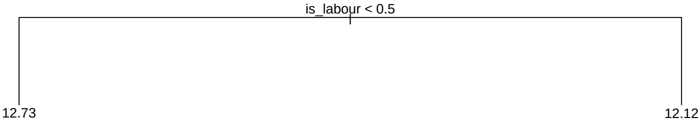


Figure 4: The pruned tree

Example: Regression Tree

Predict for test set:

- MSE for pruned: 1.922
- MSE for full: 1.945

Example: Regression Tree

Predict for test set:

- MSE for pruned: 1.922
- MSE for full: 1.945

So, pruning helped us avoid some overfitting.

Example: Regression Tree

Predict for test set:

- MSE for pruned: 1.922
- MSE for full: 1.945

So, pruning helped us avoid some overfitting.

(Typical pred error of $\sqrt{1.922} \approx 1.386$)

Example: Regression Tree

Predict for test set:

- MSE for pruned: 1.922
- MSE for full: 1.945

So, pruning helped us avoid some overfitting.

(Typical pred error of $\sqrt{1.922} \approx 1.386$)

(Bigger than IQR of 1.183, but range covers $[6.98, 16.3]$.)

Example: Regression Tree

Predict for test set:

- MSE for pruned: 1.922
- MSE for full: 1.945

So, pruning helped us avoid some overfitting.

(Typical pred error of $\sqrt{1.922} \approx 1.386$)

(Bigger than IQR of 1.183, but range covers $[6.98, 16.3]$.)

(Pretty good for 1 split!?)

Random Forests

Heterogeneous Treatment Effects

Variable Selection

Slide Title

Material.

Thanks!

rtm@american.edu
www.ryantmoore.org

References I

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16 (3): 199–215. http://www.jstor.org/stable/2676681.

D'Agostino McGowan, Lucy. 2023. *quartets: Datasets to Help Teach Statistics*. https://r-causal.github.io/quartets/.

Hernán, Miguel A. 2018. "The c-Word: Scientific Euphemisms Do Not Improve Causal Inference from Observational Data." *American Journal of Public Health* 108 (5): 616–19.

Hernán, Miguel A., John Hsu, and Brian Healy. 2019. "A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks." *CHANCE* 32 (1): 42–49. https://doi.org/10.1080/09332480.2019.1579578.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning with Applications in R*. 2nd ed. New York, NY: Springer.