Social Science Inquiry II

Week 9: Beyond linear regression

Molly Offer-Westort

Department of Political Science,
University of Chicago

Winter 2023

Loading packages for this class

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

- Housekeeping

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?

- 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:

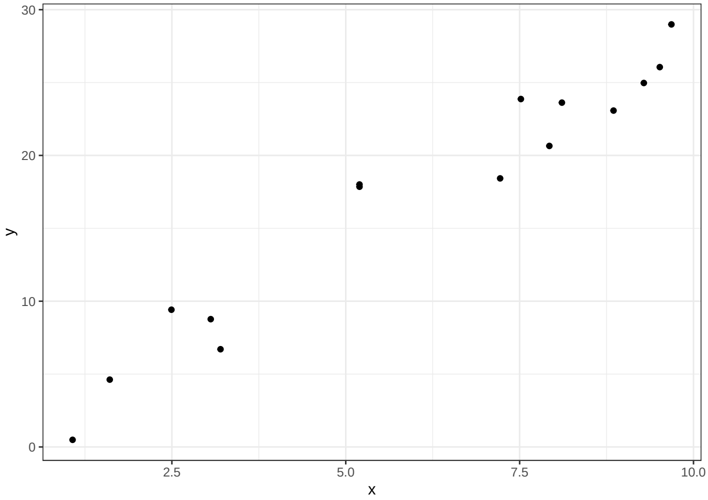
$$\hat{\varepsilon}^2_{N+1} = \left(\hat{Y}_{N+1} - Y_{N+1}\right)^2$$

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

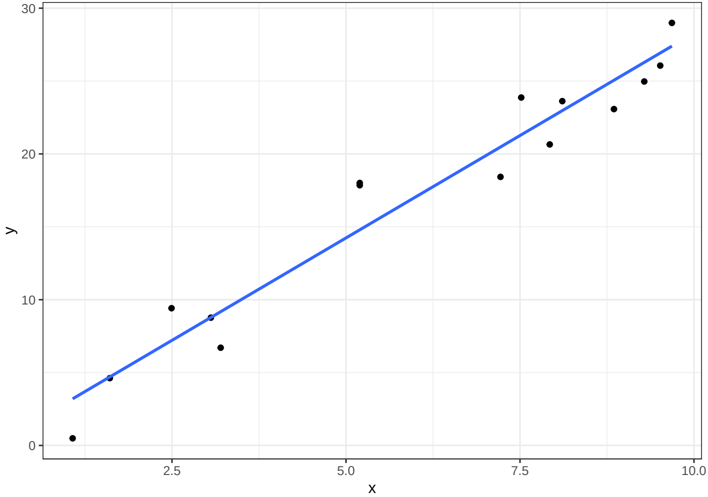
Some ML tools

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

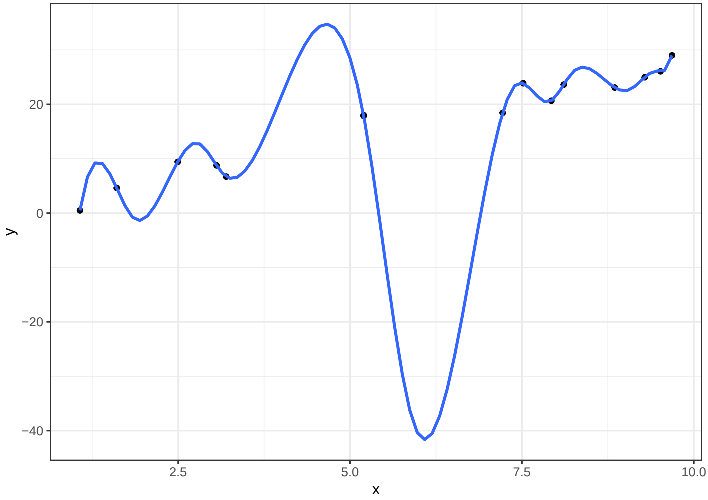
Suppose we would like to fit a model to the following data:



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?
- 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

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:

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

- 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:

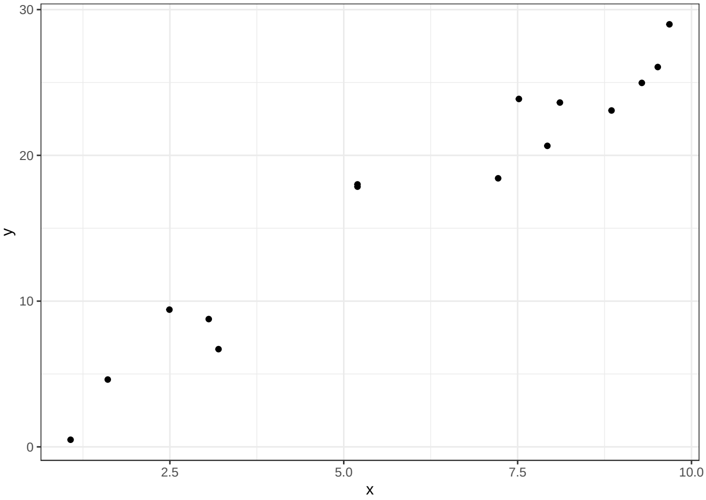
$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(Y_i - \beta_0 + \sum_{k=1}^{K} X_{ki} \beta_k \right)^2 + \lambda \sum_{k=1}^{K} |\beta_k|^q \right\}$$

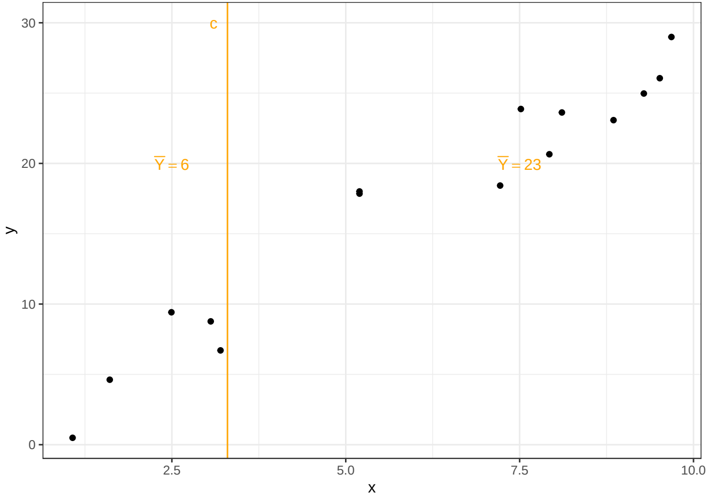
?

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
2.5
5.0
7.5
10.0
x


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

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 $X_j \ldots$

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.
  - 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 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?
- Cross-validation
- Bootstrapping
- Applying these solutions to prediction under multiple linear regression

References I