Ch 5.1.1-3: Leave one out and k-Fold Cross-Validation
Lecture 13 - CMSE 381

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

Mon, Feb 16, 2026

Last time:

- Exam

Announcements:

- Exam 1 grades ....hopefully soon

11	F	2/6	Multiple Logistic Regression / Multinomial Logistic Regression	4.3.4-5	HW #2 Due Mon 2/9	
	M	2/9	*Project Day & Review*			
	W	2/11	**Midterm #1**			
12	F	2/13	**Class not held**			
13	M	2/16	Leave one out and k-fold CV	5.1.1-3		
14	W	2/18	More k-fold CV	5.1.4-5		Q5
15	F	2/20	k-fold CV for classification	5.1.5		
16	M	2/23	Subset selection	6.1		
17	W	2/25	Shrinkage: Ridge	6.2.1		
18	F	2/27	Shrinkage: Lasso	6.2.2	HW #3 Due Sun 3/1	
	M	3/2	Spring Break			
	W	3/4	Spring Break			
	F	3/6	Spring Break			
19	M	3/9	PCA	6.3		
20	W	3/11	PCR	6.3		Q6

Covered in this lecture

- Validation Set
- LOO CV
- K-fold CV

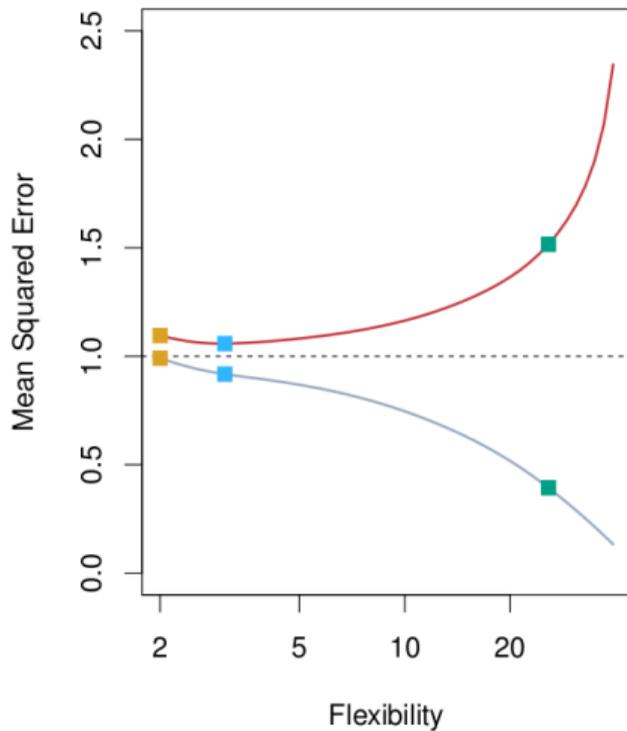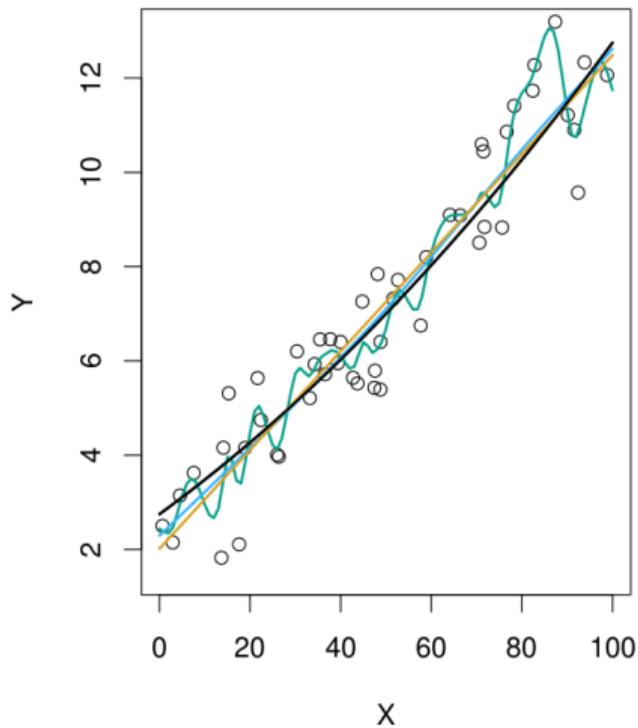# Section 1

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?
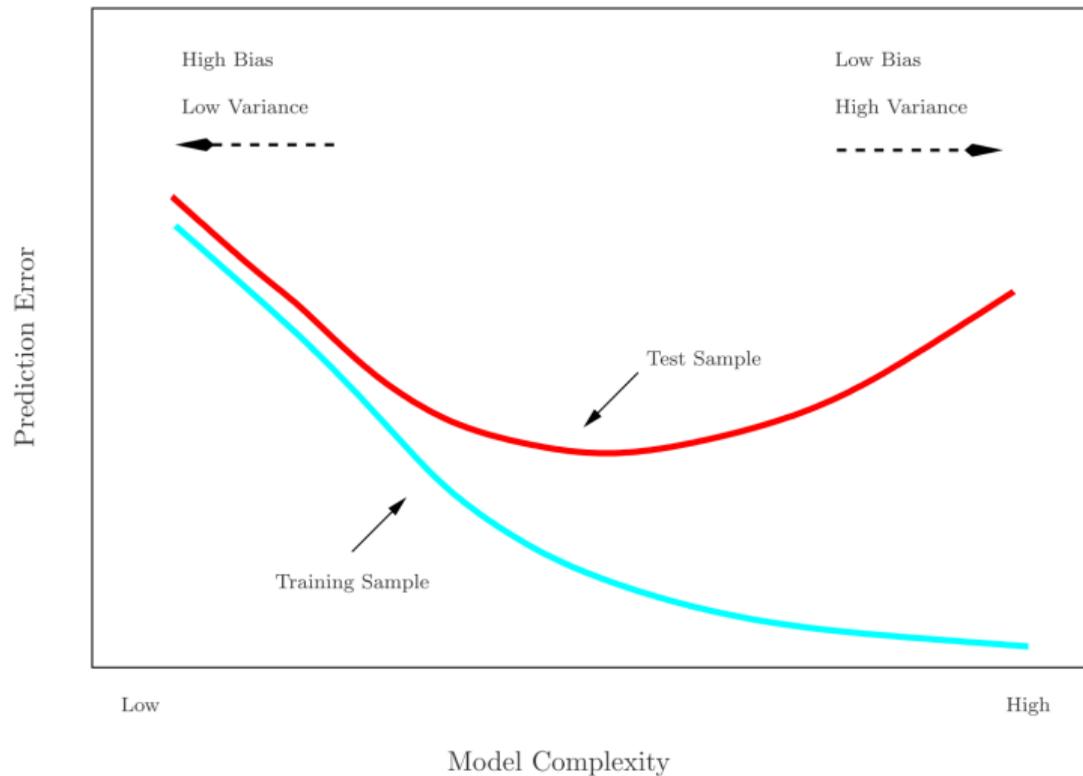
Training Error vs Testing Error
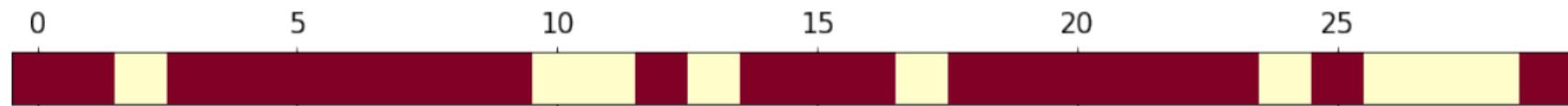
Training Error

Testing Error

Throw-back

Model tradeoffs

High Bias
Low Variance

Low Bias
High Variance

Prediction Error

Test Sample

Training Sample

Low
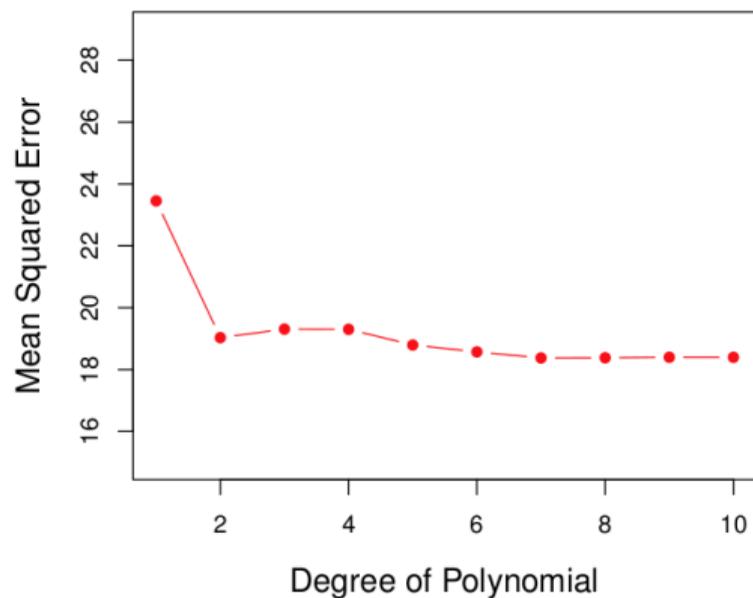
High

Model Complexity

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
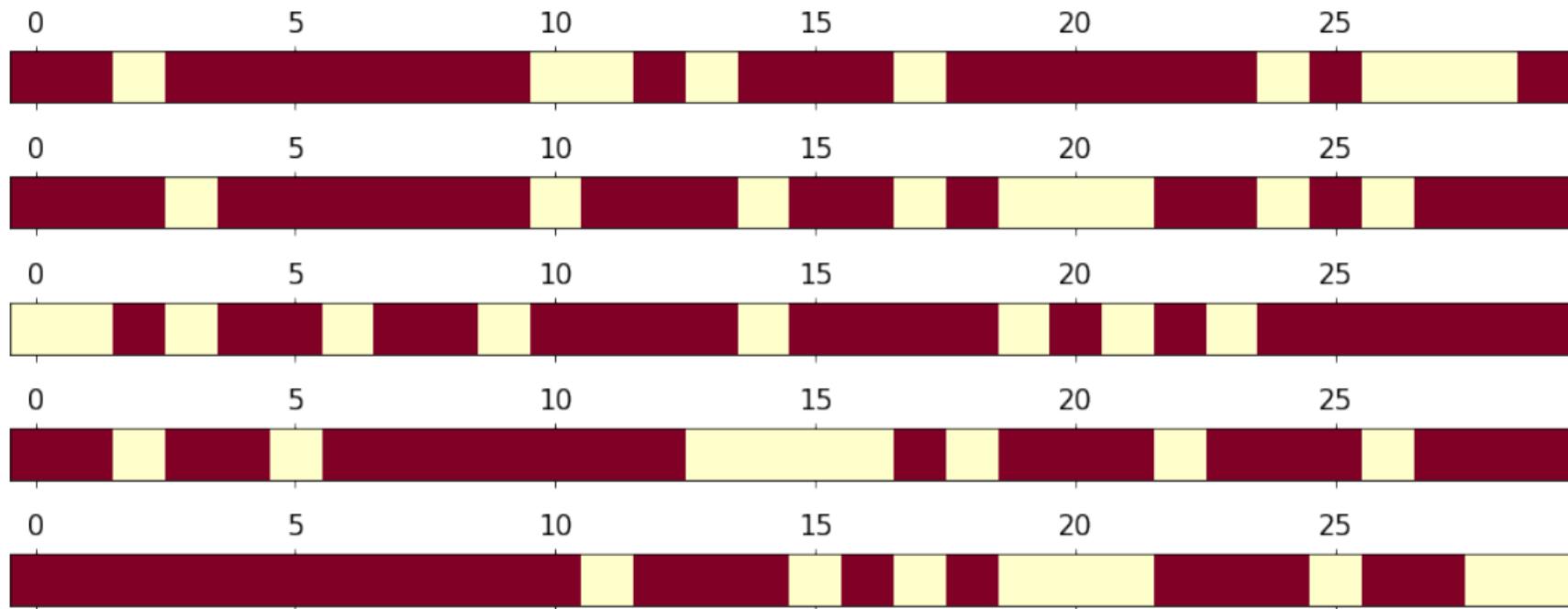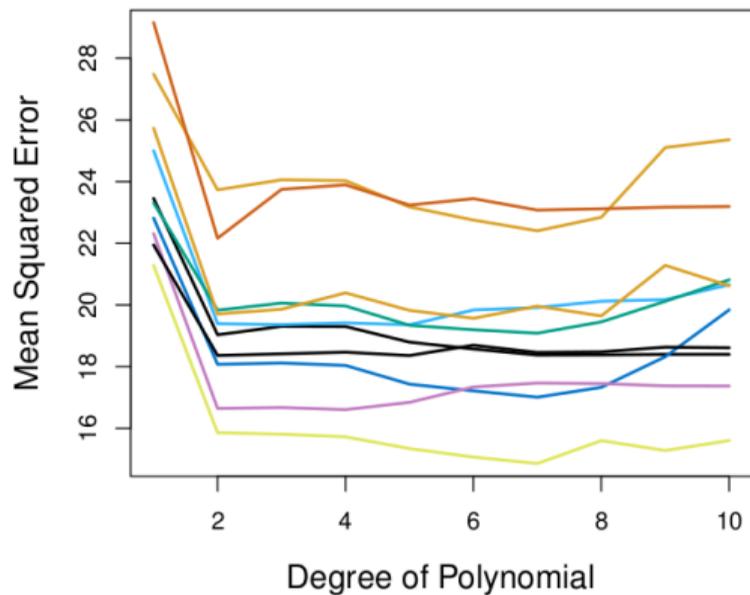- Evaluate quality (e.g. MSE)

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
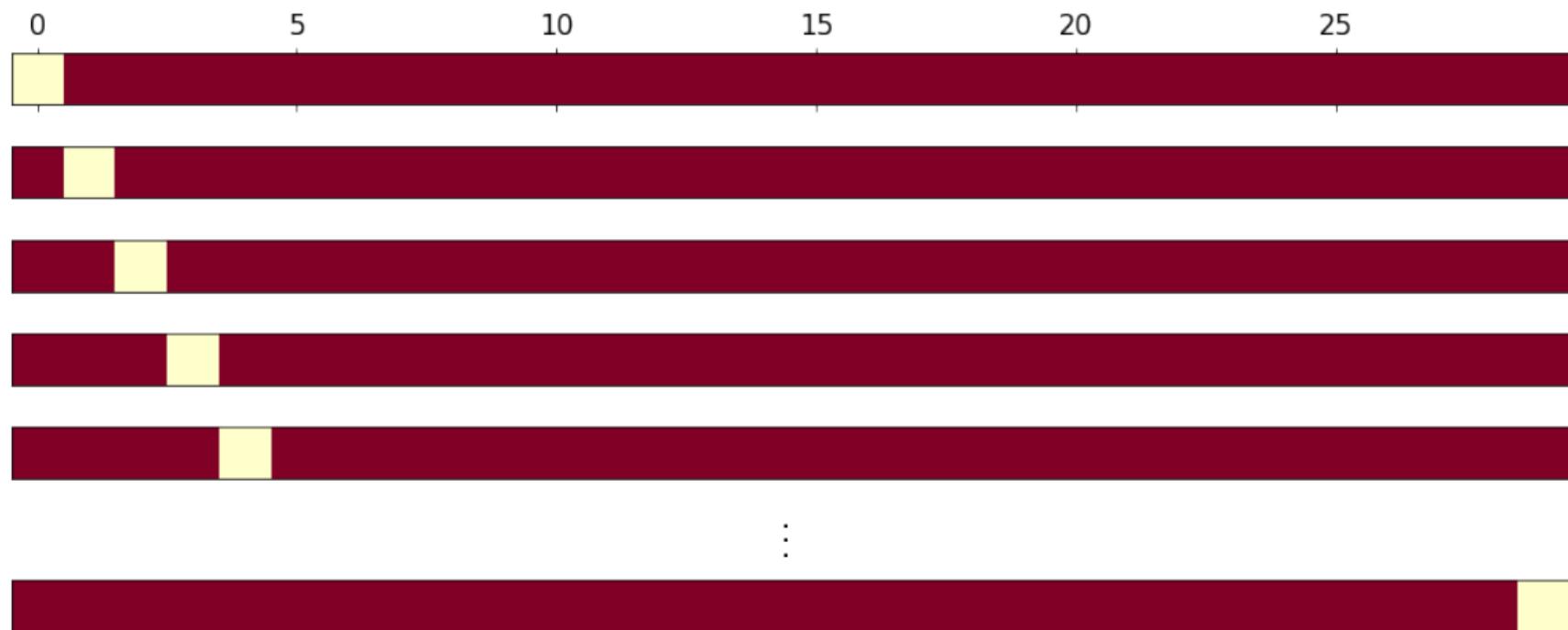
Again example with `auto` data

Coding example in jupyter notebook

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: $\{(x_2, y_2), \cdots, (x_n, y_n)\}$
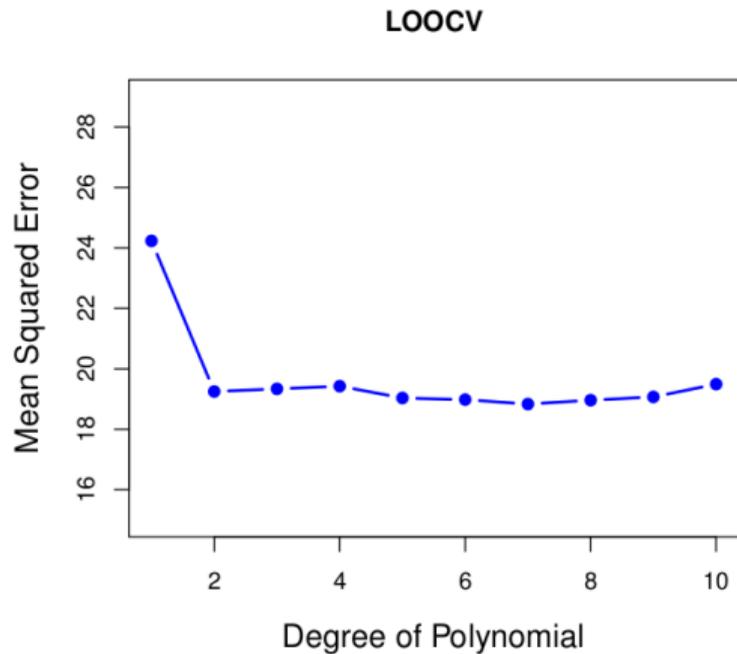- Calculate $\mathrm{MSE}_1 = (y_1 - \hat{y}_1)^2$

- 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:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{MSE}_i$$
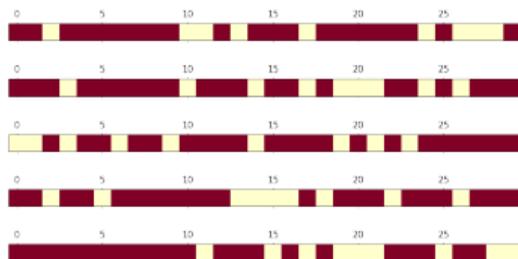
Again example with `auto` data

LOOCV

Mean Squared Error vs. Degree of Polynomial

Do the LOOCV coding section

LOOCV Pros and Cons

Advantages:

Disadvantages:

TL;DR
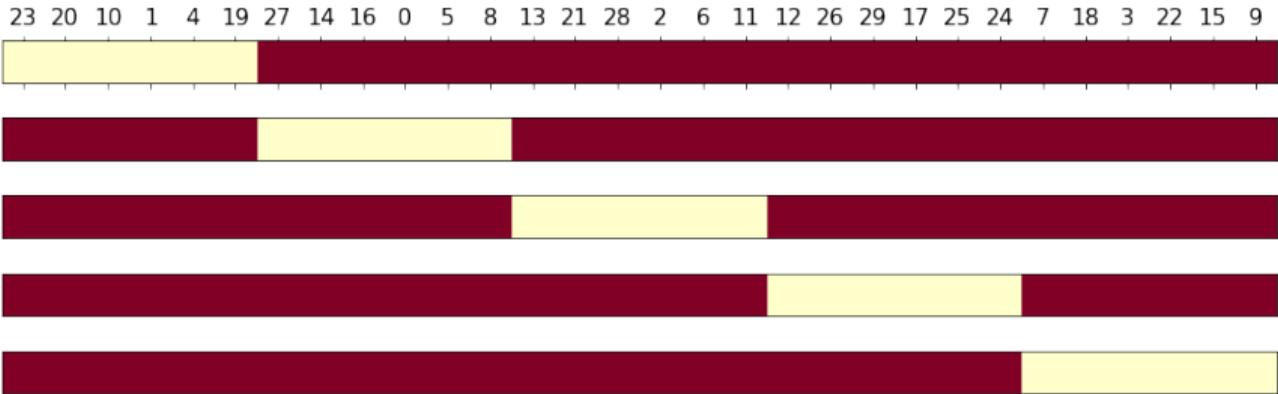
Validation set

LOO-CV

⋮
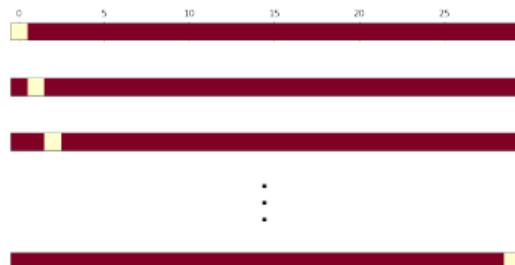
LOO-CV Score

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

Section 3

k-Fold CV

The idea

Mathy version

- Randomly split data into k-groups (folds)
- Approximately equal sized. For the sake of notation, say each set has ℓ points

- Remove ith fold U_i and reserve for testing.
- Train the model on remaining points
- Calculate
  $\mathrm{MSE}_i = \frac{1}{\ell} \sum_{(x_j, y_j) \in U_i} (y_j - \hat{y}_j)^2$

- Rinse and repeat

Return

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

By hand first!

There are 10 students in the class, and we have data points for each. They have already been randomly permuted below. Write down the training/testing sets for a 3-fold CV

	Fold 1	**Fold 2**	**Fold 3**
Damien			
Alice			
Greta			
Jasmin			
Benji			
Inigo			
Frank			
Carina			
Enrique			
Hubert			

Coding - Building k-fold CV

Pros and Cons

Pros: **Cons:**

Comparison
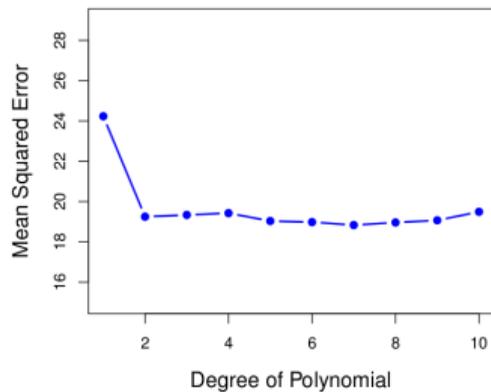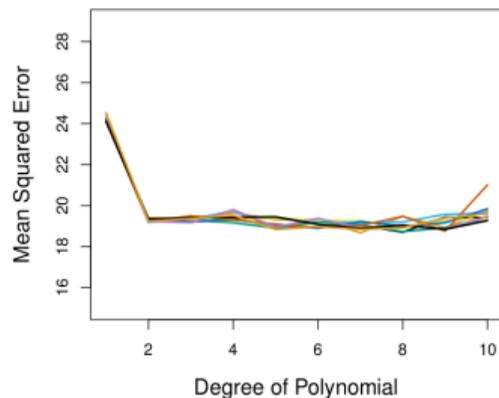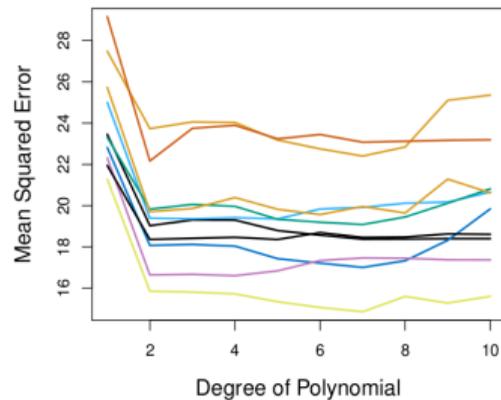
Next time

11	F	2/6	Multiple Logistic Regression / Multinomial Logistic Regression	4.3.4-5	HW #2 Due Mon 2/9	
	M	2/9	*Project Day & Review*			
	W	2/11	**Midterm #1**			
12	F	2/13	**Class not held**			
13	M	2/16	Leave one out and k-fold CV	5.1.1-3		
14	W	2/18	More k-fold CV	5.1.4-5		Q5
15	F	2/20	k-fold CV for classification	5.1.5		
16	M	2/23	Subset selection	6.1		
17	W	2/25	Shrinkage: Ridge	6.2.1		
18	F	2/27	Shrinkage: Lasso	6.2.2	HW #3 Due Sun 3/1	
	M	3/2	Spring Break			
	W	3/4	Spring Break			
	F	3/6	Spring Break			
19	M	3/9	PCA	6.3		
20	W	3/11	PCR	6.3		Q6