RESAMPLING METHODS

Validation

Bootstrap

AGENDA

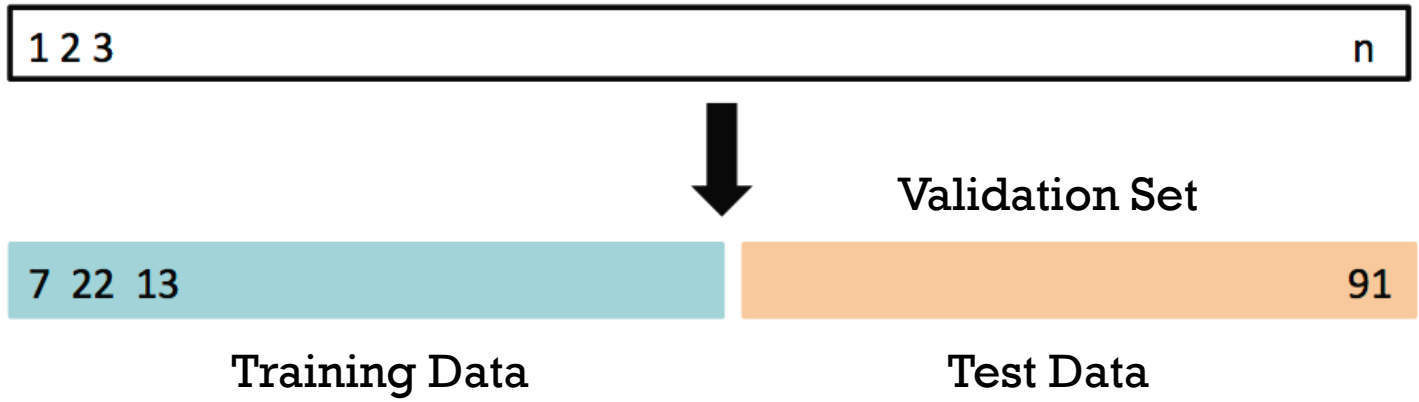
- The Validation Set Approach
- Leave-One-Out Cross Validation (LOOCV)
- K-fold Cross Validation (CV)
- The Bootstrap

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

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

AUTO DATA

- Gas mileage, horsepower, and other information for 392 vehicles.

`mpg`
: miles per gallon

`cylinders`
: Number of cylinders between 4 and 8

`displacement`
: Engine displacement (cu. inches)

`horsepower`
: Engine horsepower

`weight`
: Vehicle weight (lbs.)

`acceleration`
: Time to accelerate from 0 to 60 mph (sec.)

`year`
: Model year (modulo 100)

`origin`
: Origin of car (1. American, 2. European, 3. Japanese)

`name`
: Vehicle name

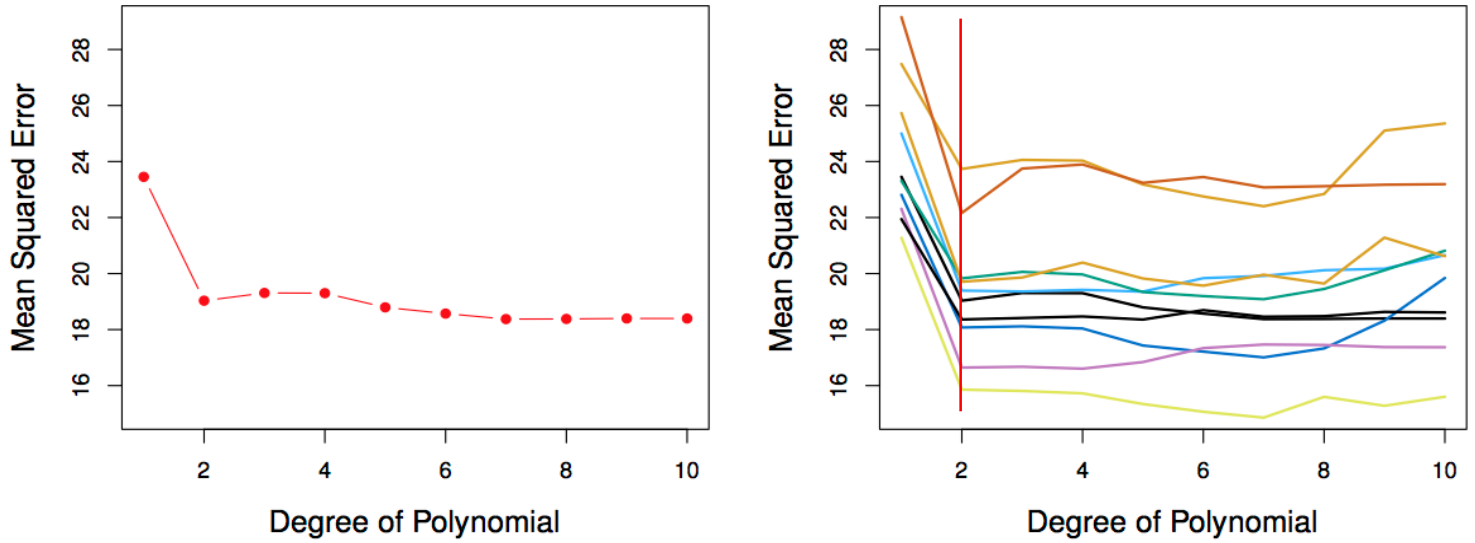
EXAMPLE: AUTO DATA

- 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

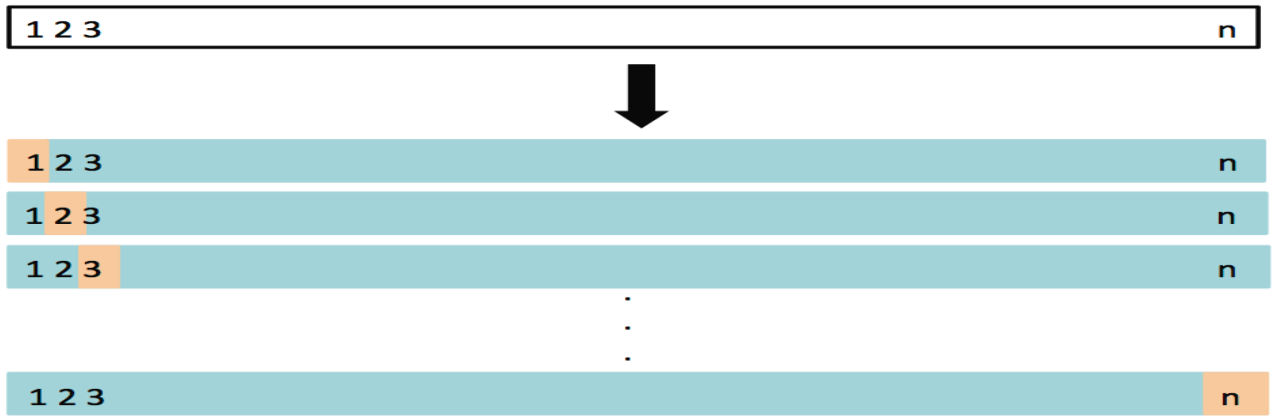
- Advantages:
  - Simple
  - Easy to implement

- 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

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

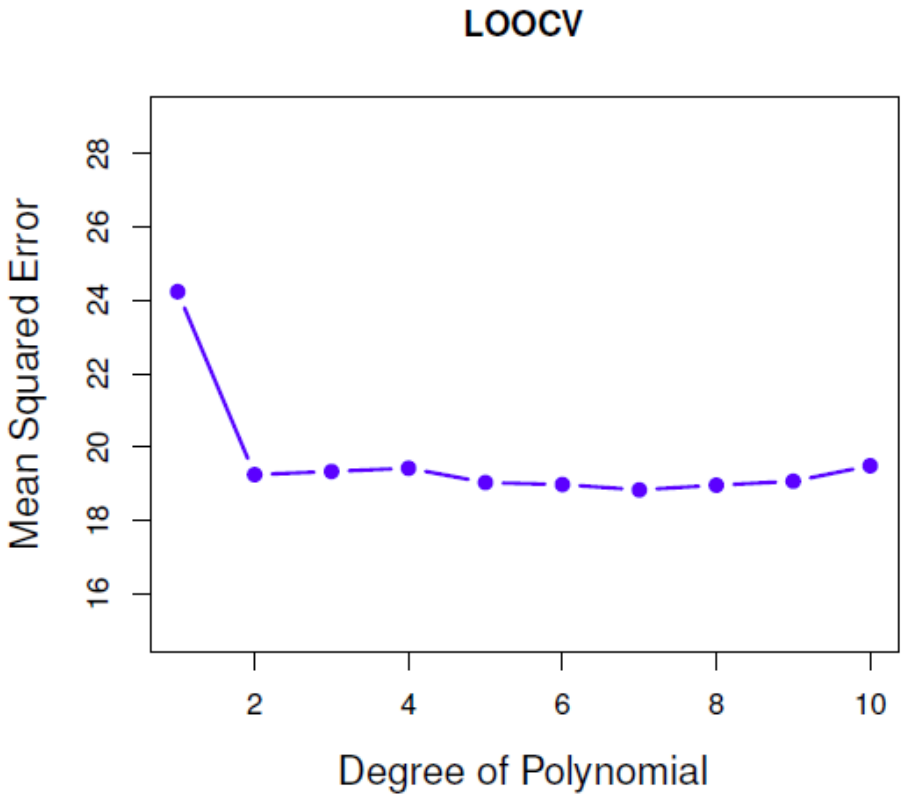


LOOCV

- Fit the model using the training data without the i^{th} point
- Compute the test-MSE MSE_i for the i^{th} 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

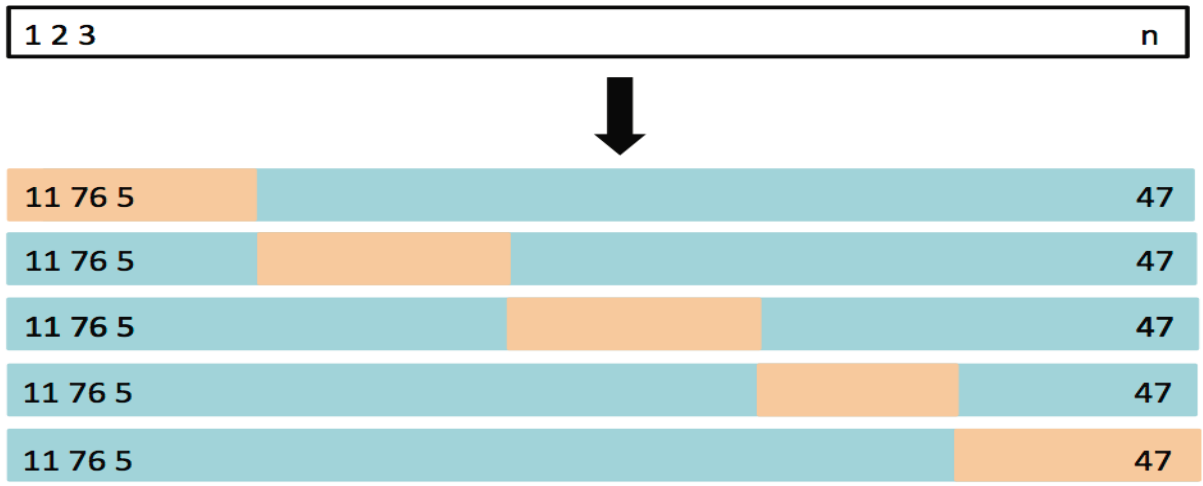
$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2, \qquad h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

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

K-FOLD CV

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.)



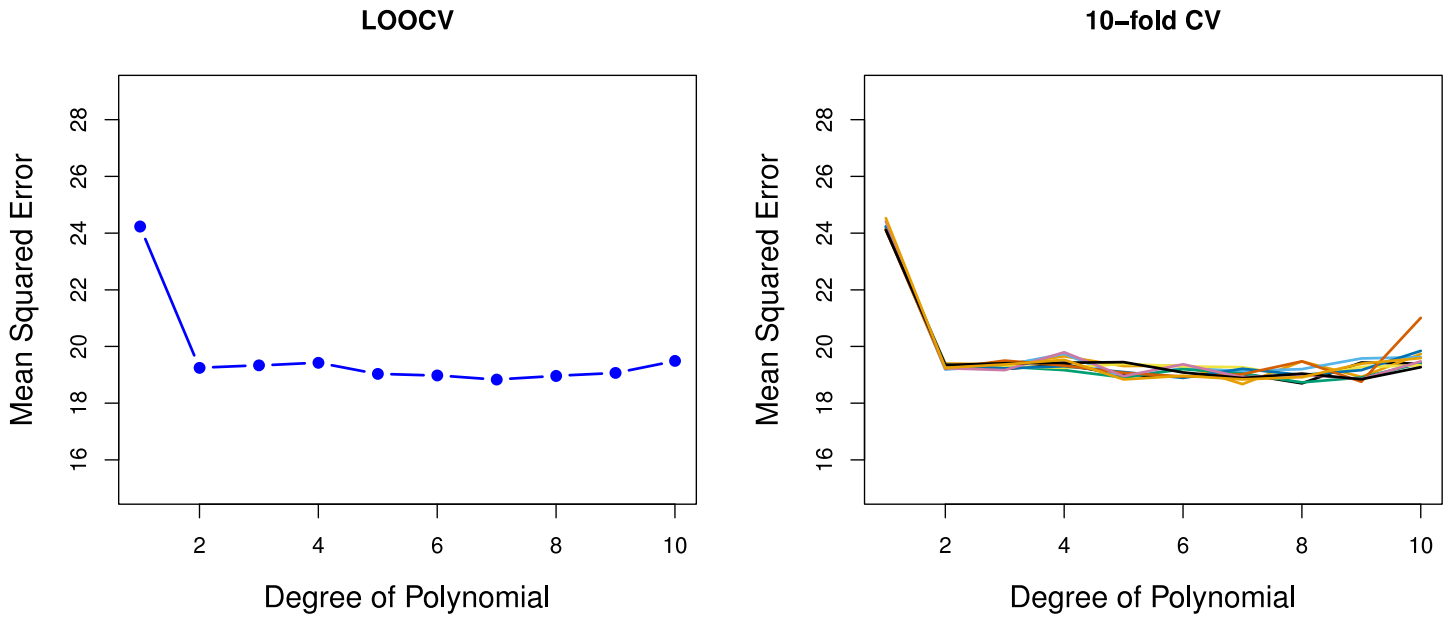
K-FOLD CV

- 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 $MSE_j, j = 1, \dots, k$
- By averaging the k different MSE's we get an estimated validation (test) error rate for new observations

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_k$$

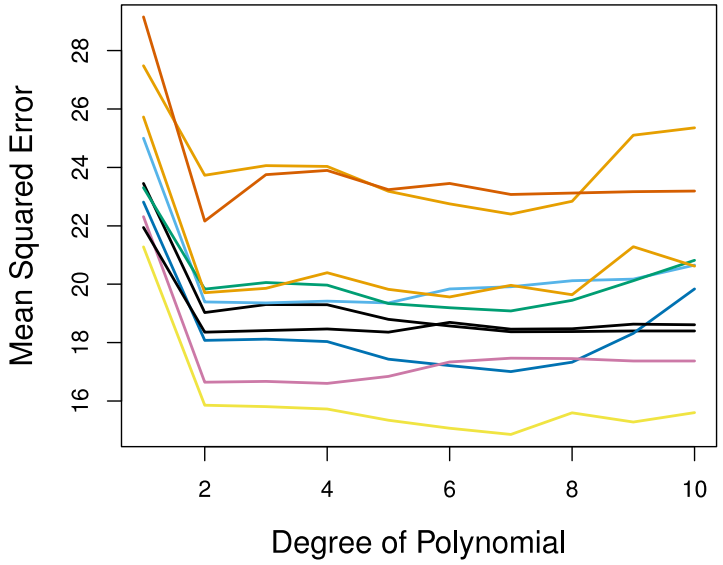
- LOOCV is a special case of k-fold CV in which $k = n$

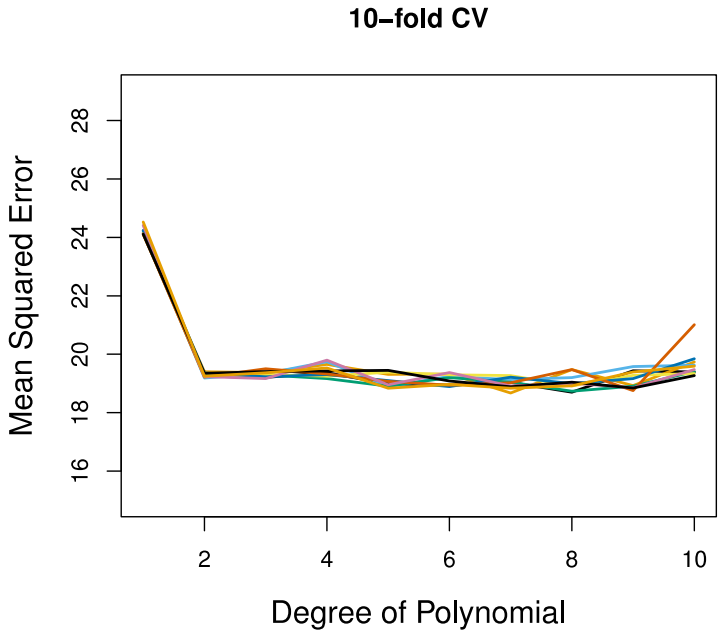
EXAMPLE: AUTO DATA



- 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!

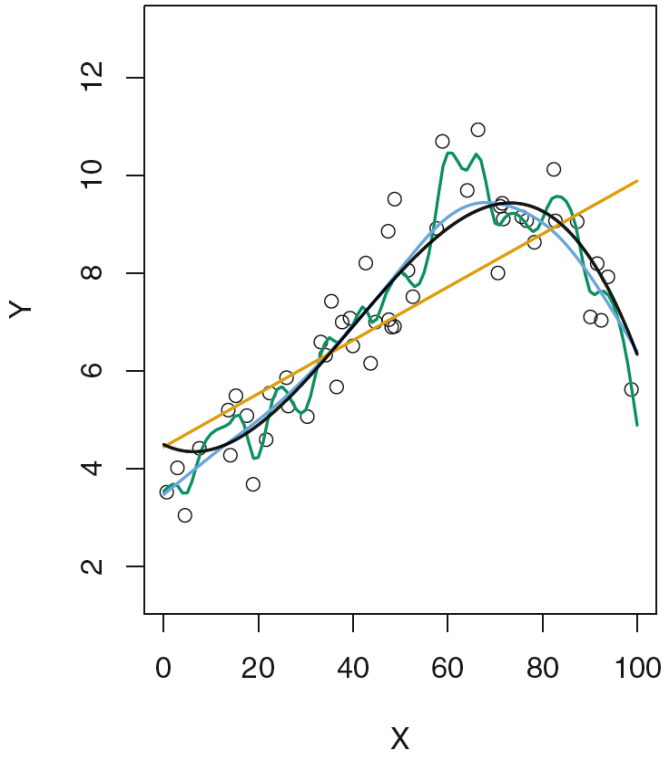
EXAMPLE: AUTO DATA



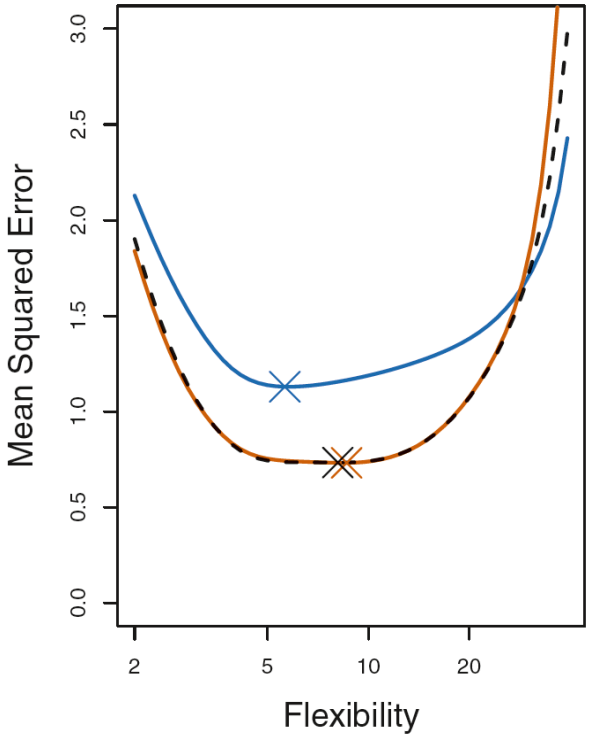


- Left: Validation Set Approach
- Right: 10-fold CV
- 10-fold CV is much more stable!

EXAMPLE: K-FOLD CV

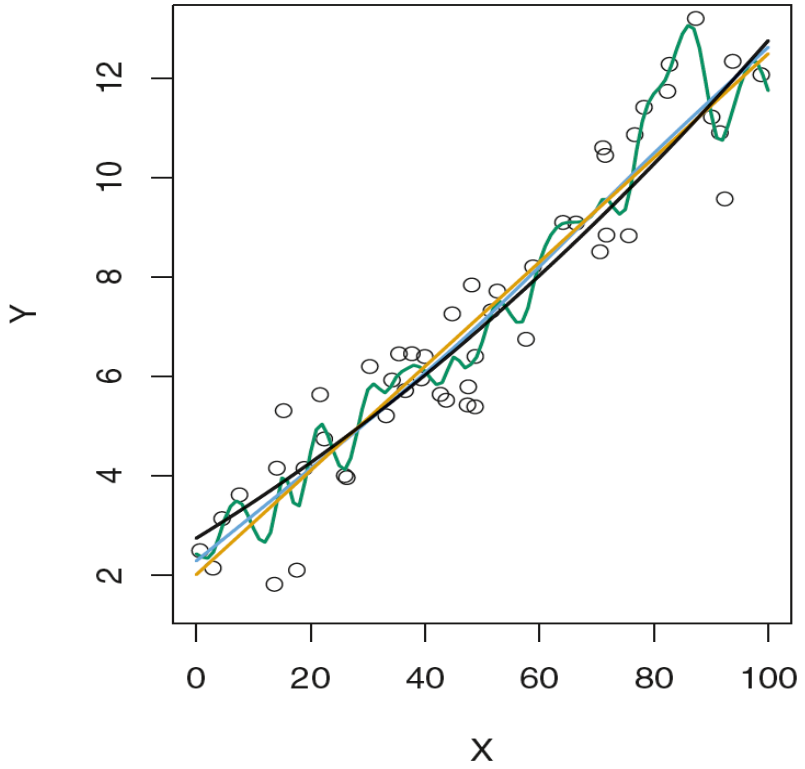


Orange: linear regression
Black: original data
Blue, Green: splines

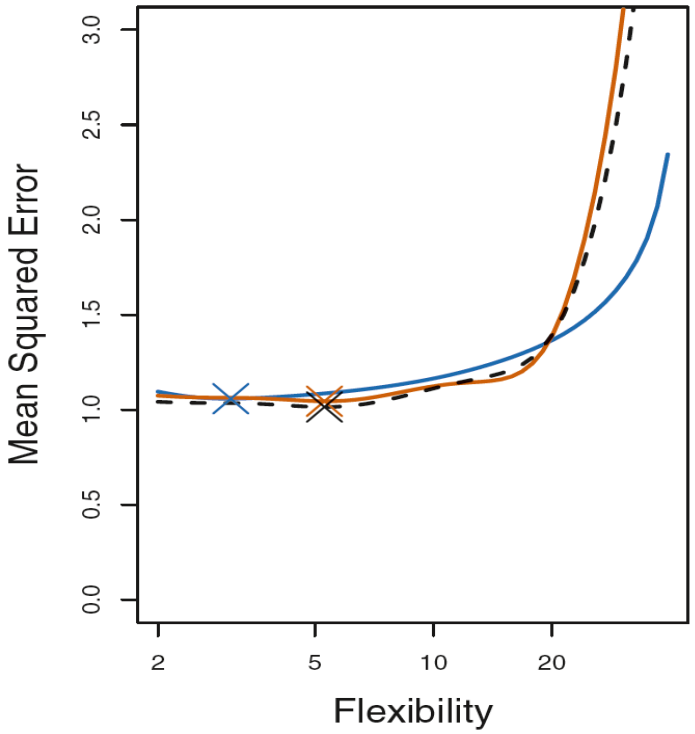


Blue: True Test MSE
Black: LOOCV MSE
Orange: 10-fold MSE

EXAMPLE: K-FOLD CV

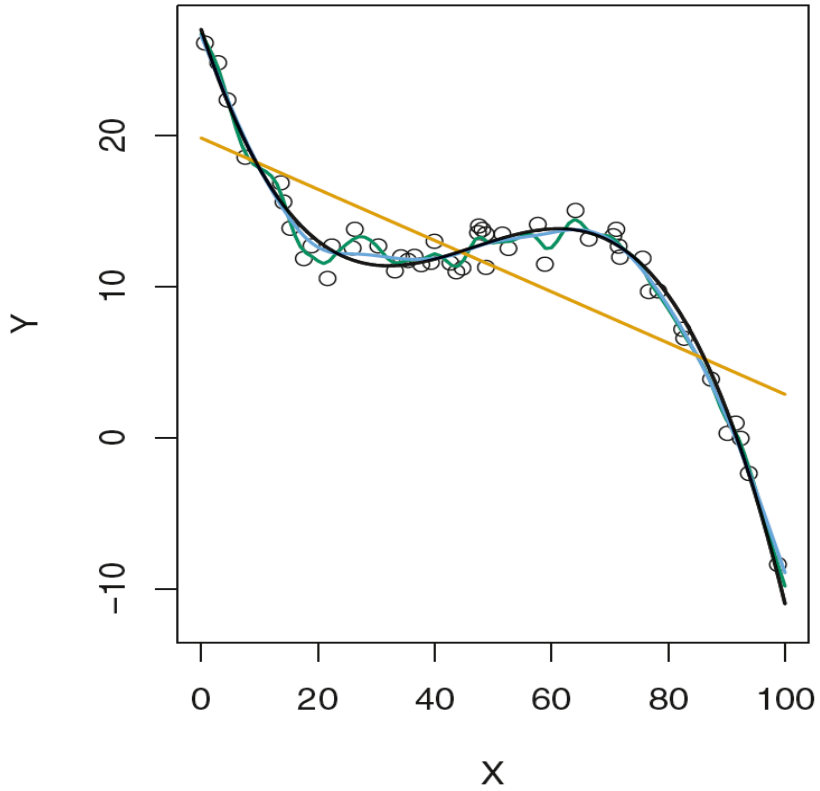


Orange: linear regression
Black: original data
Blue, Green: splines

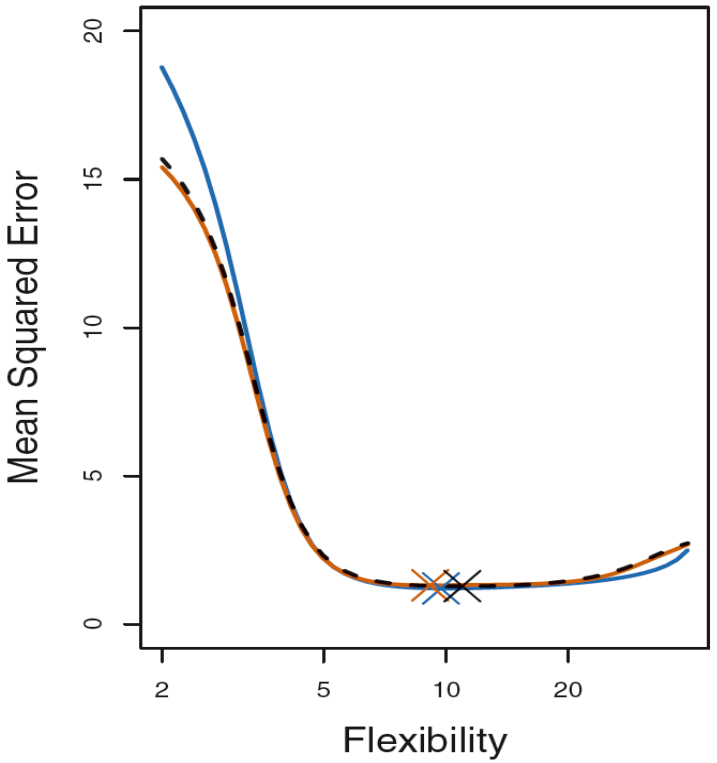


Blue: True Test MSE
Black: LOOCV MSE
Orange: 10-fold MSE

EXAMPLE: K-FOLD CV



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



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

COMPARISON

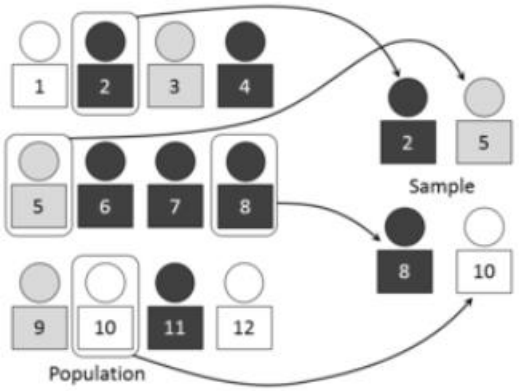
- Complexity:
  - LOOCV is more complicated than k-fold CV ($k < n$)
  - k-fold CV (k < n) is more complicated than the validation set approach
- Variance:
  - LOOCV has higher variance than k-fold CV (when $k < n$)
  - k-fold CV (k < n) has higher variance than the validation set approach

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

POPULATION AND SAMPLE



Population mean is μ. Let $\mu_1, \mu_2, \ldots, \mu_m$ are the averages of the m different samples. Then,

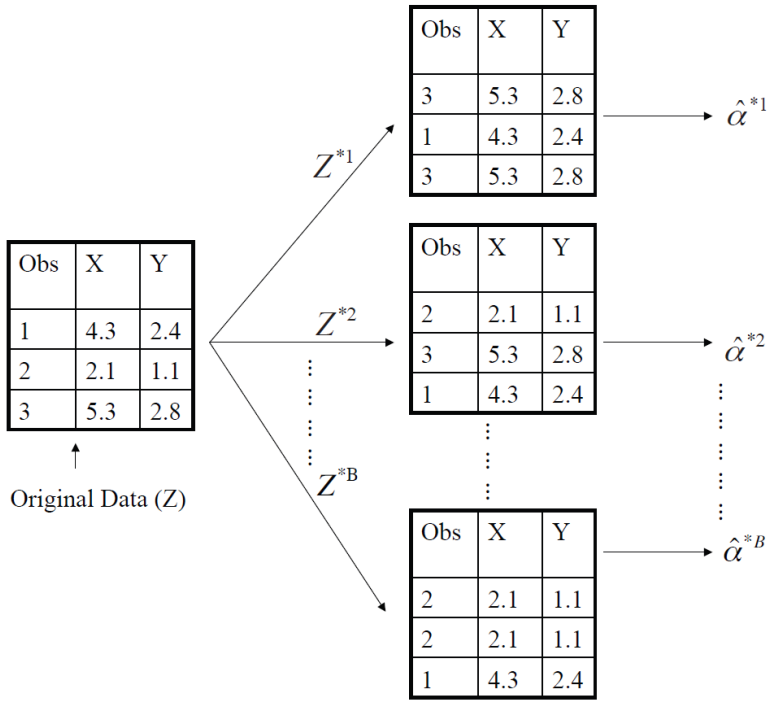
$$-z * \frac{\sigma}{\sqrt{m}} \leq \frac{\mu_1 + \cdots + \mu_m}{m} - \mu \leq z * \frac{\sigma}{\sqrt{m}}$$

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



- 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

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

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