RESAMPLING METHODS

Validation

Bootstrap

AGENDA

➢The Validation Set Approach

➢Leave-One-Out Cross Validation (LOOCV)

➢K-fold Cross Validation (CV)

➢The Bootstrap

WHAT ARE RESAMPLING METHODS?

- How to choose an appropriate model?

- How to discover overfitting?

- How to estimate the test MSE?
    - Splitting (resampling) the entire set into training and validation sets
    - Using training set for model construction
    - Using validation set for model performance evaluation (test MSE estimate) and flexibility selection

- Performance evaluation is known as model assessment

- Flexibility selection is known as model selection

VALIDATION SET APPROACH

4

THE MAIN IDEA
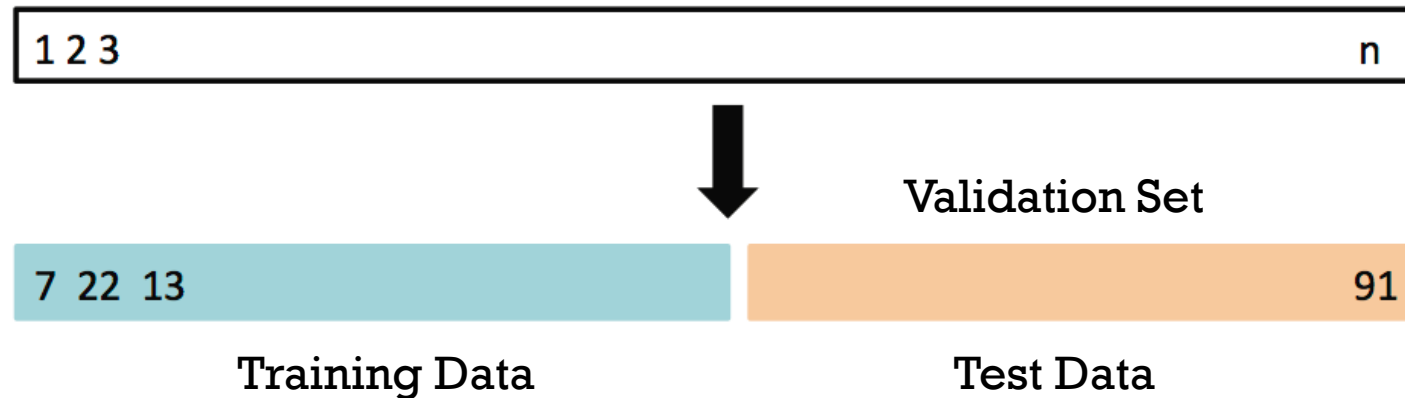
- Randomly divide the available set of observations into two parts – training set and validation set

- Then, use the training part to build the model and validation part for test error estimate

EXAMPLE: AUTO DATA

- In R
  - use library(ISLR)
  - Execute edit(Auto)

Data Editor

File   Edit   Help

	row.names	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
1	1	18	8	307	130	3504	12	70	1
2	2	15	8	350	165	3693	11.5	70	1
3	3	18	8	318	150	3436	11	70	1
4	4	16	8	304	150	3433	12	70	1
5	5	17	8	302	140	3449	10.5	70	1
6	6	15	8	429	198	4341	10	70	1
7	7	14	8	454	220	4354	9	70	1
8	8	14	8	440	215	4312	8.5	70	1
9	9	14	8	455	225	4425	10	70	1
10	10	15	8	390	190	3850	8.5	70	1
11	11	15	8	383	170	3563	10	70	1
12	12	14	8	340	160	3609	8	70	1
13	13	15	8	400	150	3761	9.5	70	1
14	14	14	8	455	225	3086	10	70	1
15	15	24	4	113	95	2372	15	70	3
16	16	22	6	198	95	2833	15.5	70	1
17	17	18	6	199	97	2774	15.5	70	1
18	18	21	6	200	85	2587	16	70	1
19	19	27	4	97	88	2130	14.5	70	3
20	20	26	4	97	46	1835	20.5	70	2
21	21	25	4	110	87	2672	17.5	70	2
22	22	24	4	107	90	2430	14.5	70	2
23	23	25	4	104	95	2375	17.5	70	2
24	24	26	4	121	113	2234	12.5	70	2

# EXAMPLE: AUTO DATA

- The goal is prediction of mpg by horsepower

- Different models:
  - $mpg \sim horsepower$
  - $mpg \sim horsepower + horsepower^2$
  - $mpg \sim horsepower + horsepower^2 + horsepower^3$
  - $mpg \sim horsepower + horsepower^2 + horsepower^3 + \cdots$

EXAMPLE: AUTO DATA

- Which model gives a better fit?
  - Randomly split data set into training (196 obs.) and validation data (196 obs.)
  - Fit all models using the training data set
  - Evaluate all models using the validation data set
  - The model with the lowest validation (test) MSE will be the winner!

RESULTS: AUTO DATA
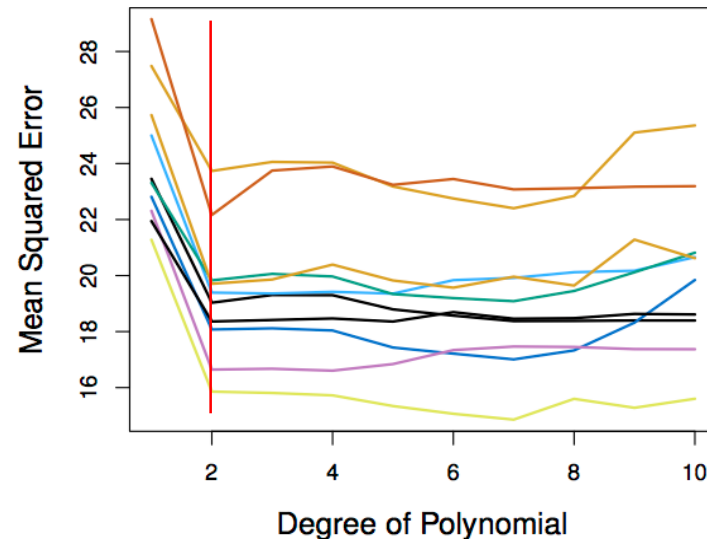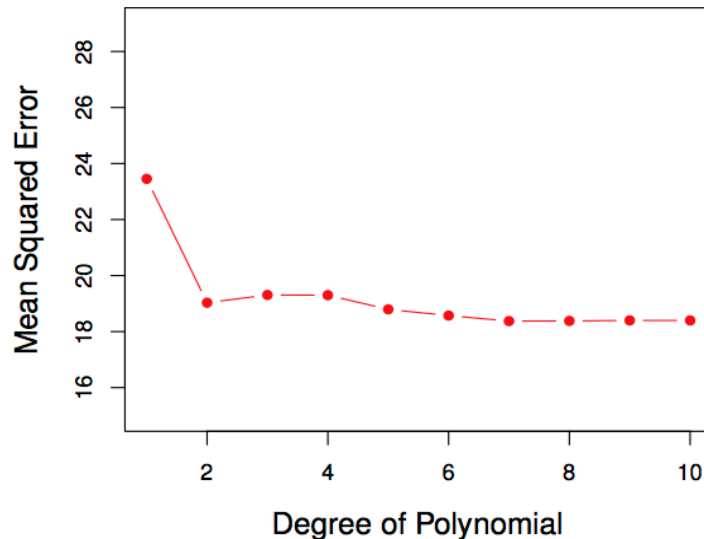
- Left: Validation error rate for a single split

- Right: Validation method repeated 10 times, each time the split is done randomly

- Impossible to estimate test MSE accurately. Possible to select the best flexibility. We need more stable approaches

THE VALIDATION SET APPROACH

- Advantages:
  - Simple
  - Easy to implement

- Disadvantages:
  - The validation MSE is not stable
  - Only a subset of observations are used to fit the model (training data). Statistical methods tend to perform worse when trained on fewer observations

LOOCV

11

LOOCV

- This method is similar to the Validation Set Approach, but it tries to address the latter's disadvantages

- Split the data set of size n into
  - Training data set (blue) size: $n - 1$
  - Test data set (beige) size: 1
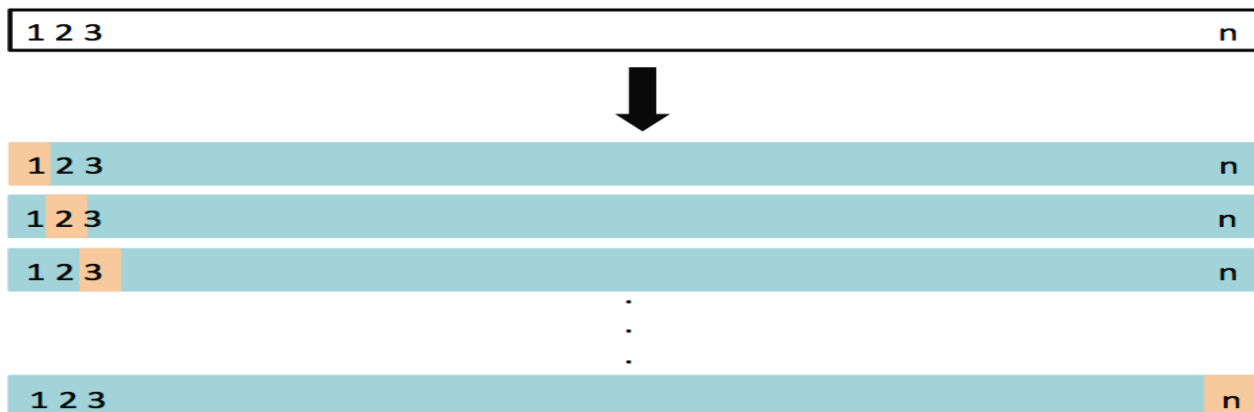
LOOCV

- Fit the model using the training data without the ith point

- Compute the test-MSE MSE_i for the ith point

- Repeat this process n times

- The MSE for the model is:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} MSE_i$$

EXAMPLE: AUTO DATA
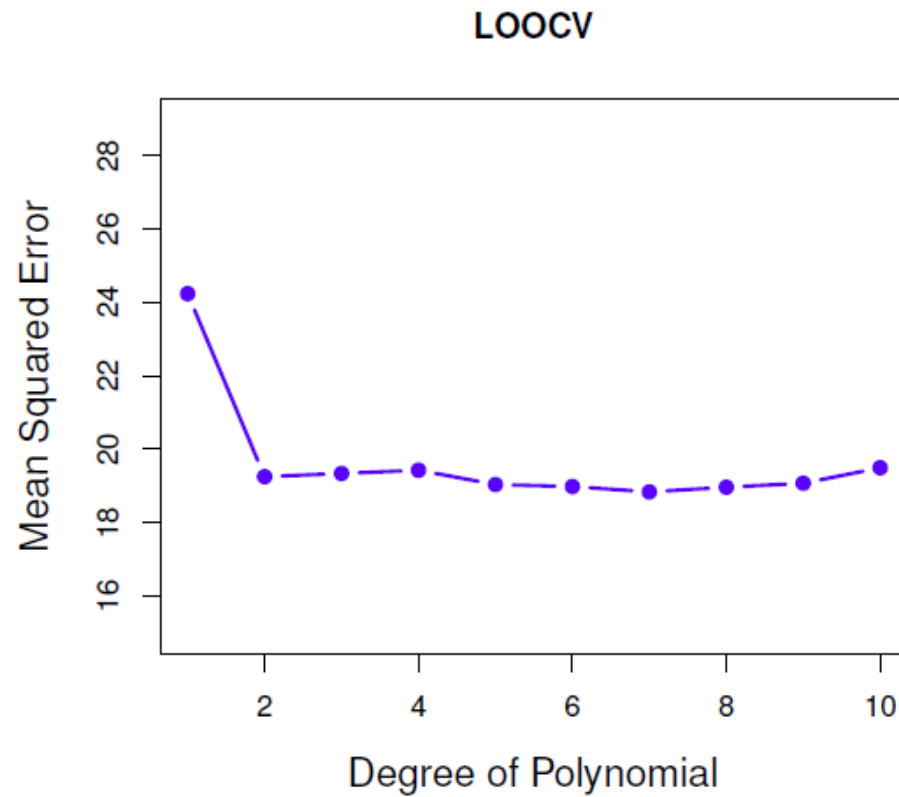
LOOCV: ADVANTAGES

- We train a statistical learning method on $n-1$ obs., i.e. almost entire dataset is used and it gives more accurate test MSE estimate

- The validation approach produces different MSEs when applied repeatedly due to randomness in the splitting process, while LOOCV will always yield the same results. It produces more stable MSE estimate

LOOCV: DRAWBACK

- LOOCV is computationally intensive.

- We fit each model n times!

LOOCV: POLYNOMIAL LEAST SQUARES

- With linear or polynomial least squares, we have an amazing formula

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2, \qquad h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

- Makes the cost of LOOCV the same as that of a single model fit!

K-FOLD CV

18

K-FOLD CV

- An alternative to LOOCV is k-fold CV

- We divide randomly the data set into k different parts (e.g. $k = 5$, or $k = 10$, etc.)
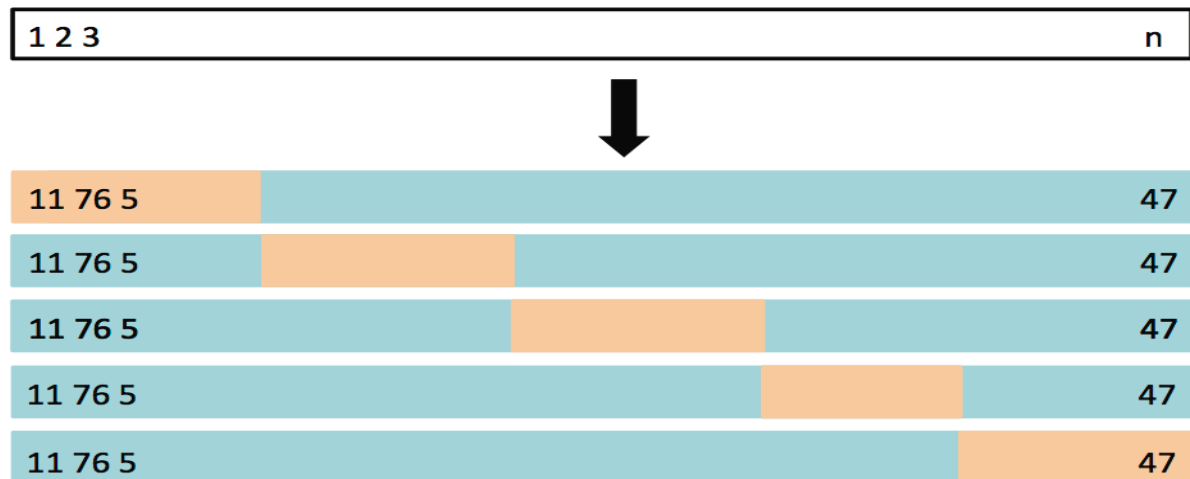
K-FOLD CV

- We then remove the first part, fit the model on the remaining $k - 1$ parts, and see how good the predictions are on the left out part (i.e. compute the MSE_1 on the first part)

- We then repeat this k different times taking out a different part each time by calculating each time the corresponding $MSE_j, j = 1, \ldots, k$

- By averaging the k different MSE's we get an estimated validation (test) error rate for new observations
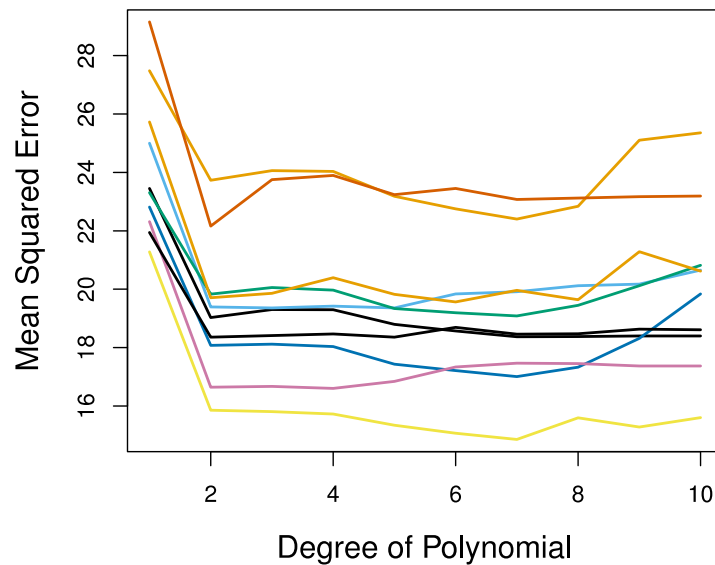
$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_k$$

- LOOCV is a special case of k-fold CV in which $k = n$
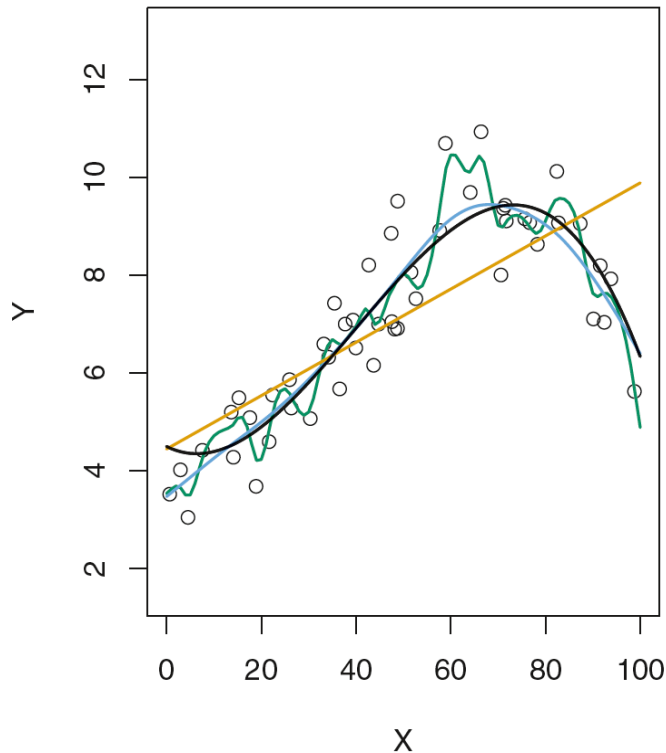
EXAMPLE: AUTO DATA
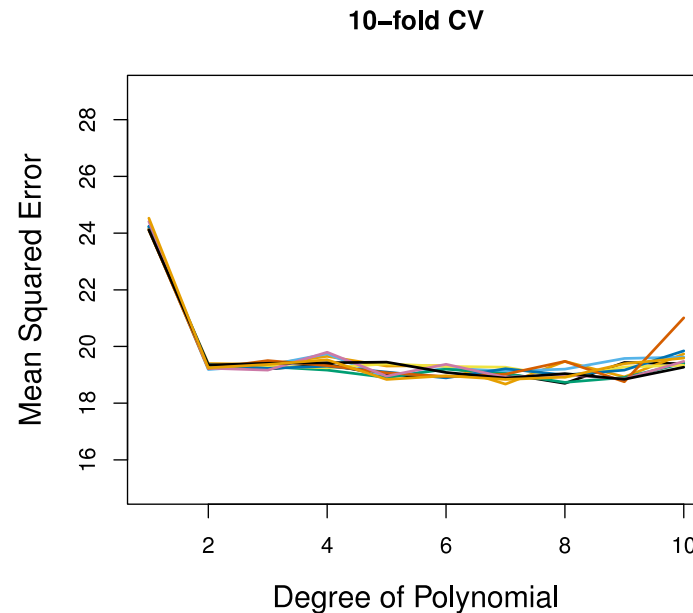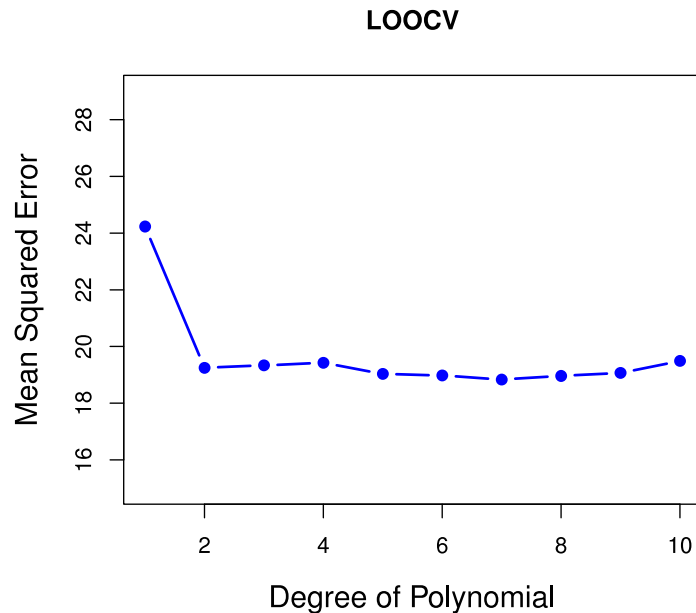
LOOCV

10−fold CV

- Left: LOOCV error curve

- Right: 10-fold CV was run many times, and the figure shows the slightly different CV error rates

- They are both stable, but LOOCV is more computationally intensive!
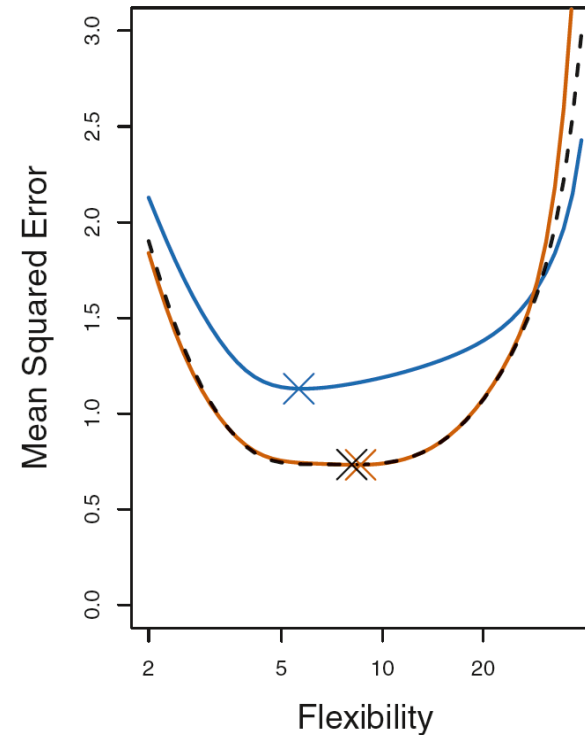
EXAMPLE: AUTO DATA
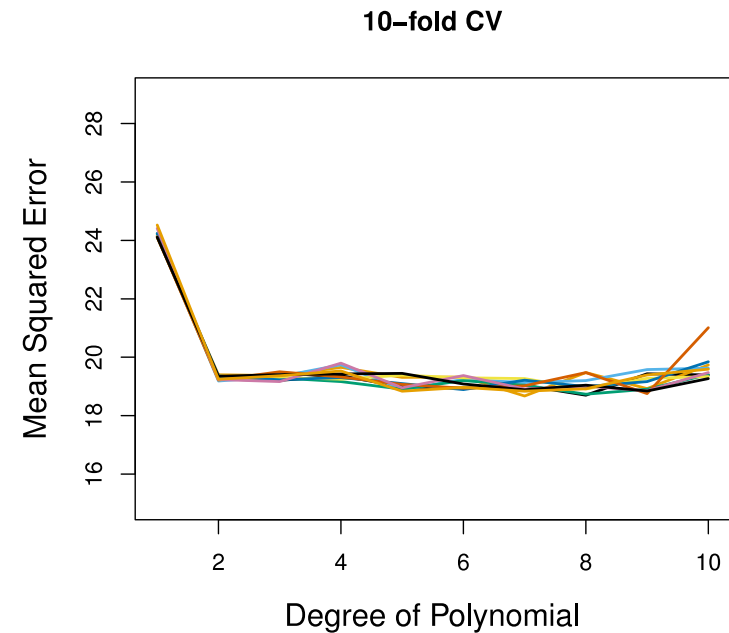
- Left: Validation Set Approach
- Right: 10-fold CV
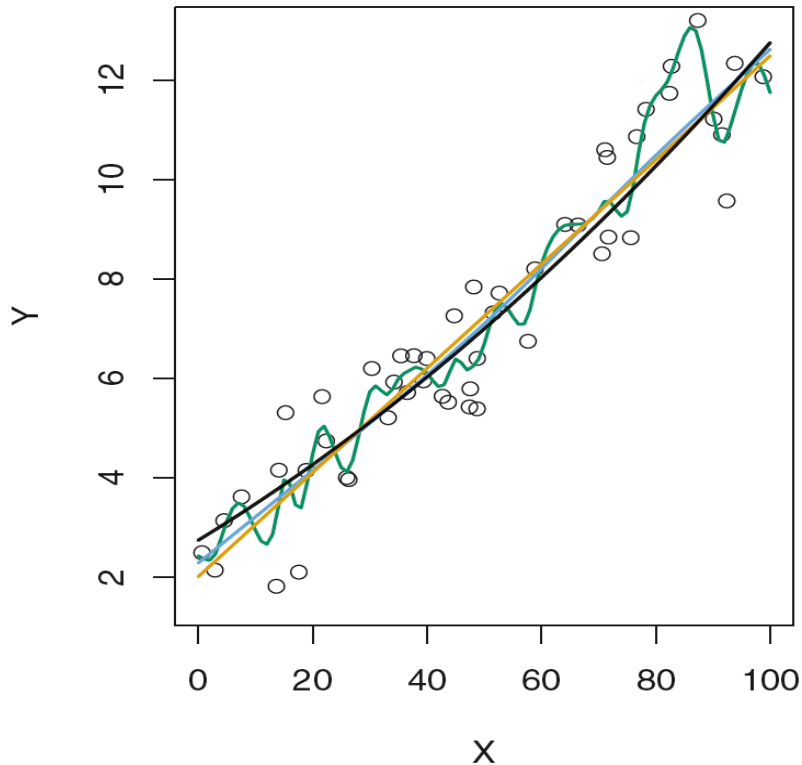- 10-fold CV is much more stable!

EXAMPLE: K-FOLD CV

Orange: linear regression
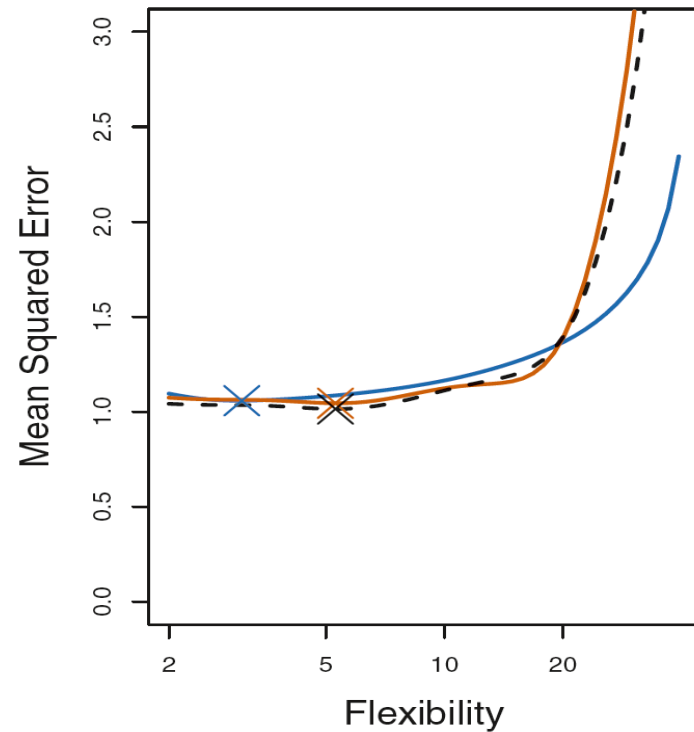Black: original data
Blue, Green: splines

Blue: True Test MSE
Black: LOOCV MSE
Orange: 10-fold MSE

23

EXAMPLE: K-FOLD CV

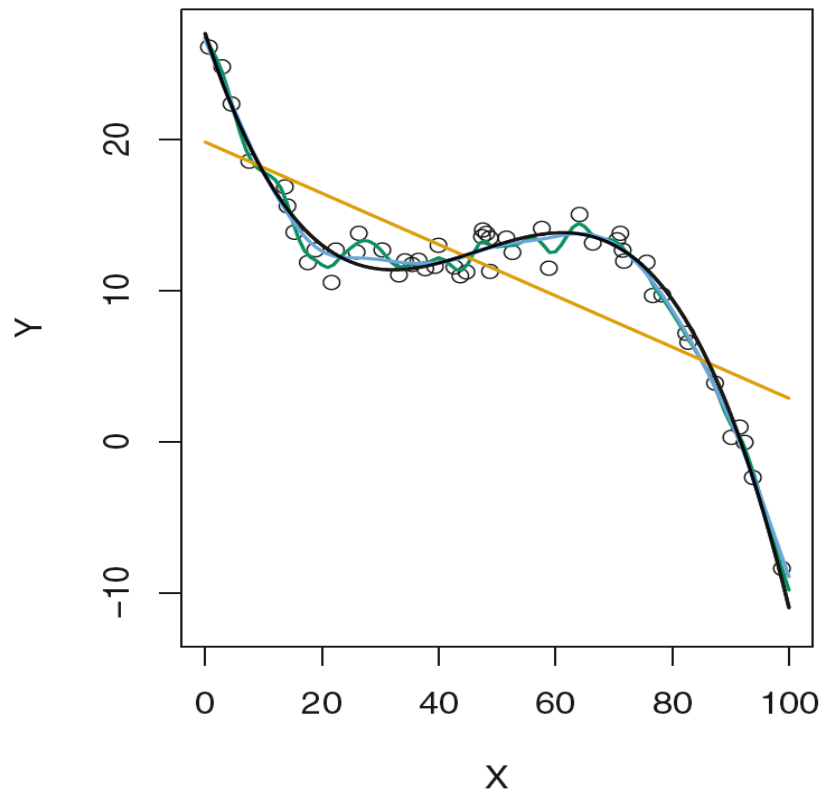Orange: linear regression
Black: original data
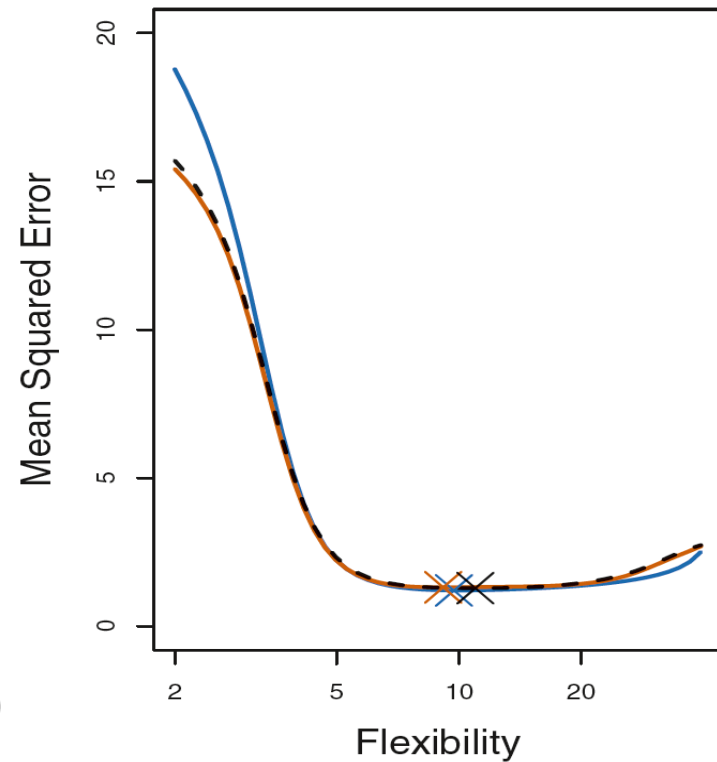Blue, Green: splines

Blue: True Test MSE
Black: LOOCV MSE
Orange: 10-fold MSE

24

EXAMPLE: K-FOLD CV

Orange: linear regression
Black: original data
Blue, Green: splines

Blue: True Test MSE
Black: LOOCV MSE
Orange: 10-fold MSE

25

COMPARISON

- Complexity:
  - LOOCV is more complicated than k-fold CV ($k < n$)
  - k-fold CV (k < n) is more complicated than the validation set approach

- Variance:
  - LOOCV has higher variance than k-fold CV (when $k < n$)
  - k-fold CV (k < n) has higher variance than the validation set approach

COMPARISON

- Bias:
  - LOOCV will give almost unbiased estimates of the test error as it uses almost the entire dataset for training
  - k-fold CV (k < n) has intermediate level of bias
  - The validation set approach is biased

- Conclusion:
  - We tend to use k-fold CV that has bias-variance trade-off
  - It has been empirically shown that $k = 5$ or $k = 10$ yield test error rate estimates that suffer neither from excessively high bias, nor from very high variance

THE BOOTSTRAP

28

POPULATION AND SAMPLE

Population mean is μ. Let $\mu_1, \mu_2, \ldots, \mu_m$ are the averages of the m different samples. Then,
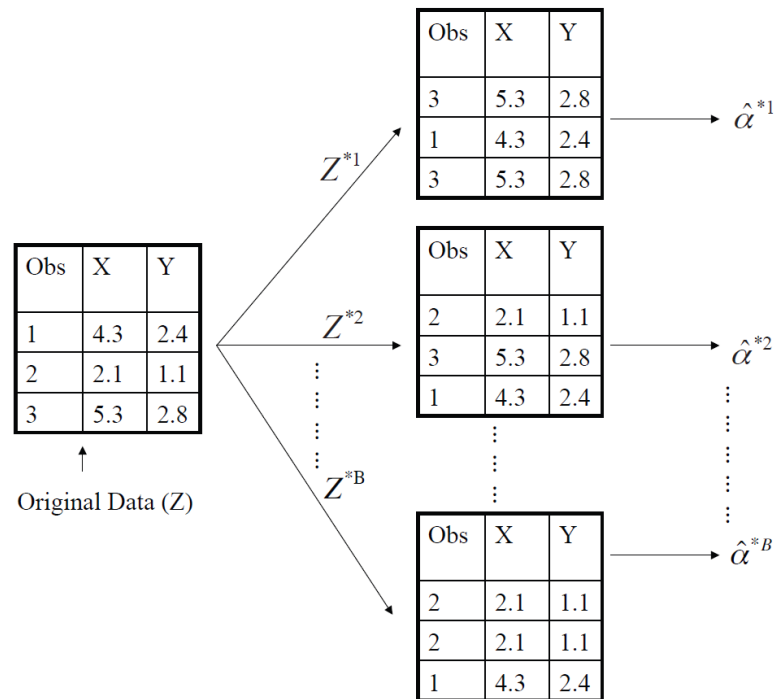
$$-z * \frac{\sigma}{\sqrt{m}} \leq \frac{\mu_1 + \cdots + \mu_m}{m} - \mu \leq z * \frac{\sigma}{\sqrt{m}}$$

where σ is the sd of the population and $z = 1,2,3$

THE BOOTSTRAP

- The bootstrapping allows us to use a computer to mimic the process of obtaining new data sets, so that we can estimate the variability of our estimate without generating additional samples

- Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set with replacement

- Each of these bootstrap data sets is created by sampling with replacement, and is the same size as our original dataset. As a result some observations may appear **more than once** in a given bootstrap data set and some **not at all**

EXAMPLE WITH 3 OBSERVATIONS

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

Z^{*1} $\longrightarrow \hat{\alpha}^{*1}$

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Original Data (Z)

Z^{*2}

Obs	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

$\longrightarrow \hat{\alpha}^{*2}$

Z^{*B}

Obs	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

$\longrightarrow \hat{\alpha}^{*B}$

- A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations. Each bootstrap data set contains n observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of α.

EXAMPLE CONTINUED

- For each bootstrap set, we calculate estimates for α

$$\hat{\alpha}_1^*, \dots, \hat{\alpha}_B^*$$

- Then we calculate the average

$$\bar{\alpha} = \frac{1}{B} \sum_{j=1}^{B} \hat{\alpha}_j^*$$

- Then we calculate the standard error of the bootstrap estimates

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1} \sum_{r=1}^{B} (\hat{\alpha}_r^* - \bar{\alpha})^2}$$

BOOTSTRAP AGGREGATION - BAGGING

- **Bagging** is designed to improve the stability and accuracy of an algorithm

- Bagging generates m new training sets by sampling with replacement

- This kind of sample is known as a bootstrap sample

- Then, the m models are fit using the above m bootstrap samples and combined by averaging the output (for regression) or voting (for classification)