Social Science Inquiry II
Week 9: Beyond linear regression

Molly Offer-Westort

Department of Political Science,
University of Chicago

Winter 2024

Loading packages for this class

```
> set.seed(60637)
> library(ggplot2)
```

- Housekeeping

Machine learning

What is it?

Machine learning

What is it?

- A body of *algorithmic* methods . . .

Machine learning

What is it?

- A body of *algorithmic* methods . . . (an algorithm is just a recipe)

Machine learning

What is it?

- A body of *algorithmic* methods ... (an algorithm is just a recipe)
- Somehow part of *artificial intelligence*...

Machine learning

What is it?

- A body of *algorithmic* methods ... (an algorithm is just a recipe)
- Somehow part of *artificial intelligence*... (basically, how computers perform tasks)

Machine learning

What is it?

- A body of *algorithmic* methods . . . (an algorithm is just a recipe)
- Somehow part of *artificial intelligence*. . . (basically, how computers perform tasks)
- In general, a flexible, *data-driven* approach to make predictions, classify data, or take decisions.

Machine learning

How do the objectives of machine learning differ from those of conventional quantitative methods in the social sciences?

Machine learning

How do the objectives of machine learning differ from those of conventional quantitative methods in the social sciences?

- In conventional quantitative methods:
  - identify and estimate a target estimand, which is often a parameter in a statistical model,

Machine learning

How do the objectives of machine learning differ from those of conventional quantitative methods in the social sciences?

- In conventional quantitative methods:
  - identify and estimate a target estimand, which is often a parameter in a statistical model, defined over some specified population of interest.

Machine learning

How do the objectives of machine learning differ from those of conventional quantitative methods in the social sciences?

- In conventional quantitative methods:
  - identify and estimate a target estimand, which is often a parameter in a statistical model, defined over some specified population of interest.
  - descriptive or causal, e.g., population employment rate, or effect of a policy

Machine learning

How do the objectives of machine learning differ from those of conventional quantitative methods in the social sciences?

- In conventional quantitative methods:
  - identify and estimate a target estimand, which is often a parameter in a statistical model, defined over some specified population of interest.
  - descriptive or causal, e.g., population employment rate, or effect of a policy
- In ML:
  - development of algorithms to make classifications or predictions.

Machine learning

How do the objectives of machine learning differ from those of conventional quantitative methods in the social sciences?

- In conventional quantitative methods:
  - identify and estimate a target estimand, which is often a parameter in a statistical model, defined over some specified population of interest.
  - descriptive or causal, e.g., population employment rate, or effect of a policy
- In ML:
  - development of algorithms to make classifications or predictions.
  - e.g., is this a picture of banana or a cat?

Machine learning

How do the objectives of machine learning differ from those of conventional quantitative methods in the social sciences?

- In conventional quantitative methods:
  - identify and estimate a target estimand, which is often a parameter in a statistical model, defined over some specified population of interest.
  - descriptive or causal, e.g., population employment rate, or effect of a policy
- In ML:
  - development of algorithms to make classifications or predictions.
  - e.g., is this a picture of banana or a cat? Will this person be more likely to click on an ad for sneakers or cookware?

Machine learning

How do the objectives of machine learning differ from those of conventional quantitative methods in the social sciences?

- In conventional quantitative methods:
  - identify and estimate a target estimand, which is often a parameter in a statistical model, defined over some specified population of interest.
  - descriptive or causal, e.g., population employment rate, or effect of a policy
- In ML:
  - development of algorithms to make classifications or predictions.
  - e.g., is this a picture of banana or a cat? Will this person be more likely to click on an ad for sneakers or cookware?
- Is there overlap between the two?

Model fit vs. prediction

- In linear regression, propose a model:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki}$$

- Select $\hat{\beta}_0 \ldots \hat{\beta}_K$ to minimize

$$\sum_{i=1}^{N} \hat{\varepsilon_i}^2 = \sum_{i=1}^{N} \left(\hat{Y}_i - Y_i \right)^2$$

Model fit vs. prediction

- For prediction tasks, we could use the same model,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki}$$

- But select $\hat{\beta}_0 \ldots \hat{\beta}_K$ to minimize squared prediction error for the next observation:

$$\hat{\varepsilon}^2_{N+1} = \left(\hat{Y}_{N+1} - Y_{N+1}\right)^2$$

Model fit vs. prediction

- For prediction tasks, we could use the same model,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki}$$

- But select $\hat{\beta}_0 \dots \hat{\beta}_K$ to minimize squared prediction error for the next observation:

$$\hat{\varepsilon}^2_{N+1} = \left(\hat{Y}_{N+1} - Y_{N+1}\right)^2$$

- Are these the same thing?

Model fit vs. prediction

- For prediction tasks, we could use the same model,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki}$$

- But select $\hat{\beta}_0 \ldots \hat{\beta}_K$ to minimize squared prediction error for the next observation:

$$\hat{\varepsilon}^2_{N+1} = \left(\hat{Y}_{N+1} - Y_{N+1}\right)^2$$

- Are these the same thing? (no)

Model fit vs. prediction

- For prediction tasks, we could use the same model,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki}$$

- But select $\hat{\beta}_0 \ldots \hat{\beta}_K$ to minimize squared prediction error for the next observation:

$$\hat{\varepsilon}^2_{N+1} = \left(\hat{Y}_{N+1} - Y_{N+1}\right)^2$$

- Are these the same thing? (no)
- If prediction is our goal, can we do better than least squares regression?

Model fit vs. prediction

- For prediction tasks, we could use the same model,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki}$$

- But select $\hat{\beta}_0 \ldots \hat{\beta}_K$ to minimize squared prediction error for the next observation:

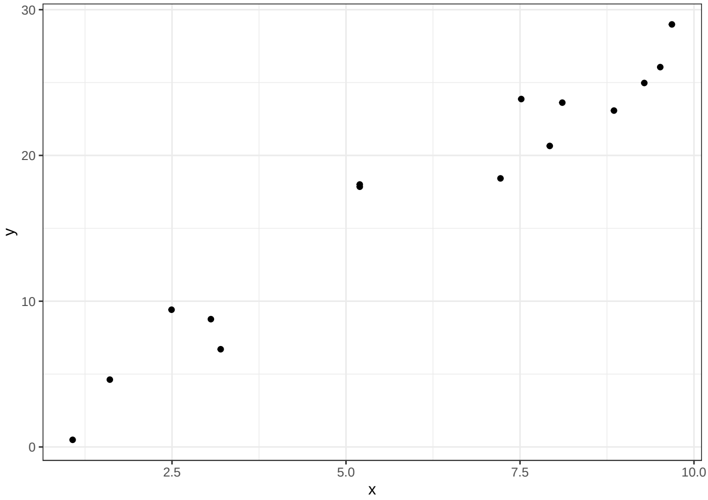
$$\hat{\varepsilon}^2_{N+1} = \left(\hat{Y}_{N+1} - Y_{N+1} \right)^2$$

- Are these the same thing? (no)
- If prediction is our goal, can we do better than least squares regression? (yes)

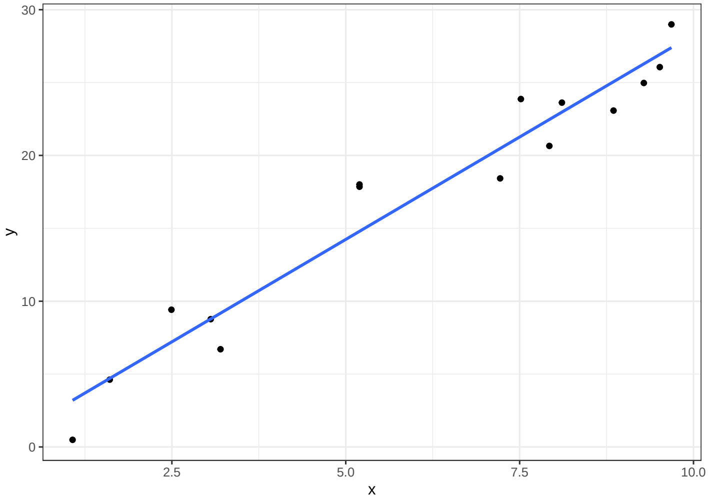
Some ML tools

- A major concern of ML: *overfit*
  - If your model fits the data *too* perfectly, it's not useful for prediction

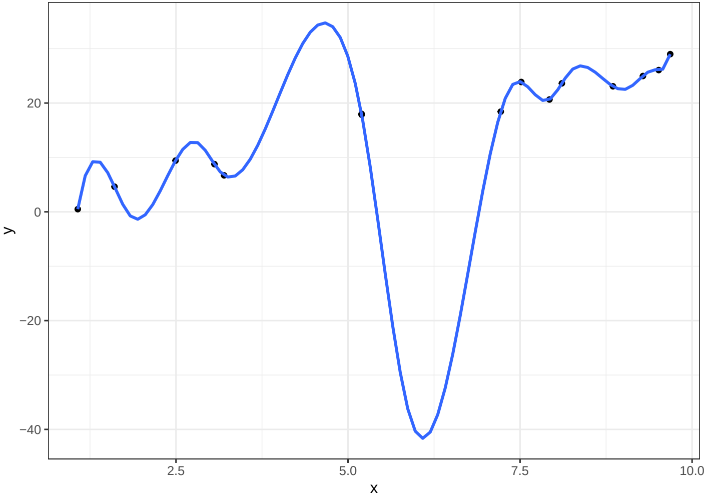
Suppose we would like to fit a model to the following data:



We could use a single line:



Or we could fit a curve that goes between every point:



Cross-validation

- If we were to draw another observation from the joint distribution of (Y, X), which one do you think would do a better job of prediction?

Cross-validation

- If we were to draw another observation from the joint distribution of (Y, X), which one do you think would do a better job of prediction?
- ML methods propose a way to check this, by separating data into training and test sets.

Cross-validation

- If we were to draw another observation from the joint distribution of (Y, X), which one do you think would do a better job of prediction?
- ML methods propose a way to check this, by separating data into training and test sets.
- You can fit different models on the training set, and then see which one does the best job of predicting response in the test set.

Cross-validation

- If we were to draw another observation from the joint distribution of (Y, X), which one do you think would do a better job of prediction?
- ML methods propose a way to check this, by separating data into training and test sets.
- You can fit different models on the training set, and then see which one does the best job of predicting response in the test set. (This is not a new idea.)

Cross-validation

- If we were to draw another observation from the joint distribution of (Y, X), which one do you think would do a better job of prediction?
- ML methods propose a way to check this, by separating data into training and test sets.
- You can fit different models on the training set, and then see which one does the best job of predicting response in the test set. (This is not a new idea.)
- There are some different ways to do this:
  - Leave-k-out
  - Leave-one-out
  - k-fold cross validation

What is our goal in fitting a model?

What is our goal in fitting a model?

- Given some data $(Y_1, \mathrm{X}_1), \ldots, (Y_N, \mathrm{X}_N)$, we fit a model, $\hat{f}(X)$.

What is our goal in fitting a model?

- Given some data $(Y_1, X_1), \ldots, (Y_N, X_N)$, we fit a model, $\hat{f}(X)$.
- Suppose our goal is prediction for the next observation.

What is our goal in fitting a model?

- Given some data $(Y_1, \mathrm{X}_1), \ldots, (Y_N, \mathrm{X}_N)$, we fit a model, $\hat{f}(X)$.
- Suppose our goal is prediction for the next observation.
- Given X_{N+1}, we want to minimize

$$L\left(Y_{N+1}, \hat{f}(X_{N+1})\right) = \left(Y_{N+1} - \hat{f}(X_{N+1})\right)^2$$

What is our goal in fitting a model?

- Given some data $(Y_1, \mathrm{X}_1), \ldots, (Y_N, \mathrm{X}_N)$, we fit a model, $\hat{f}(X)$.
- Suppose our goal is prediction for the next observation.
- Given X_{N+1}, we want to minimize

$$L\left(Y_{N+1}, \hat{f}(X_{N+1})\right) = \left(Y_{N+1} - \hat{f}(X_{N+1})\right)^2$$

- We may be interested not just in how our method performs on one specific observation, but how it performs in expectation

$$\mathrm{Err} = \mathrm{E}\left[L(Y, \hat{f}(X))\right]$$

What is our goal in fitting a model?

- Given some data $(Y_1, \mathrm{X}_1), \ldots, (Y_N, \mathrm{X}_N)$, we fit a model, $\hat{f}(X)$.
- Suppose our goal is prediction for the next observation.
- Given X_{N+1}, we want to minimize

$$L\left(Y_{N+1}, \hat{f}(X_{N+1})\right) = \left(Y_{N+1} - \hat{f}(X_{N+1})\right)^2$$

- We may be interested not just in how our method performs on one specific observation, but how it performs in expectation

$$\mathrm{Err} = \mathrm{E}\left[L(Y, \hat{f}(X))\right]$$

- What should the expectation be taken over?

What is our goal in fitting a model?

- Given some data $(Y_1, \mathrm{X}_1), \ldots, (Y_N, \mathrm{X}_N)$, we fit a model, $\hat{f}(X)$.
- Suppose our goal is prediction for the next observation.
- Given X_{N+1}, we want to minimize

$$L\left(Y_{N+1}, \hat{f}(X_{N+1})\right) = \left(Y_{N+1} - \hat{f}(X_{N+1})\right)^2$$

- We may be interested not just in how our method performs on one specific observation, but how it performs in expectation

$$\mathrm{Err} = \mathrm{E}\left[L(Y, \hat{f}(X))\right]$$

- What should the expectation be taken over? Can/should we hold the data we used for fitting the model fixed?

What is our goal in fitting a model?

- Difference between conditional error and expected test error

What is our goal in fitting a model?

- Difference between conditional error and expected test error
  - Conditional test error:

$$\mathrm{Err}_{\mathcal{T}} = \mathrm{E}\left[L(Y, \hat{f}(X))|\mathcal{T}\right]$$

    Training set $\mathcal{T}$ is fixed.

What is our goal in fitting a model?

- Difference between conditional error and expected test error
  - Conditional test error:

$$\mathrm{Err}_{\mathcal{T}} = \mathrm{E}\left[L(Y, \hat{f}(X))|\mathcal{T}\right]$$

    Training set $\mathcal{T}$ is fixed.
  - Expected test error:

$$\mathrm{Err} = \mathrm{E}\left[L(Y, \hat{f}(X))\right] = \mathrm{E}\left[\mathrm{E}\left[L(Y, \hat{f}(X))|\mathcal{T}\right]\right]$$

What is our goal in fitting a model?

- Difference between conditional error and expected test error
  - Conditional test error:

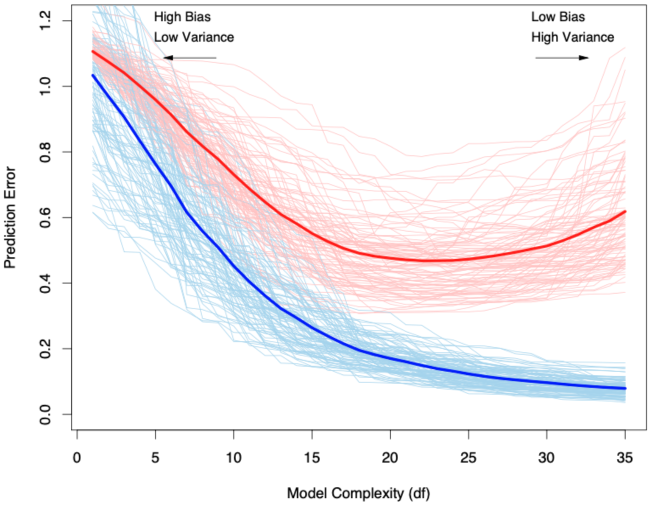
$$\mathrm{Err}_{\mathcal{T}} = \mathrm{E}\left[L(Y, \hat{f}(X))|\mathcal{T}\right]$$

    Training set $\mathcal{T}$ is fixed.
  - Expected test error:

$$\mathrm{Err} = \mathrm{E}\left[L(Y, \hat{f}(X))\right] = \mathrm{E}\left[\mathrm{E}\left[L(Y, \hat{f}(X))|\mathcal{T}\right]\right]$$

- We may be interested in $\mathrm{Err}_{\mathcal{T}}$, in practice most estimating methods will give us estimates of Err.

What is our goal in fitting a model?



Hastie et al. (2009)

Blue is in-sample error, red is out-of-sample error.

What is our goal in fitting a model?

- We may want to use expected test error to **select among models**, or versions of models.

What is our goal in fitting a model?

- We may want to use expected test error to **select among models**, or versions of models.
- And, once we have selected a version of a model, we may want to **assess** how a selected model performs.

What is our goal in fitting a model?

- We can't measure expected test error directly.

What is our goal in fitting a model?

- A procedure that allows us to estimate it:
  - Split data into three parts

What is our goal in fitting a model?

- A procedure that allows us to estimate it:
  - Split data into three parts

  - Fit models to the training set.

What is our goal in fitting a model?

- A procedure that allows us to estimate it:
  - Split data into three parts

  - Fit models to the training set.
  - Estimate prediction error of models in validation set.

What is our goal in fitting a model?

- A procedure that allows us to estimate it:
  - Split data into three parts

  - Fit models to the training set.
  - Estimate prediction error of models in validation set.
  - Select model with minimum error in validation set.

What is our goal in fitting a model?

- A procedure that allows us to estimate it:
  - Split data into three parts

  - Fit models to the training set.
  - Estimate prediction error of models in validation set.
  - Select model with minimum error in validation set.
  - Then get generalization error of just that model on test set.

What is our goal in fitting a model?

- A procedure that allows us to estimate it:
  - Split data into three parts

  - Fit models to the training set.
  - Estimate prediction error of models in validation set.
  - Select model with minimum error in validation set.
  - Then get generalization error of just that model on test set.
- Why do we need to estimate the prediction error of the selected model *again*?

What is our goal in fitting a model?

- A procedure that allows us to estimate it:
  - Split data into three parts

  - Fit models to the training set.
  - Estimate prediction error of models in validation set.
  - Select model with minimum error in validation set.
  - Then get generalization error of just that model on test set.
- Why do we need to estimate the prediction error of the selected model *again*? Winner's curse.

Cross-validation.

- We can potentially get more out of our data by cross-validating.

Version 1	Training	Validation
Version 2	Validation	Training

Cross-validation.

- We can potentially get more out of our data by cross-validating.

Version 1	Training	Validation
Version 2	Validation	Training

$$\widehat{\mathrm{Err}}_{CV} = \sum_{i=1}^{N} L\left(y_i, \hat{f}^{-k(i)}(x_i)\right)$$

$\hat{f}^{-k(i)}$ are the fits from the folds k that do not contain i.

K-fold cross validation.

Version 1	Training	Training	Training	Training	Validation
Version 2	Training	Training	Training	Validation	Training
Version 3	Training	Training	Validation	Training	Training
Version 4	Training	Validation	Training	Training	Training
Version 5	Validation	Training	Training	Training	Training

$$\widehat{\mathrm{Err}}_{CV} = \sum_{i=1}^{N} L\left(y_i, \hat{f}^{-k(i)}(x_i)\right)$$

$\hat{f}^{-k(i)}$ are the fits from the folds k that do not contain i.

Cross-validion.

- How do we pick K?

Cross-validion.

- How do we pick K?
- $K = N$?

Cross-validion.

- How do we pick K?
- $K = N$? Low bias, possibly high variance (our prediction sets are very similar).

Cross-validion.

- How do we pick K?
- $K = N$? Low bias, possibly high variance (our prediction sets are very similar).
- $K = 5$?

Cross-validion.

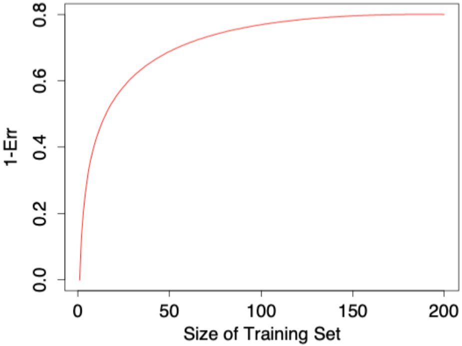
- How do we pick K?
- $K = N$? Low bias, possibly high variance (our prediction sets are very similar).
- $K = 5$? Lower variance, possibly higher bias.

Cross-validion.

- How do we pick K?
- $K = N$? Low bias, possibly high variance (our prediction sets are very similar).
- $K = 5$? Lower variance, possibly higher bias. How much does the prediction change as we change the size of the data set?

Cross-validion.

- How do we pick K?
- $K = N$? Low bias, possibly high variance (our prediction sets are very similar).
- $K = 5$? Lower variance, possibly higher bias. How much does the prediction change as we change the size of the data set?



Hastie et al. (2009)

Cross-validion.

- Rule of thumb is often 5 or 10.

Regularization

- Overfit can become a real problem when we have a lot of predictors (K) relative to our number of observations (N)

Regularization

- Overfit can become a real problem when we have a lot of predictors (K) relative to our number of observations (N)
- This is a common problem when we think about an industry setting, where for every customer a business might have a large number of measurements.

Regularization

- Overfit can become a real problem when we have a lot of predictors (K) relative to our number of observations (N)
- This is a common problem when we think about an industry setting, where for every customer a business might have a large number of measurements. Which ones should they use to predict an outcome?

Regularization

- Consider a model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

Regularization

- Consider a model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

- With $K \geq N$, even if every β_k is non-zero, we won't be able to make good predictions with all of our $\hat{\beta}_k$–and when we care about prediction, that's not our goal, anyhow.

Regularization

- Consider a model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

- With $K \geq N$, even if every β_k is non-zero, we won't be able to make good predictions with all of our $\hat{\beta}_k$–and when we care about prediction, that's not our goal, anyhow.
- With regularization, we shrink some of the $\hat{\beta}_k$ nearly all the way or all the way to zero.

Regularization

- Consider a model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

- With $K \geq N$, even if every β_k is non-zero, we won't be able to make good predictions with all of our $\hat{\beta}_k$–and when we care about prediction, that's not our goal, anyhow.
- With regularization, we shrink some of the $\hat{\beta}_k$ nearly all the way or all the way to zero.
- For *ridge regression* or *lasso*, we select the $\hat{\beta}_k$ using:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(Y_i - \beta_0 + \sum_{k=1}^{K} X_{ki} \beta_k \right)^2 + \lambda \sum_{k=1}^{K} |\beta_k|^q \right\}$$

Hastie et al. (2009)

Regression Trees

- Suppose we have joint data, (Y, X), with just one predictor, X.

Regression Trees

- Suppose we have joint data, (Y, X), with just one predictor, X.
- Our goal is to pick some value of c so that we can split the data into two sub-samples . . .
  - $X_i \leq c$
  - $X_i > c$

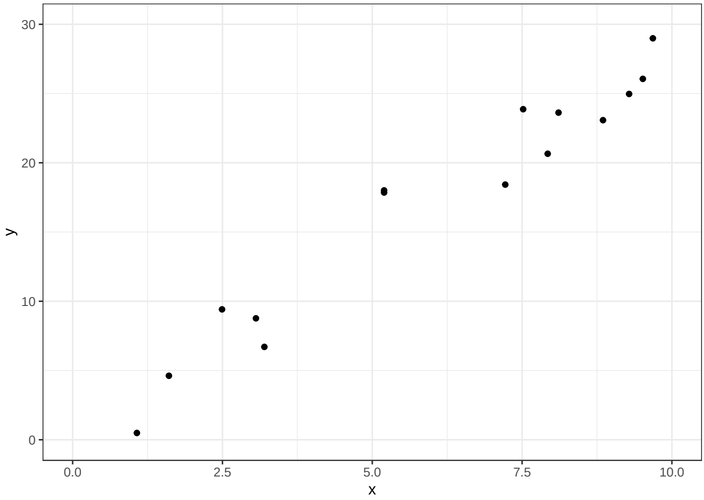
Regression Trees

- Suppose we have joint data, (Y, X), with just one predictor, X.
- Our goal is to pick some value of c so that we can split the data into two sub-samples ...
  - $X_i \leq c$
  - $X_i > c$
- ... and for each sub-sample, predict $\hat{Y}$ as the mean of the Y_i within each sample.

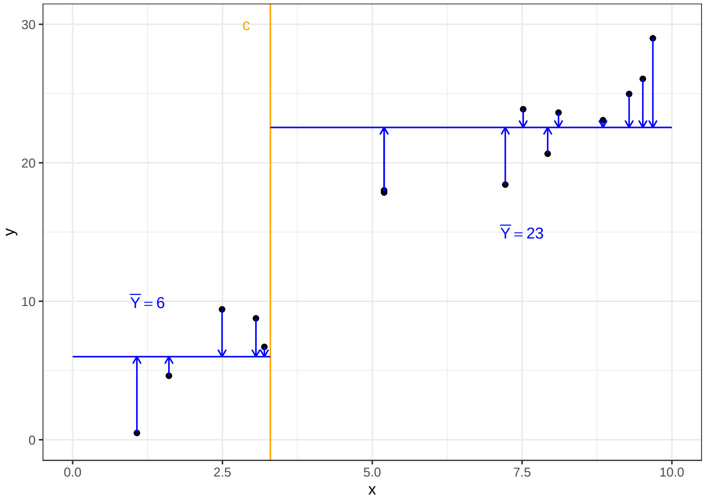
Regression Trees

- Suppose we have joint data, (Y, X), with just one predictor, X.
- Our goal is to pick some value of c so that we can split the data into two sub-samples …
  - $X_i \leq c$
  - $X_i > c$
- …and for each sub-sample, predict $\hat{Y}$ as the mean of the Y_i within each sample.
- We want to pick c to minimize:

$$Q = \sum_{i:X_i \leq c} (Y_i - \bar{Y}_{\text{lower}})^2 + \sum_{i:X_i > c} (Y_i - \bar{Y}_{\text{upper}})^2$$



30
20
10
0
y
0.0
2.5
5.0
7.5
10.0
x


c
$\overline{Y} = 6$
$\overline{Y} = 23$
30
20
10
0
y
0.0
2.5
5.0
7.5
10.0
x

Regression Trees

- Now suppose we have joint data, $(Y, X_1, \ldots, X_k)$.

Regression Trees

- Now suppose we have joint data, $(Y, X_1, \ldots, X_k)$.
- We will do the same approach to finding thresholds to minimize prediction error, but we'll want to pick which X_k we use for threshholding, as well.

Regression Trees

- Now suppose we have joint data, $(Y, X_1, \ldots, X_k)$.
- We will do the same approach to finding thresholds to minimize prediction error, but we'll want to pick which X_k we use for threshholding, as well.
- Generally, we'll define the depth of the tree as 2 or three variables; first we'll split on X_k, then we'll split on $X_j \ldots$

Honesty

- Returning to (causal) inference...

Honesty

- Returning to (causal) inference. . . we might like to use these methods to get valid inference, potentially on causal targets.

An honest tree algorithm

1. Split the sample into two folds.

An honest tree algorithm

1. Split the sample into two folds.
2. Use the first fold to learn splits of the tree.

An honest tree algorithm

1. Split the sample into two folds.
2. Use the first fold to learn splits of the tree.
3. Estimate response within leaves using the second fold.

An honest tree algorithm

1. Split the sample into two folds.
2. Use the first fold to learn splits of the tree.
3. Estimate response within leaves using the second fold.

- This can result in some leaves being empty.

An honest tree algorithm

1. Split the sample into two folds.
2. Use the first fold to learn splits of the tree.
3. Estimate response within leaves using the second fold.

- This can result in some leaves being empty. Prune them?

An honest tree algorithm

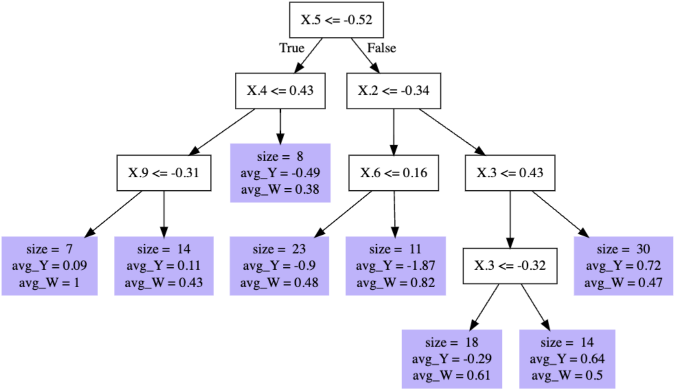
1. Split the sample into two folds.
2. Use the first fold to learn splits of the tree.
3. Estimate response within leaves using the second fold.

- This can result in some leaves being empty. Prune them?
- This procedure reduces bias relative to those proposed by Breiman (2001).

An honest tree algorithm

```
> library(grf)
> set.seed(60637)
> n <- 500
> p <- 10
> X <- matrix(rnorm(n * p), n, p)
> W <- rbinom(n, 1, 0.5)
> Y <- pmax(X[, 1], 0) * W + X[, 2] +
+   pmin(X[, 3], 0) + rnorm(n)
> c.forest <- causal_forest(X, Y, W)
```

An honest tree



Matrix completion: the Nexflix problem

- Netflix has data on viewers, their characteristics, and how they rate movies.

Matrix completion: the Nexflix problem

- Netflix has data on viewers, their characteristics, and how they rate movies.
- The question: how to best recommend to them movies that they have not yet rated?

Matrix completion: the Nexflix problem

- Netflix has data on viewers, their characteristics, and how they rate movies.
- The question: how to best recommend to them movies that they have not yet rated?
- The challenge: come up with the best recommendation algorithm, winner gets $1 million.

Matrix completion: the Nexflix problem

- Netflix has data on viewers, their characteristics, and how they rate movies.
- The question: how to best recommend to them movies that they have not yet rated?
- The challenge: come up with the best recommendation algorithm, winner gets $1 million.
- This can be framed as a matrix completion problem: put users on rows, movies on columns, predict all of the missing rankings.

Causal inference

- Machine learning tools can be super useful for causal inference.

Causal inference

- Machine learning tools can be super useful for causal inference.
  - Fit prediction models separately to treatment and control, so we can do a better job of estimating treatment effects at different covariate values.

Causal inference

- Machine learning tools can be super useful for causal inference.
  - Fit prediction models separately to treatment and control, so we can do a better job of estimating treatment effects at different covariate values.
  - Learn which covariates to include in a (causal) regression model.

Causal inference

- Machine learning tools can be super useful for causal inference.
  - Fit prediction models separately to treatment and control, so we can do a better job of estimating treatment effects at different covariate values.
  - Learn which covariates to include in a (causal) regression model.
  - For observational data, predict propensity to be in treatment vs. control group, based on covariates.

Causal inference: no free lunch

- Machine learning does not solve the fundamental problem of causal inference.

Causal inference: no free lunch

- Machine learning does not solve the fundamental problem of causal inference.
- Causal interpretations are based on assumptions about the data generating process, or knowledge of assignment procedures. These are outside the realm of machine learning methods.

Prediction error

- ML methods may do a good job of producing estimates, but how do we account for inference?

Prediction error

- ML methods may do a good job of producing estimates, but how do we account for inference?
- Cross-validation

Prediction error

- ML methods may do a good job of producing estimates, but how do we account for inference?
- Cross-validation
- Bootstrapping

Prediction error

- ML methods may do a good job of producing estimates, but how do we account for inference?
- Cross-validation
- Bootstrapping
- Applying these solutions to prediction under multiple linear regression

References I

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. Annual Review of Economics, 11:685–725.

Breiman, L. (2001). Random forests. Machine learning, 45:5–32.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction, volume 2. Springer.