Ch 5.1.1-2: Leave One Out Cross-validation

Lecture 12 - CMSE 381

Prof. Elizabeth Munch

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Mon, Oct 2, 2023

Announcements

Last time:

- Exam

Announcements:

- Fourth homework due next monday
- Office hours
- Drops

Covered in this lecture

- LOO CV
- Outliers
- Leverage statistic

Section 1

Validation set

What's the problem?

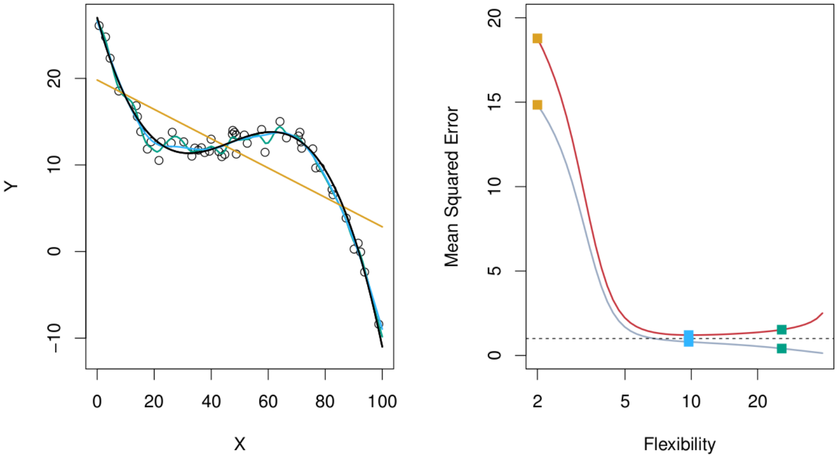
- How well is my ML method doing? *Model Assessment*
- Which method is best for our data?
- How many features should I use? Which ones? *Model selection*
- What is the uncertainty in the learned parameters?

Training Error vs Testing Error

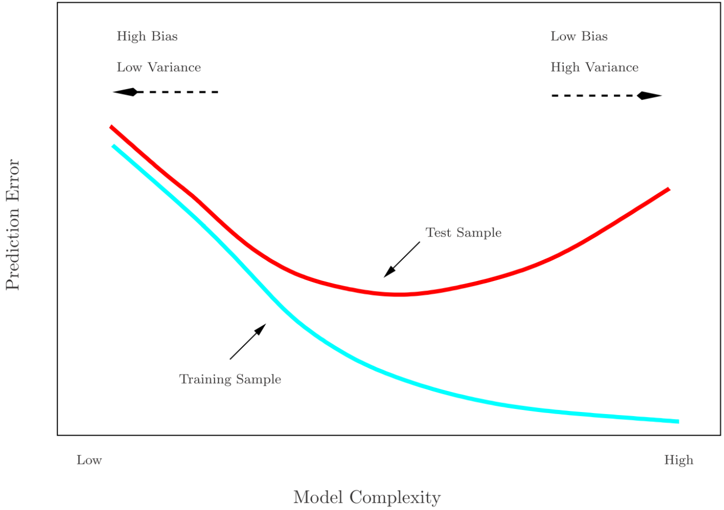
Training Error

Testing Error

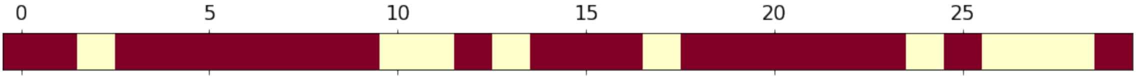
Throw-back Monday



Model tradeoffs



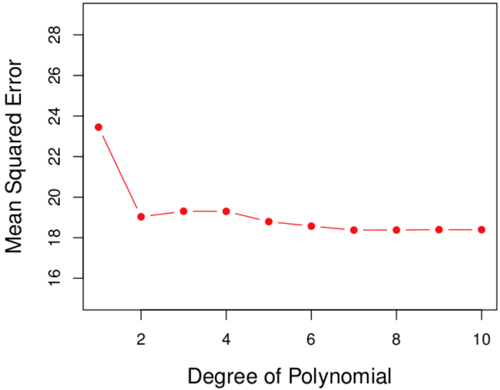
Validation set approach



- Divide randomly into two parts:
  - Training set
  - Validation/Hold-out/Testing set
- Fit model on training set
- Use fitted model to predict response for observations in the test set
- Evaluate quality (e.g. MSE)

Coding example in jupyter notebook

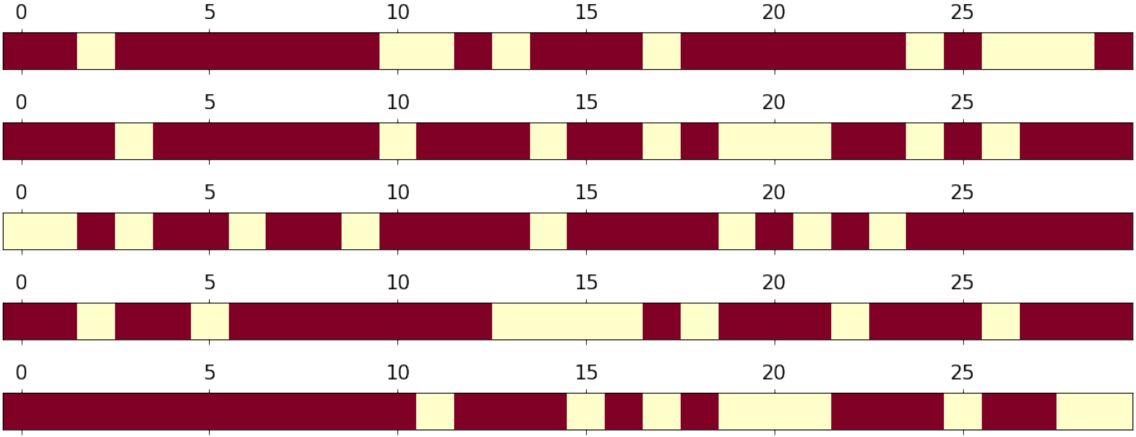
Example with the `auto` data



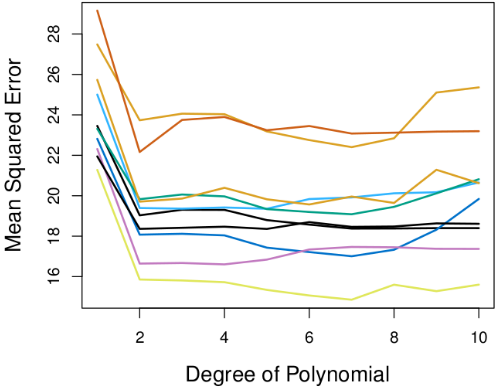
Predicting `mpg` using `horsepower`:

$$\texttt{mpg} = \beta_0 + \beta_1 \texttt{hp} + \beta_2 \texttt{hp}^2 + \cdots + \beta_p \texttt{hp}^p$$

Rinse and repeat



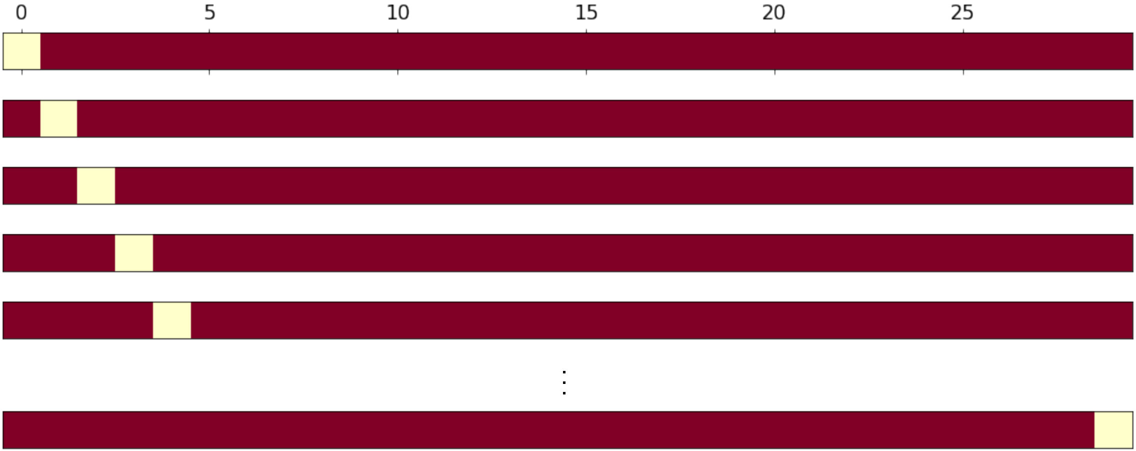
Again example with auto data



Section 2

Leave-One-Out Cross-Validation (LOOCV)

The idea



The idea in mathy words

- Remove (x_1, y_1) for testing.
- Train the model on $n-1$ points: $\{(x_2, y_2), \cdots, (x_n, y_n)\}$
- Calculate $\mathrm{MSE}_1 = (y_1 - \hat{y}_1^2)$

- Remove (x_2, y_2) for testing.
- Train the model on $n-1$ points: $\{(x_1, y_1), (x_3, y_3), \cdots, (x_n, y_n)\}$
- Calculate $\mathrm{MSE}_2 = (y_2 - \hat{y}_2^2)$

- Rinse and repeat

Return the score:

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

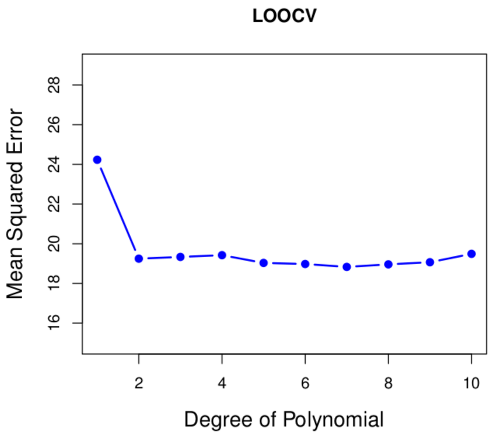
Do the LOOCV coding section

LOOCV Pros and Cons

Advantages:

Disadvantages:

Again example with auto data

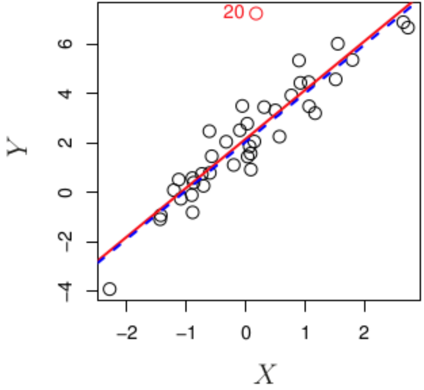


Section 3

The one time you can cheat (by not computing every model fit)

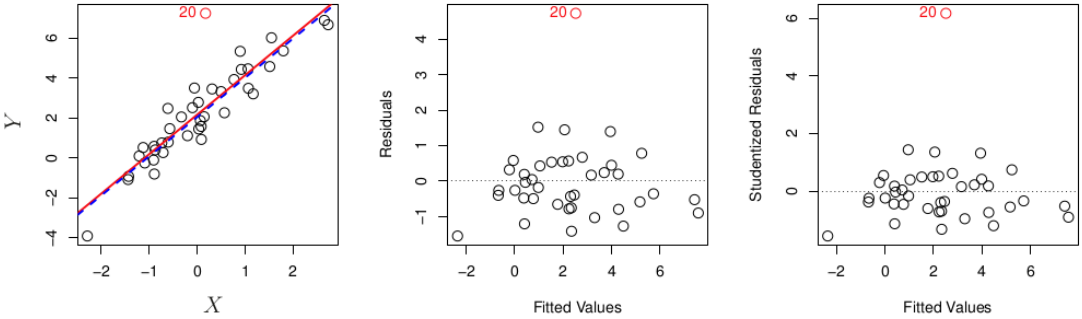
Outliers

An *outlier* is a point for which y_i is far from the value predicted by the model.



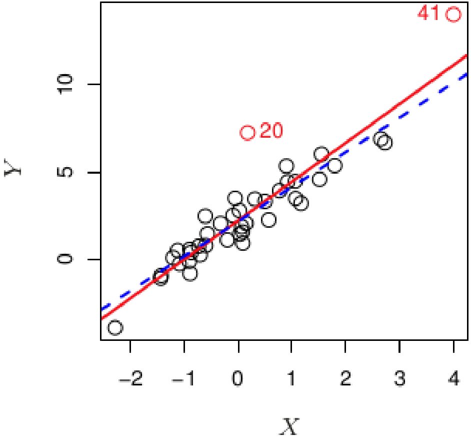
Residuals

Residuals



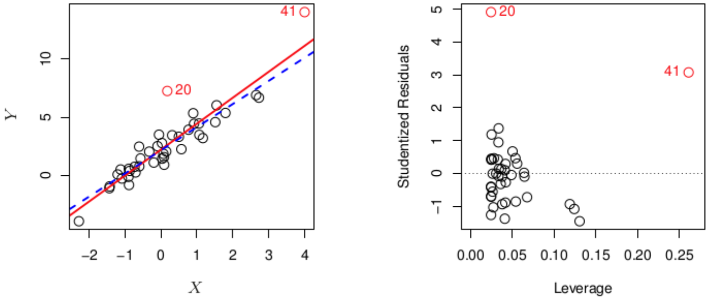
20
Y
X
−4
−2
0
2
4
6
−2
−1
0
1
2
20
Residuals
Fitted Values
−1
0
1
2
3
4
−2
0
2
4
6
20
Studentized Residuals
Fitted Values
0
2
4
6
−2
0
2
4
6

High Leverage



Observations with *high leverage* have an unusual value for x_i.

Leverage statistic



Version for $p = 1$

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}$$

Leverage statistic properties

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}$$

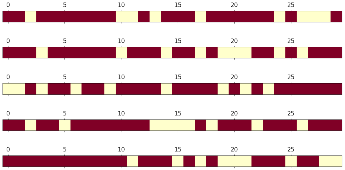
Speeding up LOOCV

Warning: This only works for least squares linear or polynomial regression.

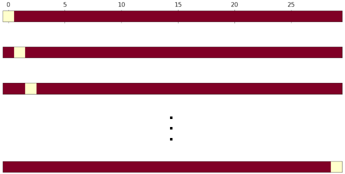
$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i = CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

TL;DR

Validation set



LOO-CV



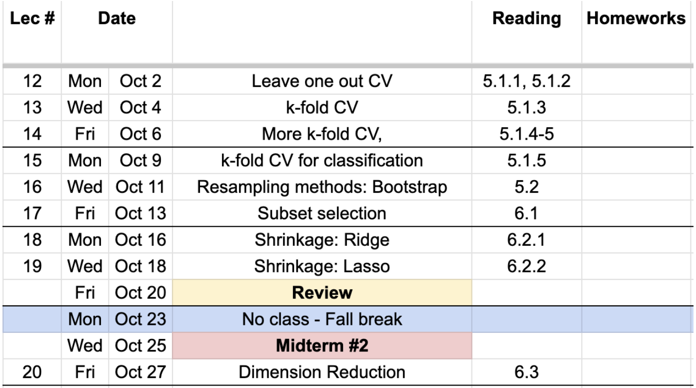
LOO-CV Score

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

Cheap trick for regression

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2$$

Next time



Lec #	Date			Reading	Homeworks
12	Mon	Oct 2	Leave one out CV	5.1.1, 5.1.2	
13	Wed	Oct 4	k-fold CV	5.1.3	
14	Fri	Oct 6	More k-fold CV,	5.1.4-5	
15	Mon	Oct 9	k-fold CV for classification	5.1.5	
16	Wed	Oct 11	Resampling methods: Bootstrap	5.2	
17	Fri	Oct 13	Subset selection	6.1	
18	Mon	Oct 16	Shrinkage: Ridge	6.2.1	
19	Wed	Oct 18	Shrinkage: Lasso	6.2.2	
	Fri	Oct 20	**Review**		
	Mon	Oct 23	No class - Fall break		
	Wed	Oct 25	**Midterm #2**		
20	Fri	Oct 27	Dimension Reduction	6.3	