Topic 4. Resampling

What Are Resampling Methods?

- Tools that involve repeatedly drawing samples from a training set and refitting a model of interest on each sample set in order to obtain more information about the fitted model
  - **Model assessment**: estimate test error rate of a learning method (when no test set is available)
  - **Model selection**: select the model with appropriate level of flexibility
- They are computationally expensive! But these days we have powerful computers.
- Two resampling methods:
  - Cross Validation
  - Bootstrapping

Cross Validation

Cross Validation

- Training error rate vs. test error rate
- Test error rate: average error when a statistical learning method is used to predict the response of a new observation that was not used in training the method
- Test error rate can be easily calculated if a designated test data set is available. Unfortunately, this is usually not the case.
- **Question:** how can we assess the test error rate using the available data set?
- Idea: hold out a subset of the data for test

Outline

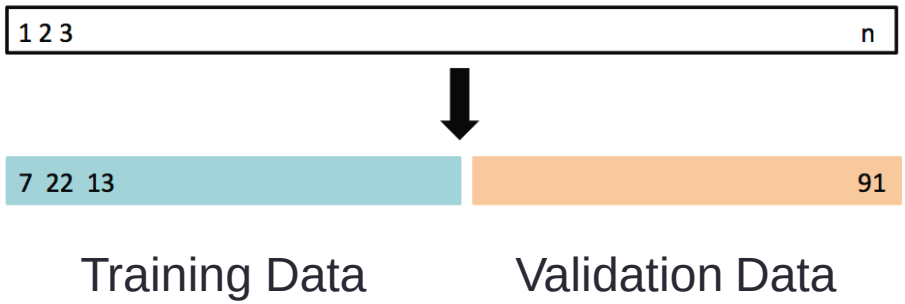
- **Cross Validation Methods**
  1. The Validation Set Approach
  2. Leave-One-Out Cross Validation
  3. K-fold Cross Validation

- **Cross Validation on Classification Problems**

- **Cross Validation Accuracy**

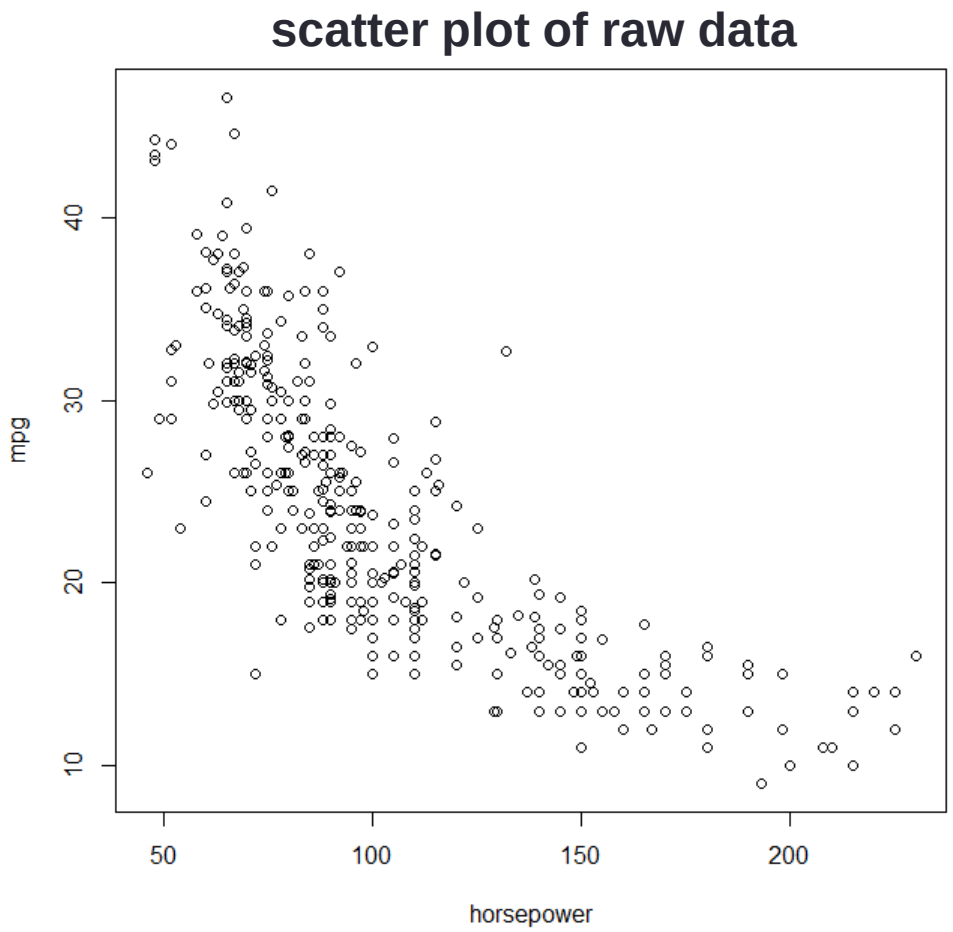
1. The Validation Set Approach

- Suppose that we have a number of possible models for a data set and want to find the best one that gives the lowest *test* (not training) error rate. No designated test set is available.
- We can randomly split the available data into two parts: a *training* set and a *validation* or *hold-out* set.
- We would then use the training set to build each possible model and choose the model that gives the lowest test error rate when applied to the validation data.



Example: Auto Data

- Suppose that we want to predict mpg from horsepower

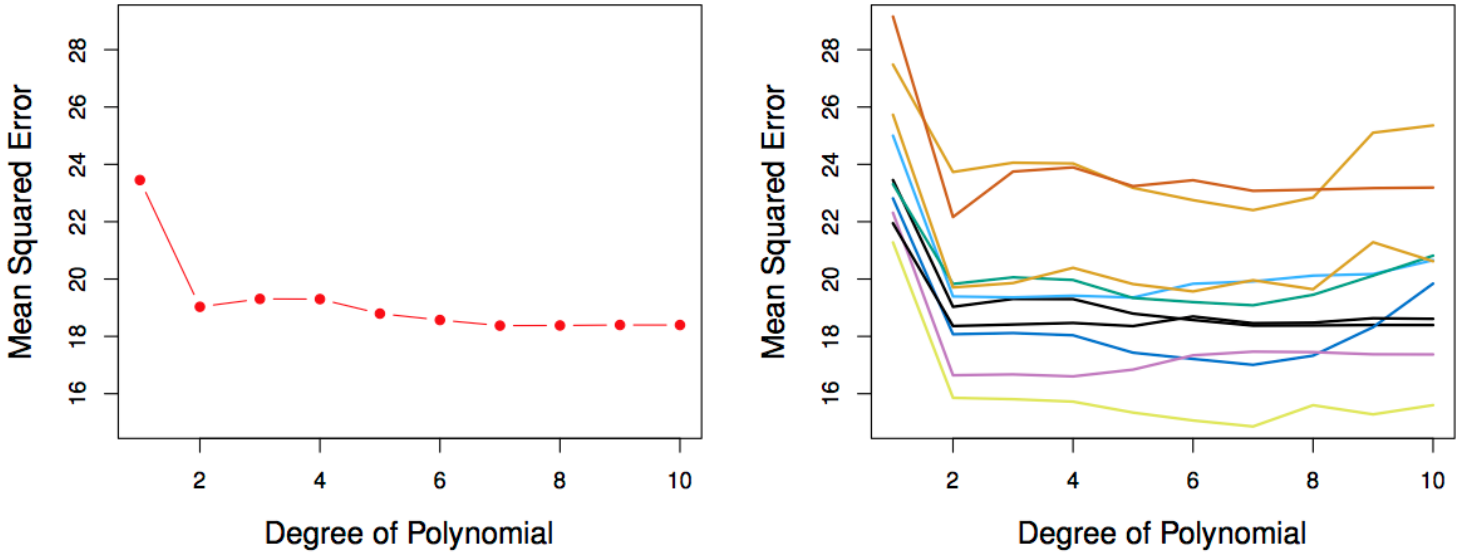


Problem

- Possible models:
  - mpg ~ horsepower (linear)
  - mpg ~ horsepower + horspower^2 (quadratic)
  - mpg ~ horsepower + horspower^2 + horspower^3 (cubic)
- Which model gives the best fit?
- The validation set approach
  - Randomly split the Auto data set (392 obs.) into training set (196 obs.) and validation set (196 obs.)
  - Fit each model using the training data set
  - Then, estimate test error rate using the validation data set
  - The model with the lowest validation (testing) MSE is the winner!

Results (Auto Data)

- Left: validation error rate for a single split
- Right: validation is repeated 10 times, each time the split is done randomly.



- The quadratic model is dramatically better than the linear model, and no benefit to use the cubic or higher-order models.
- A lot of variability among the MSE's from the 10 validations … Not good! We need more stable methods.

Advantages/Disadvantages of Validation Set Approach

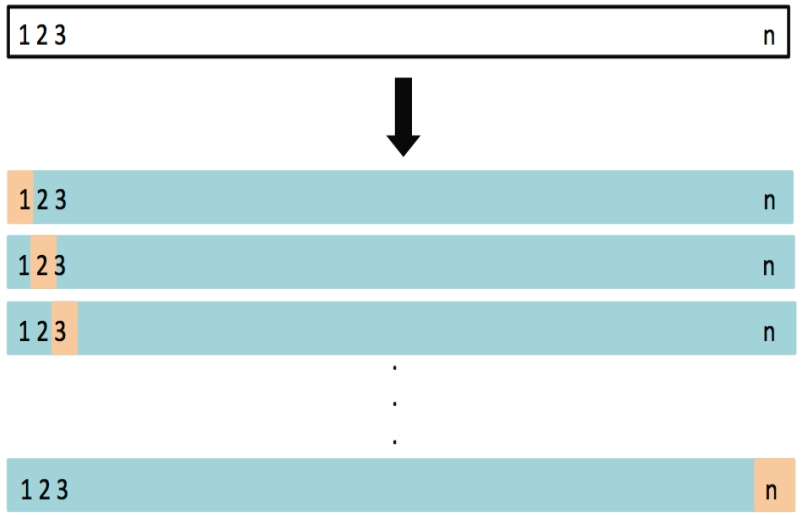
- Advantages
  - Conceptually simple
  - Easy to implement

- Disadvantages
  - The validation MSE can be highly variable.
  - Only a subset of observations are used to fit the model (training data). Statistical methods tend to perform worse when trained on fewer observations. This suggests that the validation set error rate may tend to *overestimate* the test error rate for the model fit on the entire data set.

2. Leave-One-Out Cross Validation (LOOCV)

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

- For each given model:
  - Split the data set of size n into
    - Training data set (blue) size: $n - 1$
    - Validation data set (beige) size: 1
  - Fit the model using the training data
  - Validate model using the validation data, and compute the corresponding MSE
  - Repeat this process n times

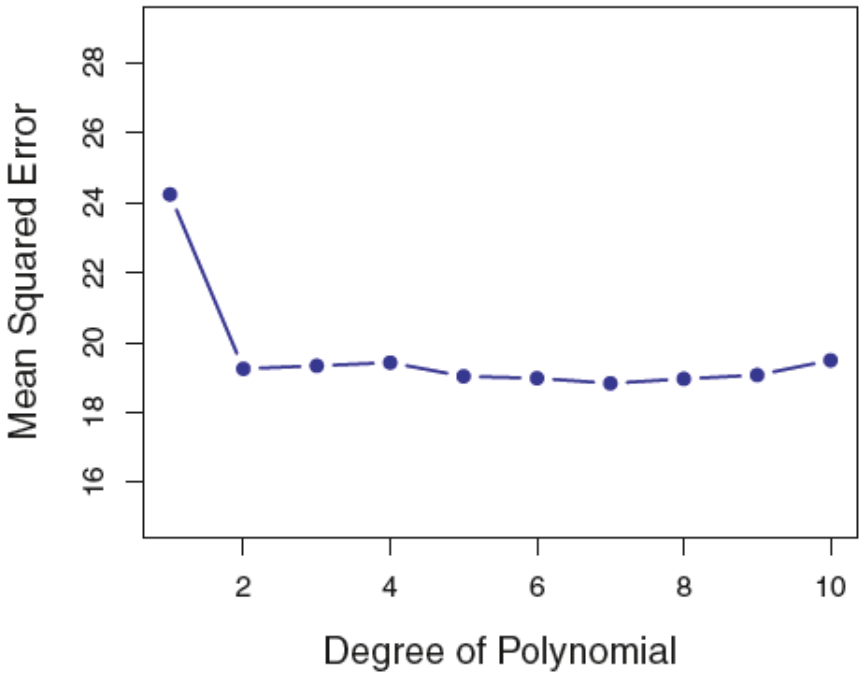


- The MSE for the model is computed as follows:

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

Results (Auto Data)

- Results: LOOCV test MSEs for different polynomial models



- Again, the quadratic model is the best.

LOOCV vs. Validation Set Approach

- LOOCV has less bias
  - We repeatedly fit the statistical learning method using training data that contains $n - 1$ observations, i.e., almost the entire data set is used.
- LOOCV produces a less variable MSE
  - The validation approach produces different MSE when applied repeatedly due to randomness in the splitting process, while performing LOOCV multiple times will always yield the same results, because we split based on one observation each time.
- LOOCV is computationally intensive (disadvantage)
  - We need to fit each model n times!

A Short Cut for LOOCV

- For least squares linear or polynomial regression, there is a simple formula for calculating the LOOCV error

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

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

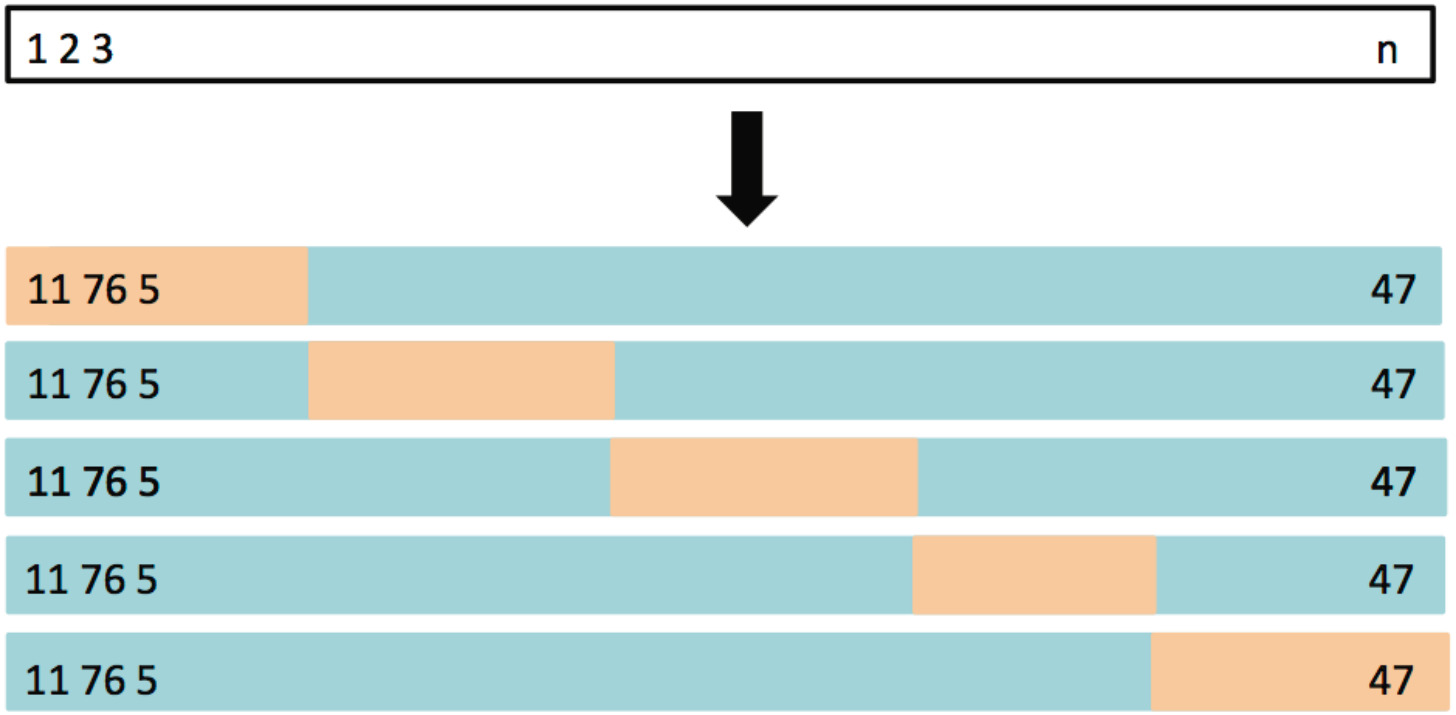
- Using this magic formula, we only need to fit one model using all the data.
- However, this formula does not hold in general. That is, for other methods (e.g., logistic regression, LDA), we still need to refit the model n times.

3. K-fold Cross Validation

- LOOCV is computationally intensive, so we can run k-fold Cross Validation (k-fold CV) instead.
- We divide the data set into k different parts (e.g., k = 5, or 10, etc.)
- 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 on the first part)
- We repeat this k different times taking out a different part each time
- By averaging the k different MSE's we get an estimated validation (test) error rate

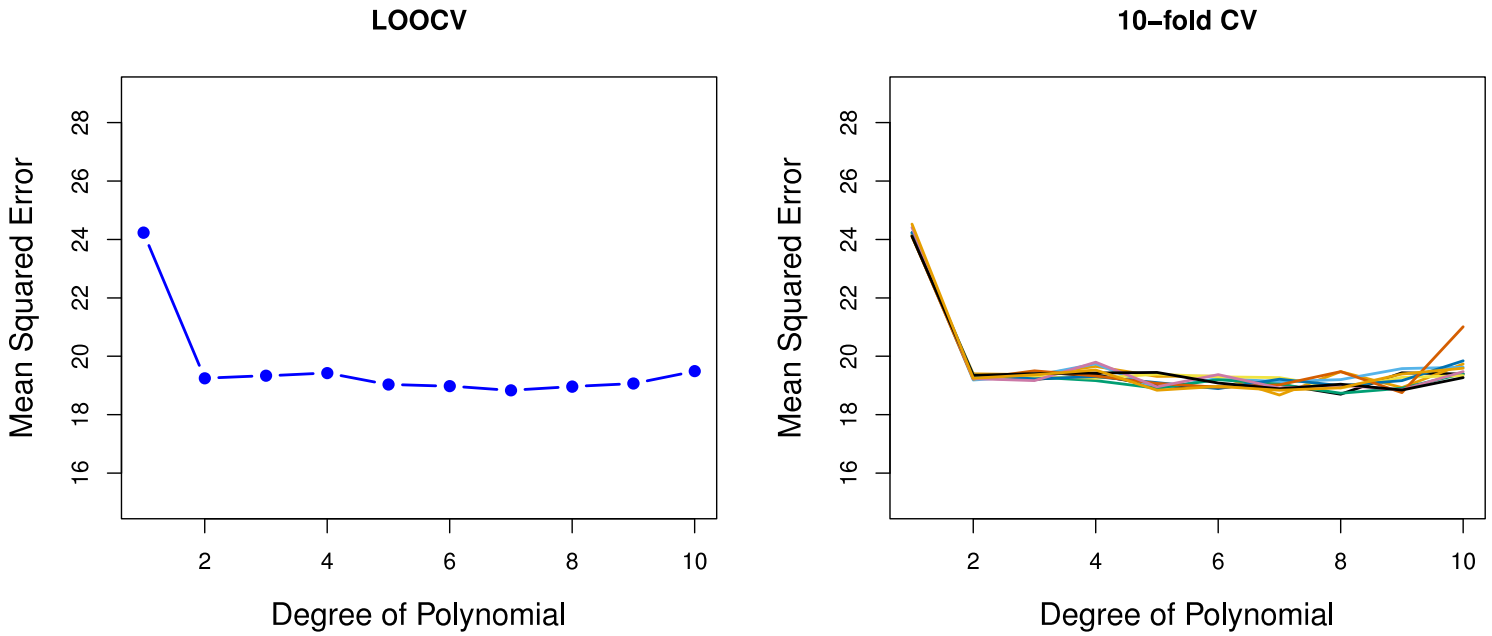
$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

Illustration of K-fold CV



Results (Auto Data): LOOCV vs. K-fold CV

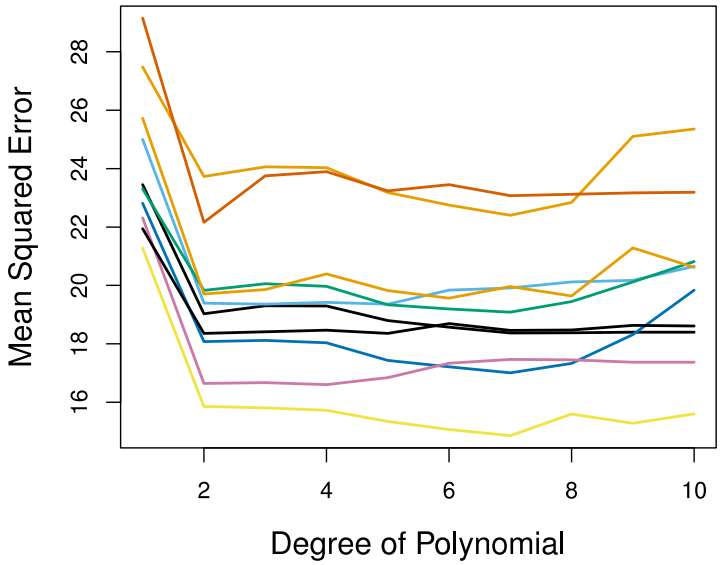
- Left: LOOCV error curve
- Right: 10-fold CV was run many times

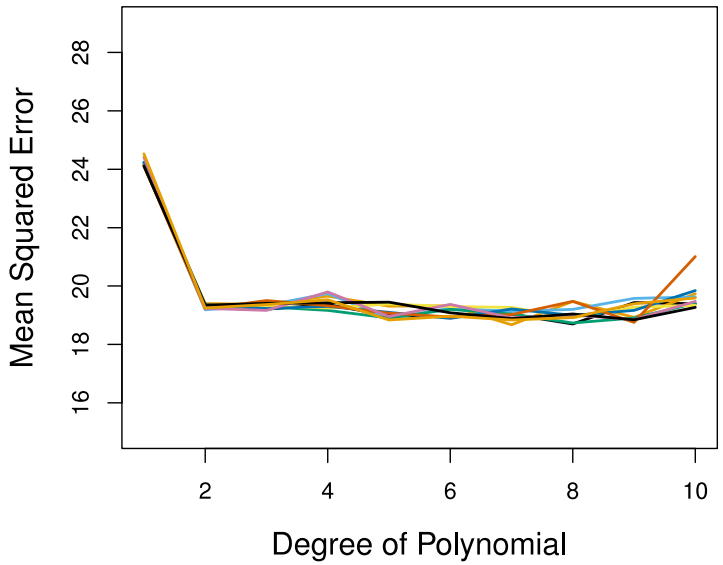


- LOOCV is a special case of k-fold, where k = n
- They are both stable, but LOOCV is more computationally intensive!

Results (Auto Data): Validation Set Approach vs. K-fold CV

- Left: Validation Set Approach
- Right: 10-fold Cross Validation Approach
- Indeed, 10-fold CV is more stable!



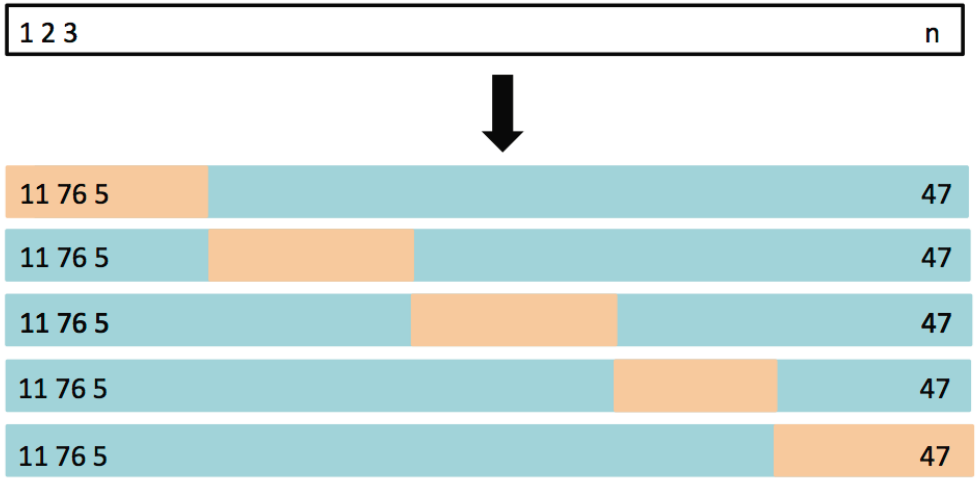


The Choice of K

- We tend to use k-fold CV with K = 5 and K = 10.
- These are the magical K's.
- It has been empirically shown that they yield test error rate estimates that suffer neither from excessively high bias, nor from very high variance.

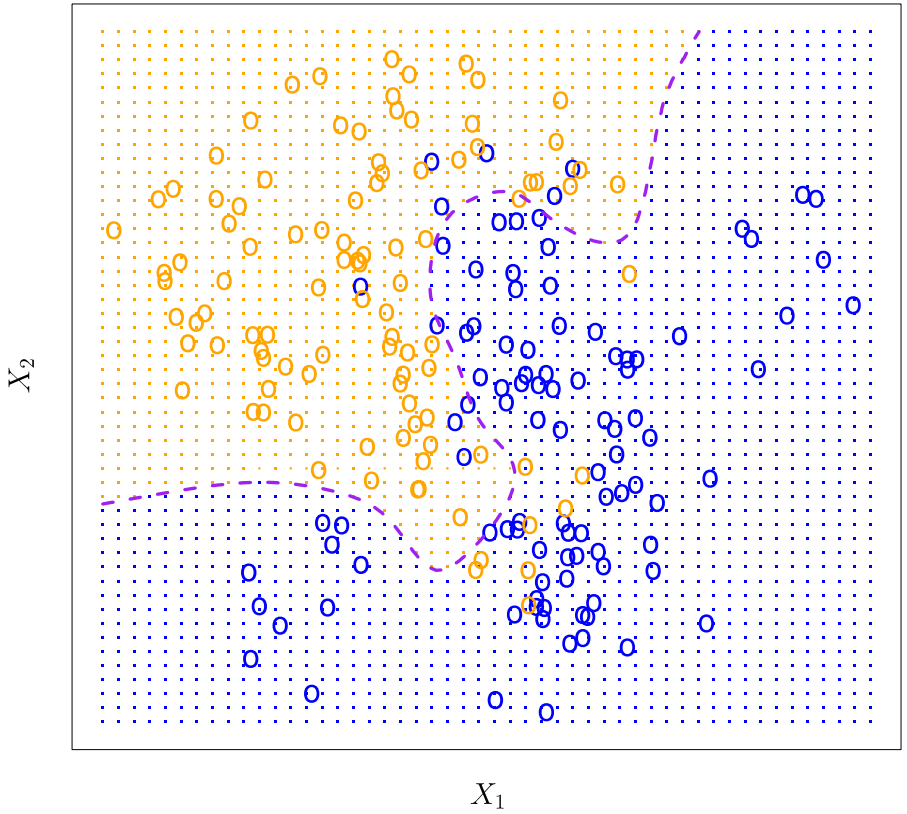
Cross-Validation on Classification

- So far, we have been dealing with CV on regression problems.
- We can use cross validation in classification in a similar manner
  - Divide data into k parts
  - Hold out one part, fit model using the remaining data and compute the test error rate (#misclassified observations!) on the hold out data
  - Repeat k times
  - CV error rate is the average over the k error rate estimates computed



Example: CV to Select Logistic Regression Models

- The data set used is simulated
- The purple dashed line is the Bayes' boundary

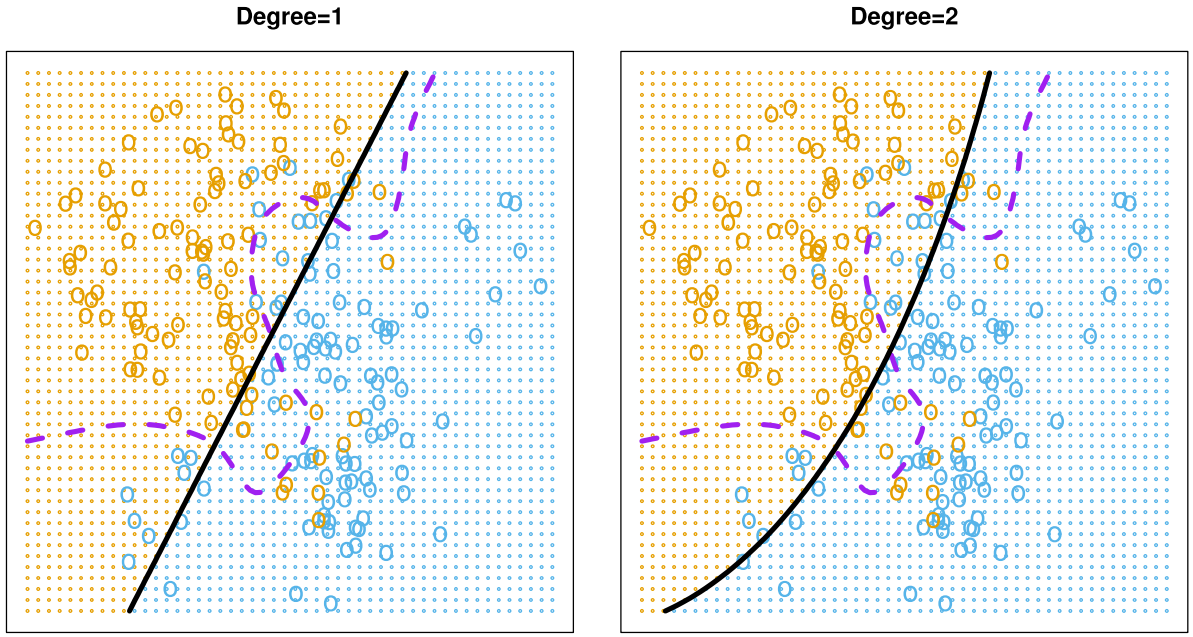


Bayes error rate: 0.133

Logistic Regression Models

- Linear logistic regression (Degree 1) is not able to fit the Bayes' decision boundary.
- Quadratic logistic regression does better than linear

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$$

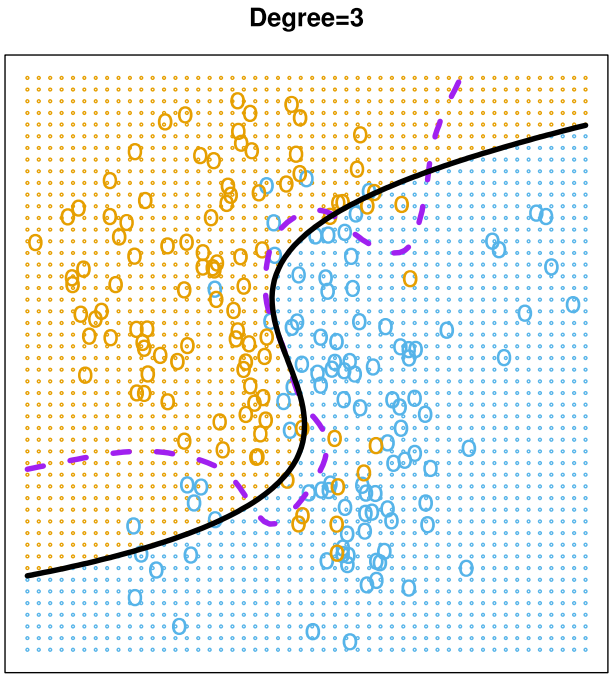


Error rate: 0.201

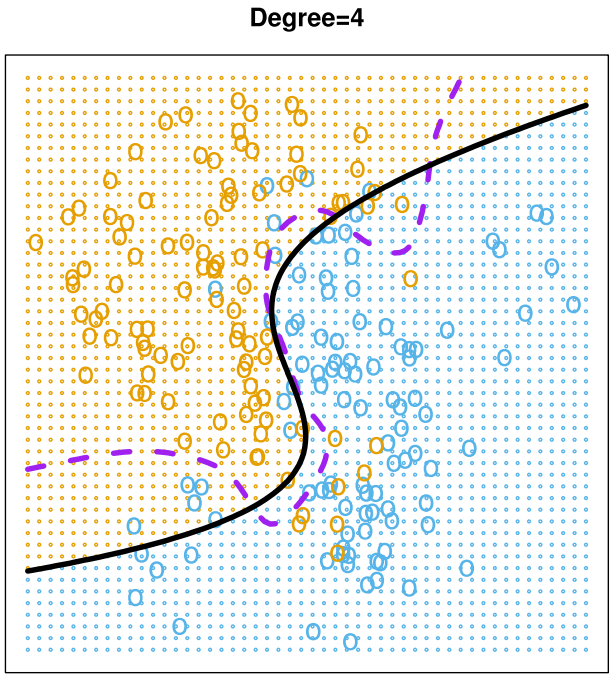
Error rate: 0.197

Higher-degree Logistic Regression

- Using cubic and quartic predictors, the accuracy of the model improves



Error Rate: 0.160



Error Rate: 0.162

CV to Choose the Order

- If the test data set is not available, the test error rates will be unknown. How to decide which one is the best among the four logistic regression models?

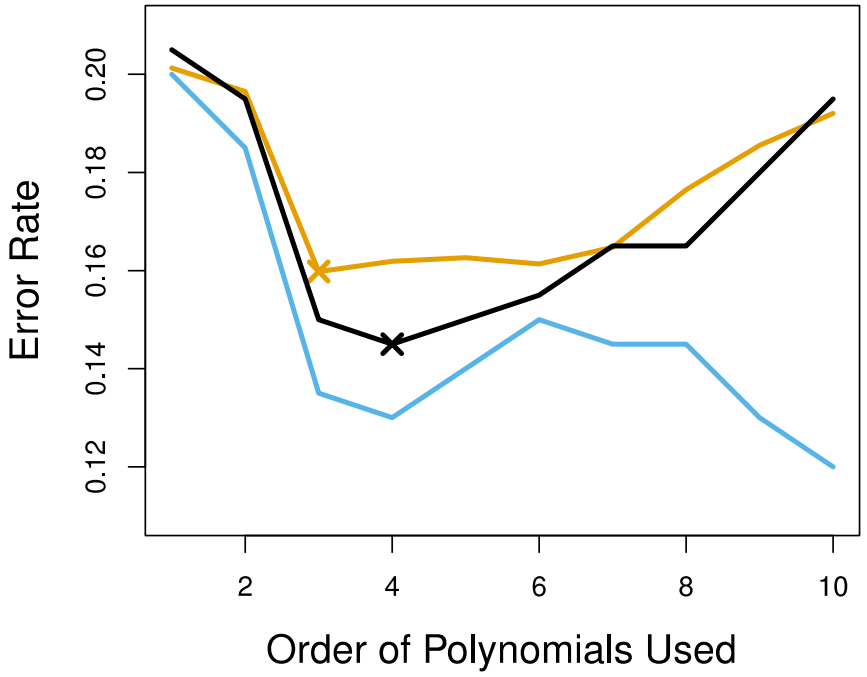
- Apply 10-fold CV on the available data

  Brown: True Test Error

  Blue: Training Error

  Black: 10-fold CV Error

  (1) CV curve approximates the true test curve well.
  (2) CV chooses the 4th order, which is close to the 3rd order chosen by the test curve.

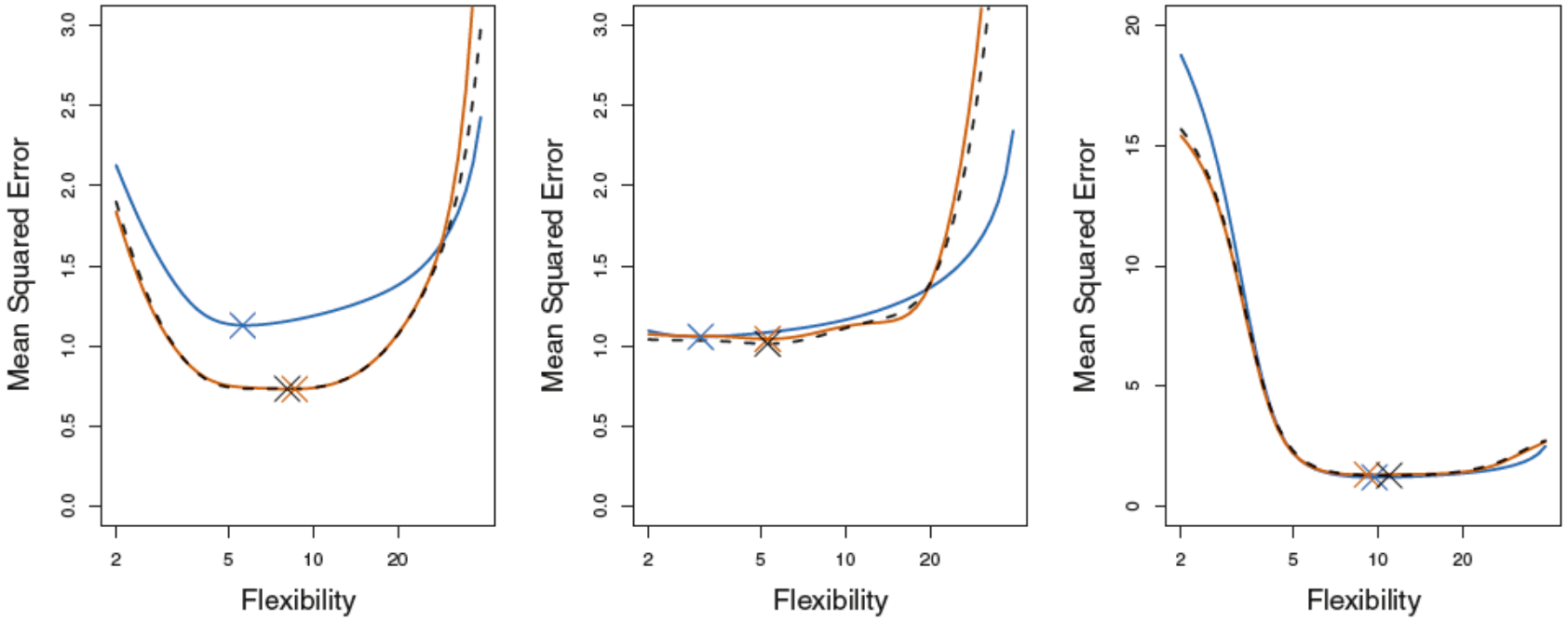


CV Estimates of Test Error vs. True Test Error

- MSEs from cross validation are estimates of test error rate.
- For real data, we do not know the true test MSE, so it is difficult to determine the accuracy of the CV estimates.
- In simulation studies, we can compute the true test MSE, and thereby evaluate the accuracy of the CV estimates.

A Simulation Study

- Three simulated cases
- Blue: True test error
- Black dashed: LOOCV estimate of test error
- Orange: 10-fold CV estimate of test error
- Crosses: minimum of each curve



- The LOOCV and 10-fold CV estimates are very similar.

Accuracy of CV Estimates: Model Assessment

- **Model assessment**: determine how well a given statistical learning method can perform on new data
- Interest is on the *estimate of the test MSE*.
- **Results** of the simulation study

  Right: CV estimates and the true test MSE are almost identical.

  Center: the two are similar at lower degrees of flexibility, while CV estimates overestimate the true test error for higher degrees of flexibility.

  Left: the CV estimates have the correct general shape, but they underestimate the true test MSE.

Accuracy of CV Estimates: Model Selection

- **Model selection**: determine the best learning method which results in the lowest test error from a number of possible methods (which have different levels of flexibility)
- Interest is on the *minimum point* in the estimated test MSE curve
- **Results** of the simulation study

  All the CV curves come close to identifying the correct level of flexibility, i.e., the flexibility level corresponding to the smallest test MSE.

Bootstrap

Motivation

- **Question**: how to quantify the uncertainty associated with a given estimator or statistical learning method?
- A simple example: estimate standard errors of coefficients from a linear regression fit
- Chapter 3 provides formulas

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  R outputs the standard errors automatically.

- For learning methods that are more complex than linear regression, such formulas may not be available. What can we do to find the standard error in their parameter estimation?

The Bootstrap

- A powerful tool to quantify the uncertainty associated with a given estimator or statistical learning method
- The power of the bootstrap lies in the fact that it can be easily applied to a wide range of statistical learning methods, including those for which a measure of variability is otherwise difficult to obtain and is not automatically output by software.

Example: the Problem

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y, which are two random variables.
- We will invest a fraction α in X, and the remaining $1 - \alpha$ in Y.
- We wish to choose α to minimize the total risk, or variance, of our investment; that is, to minimize $Var(\alpha X + (1 - \alpha)Y)$.
- One can show that the value minimizing the risk is

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

- **Problem:** how to estimate α?

Example: Estimating the Optimal α

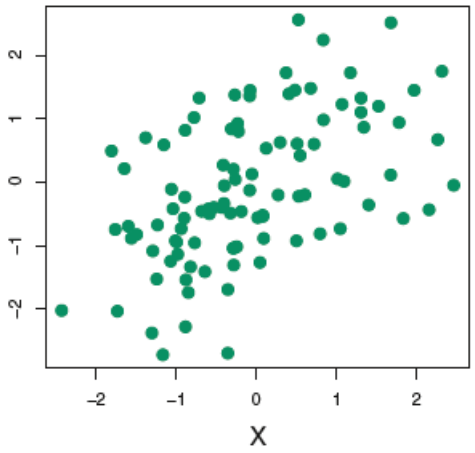
- In reality, the quantities $\sigma_X^2, \sigma_Y^2, \sigma_{XY}$ are unknown.
- We can compute estimates for them, $\hat{\sigma}_X^2, \hat{\sigma}_Y^2, \hat{\sigma}_{XY}$, using a data set that contains past measurements for X and Y.
- We then estimate the value of the optimal α by

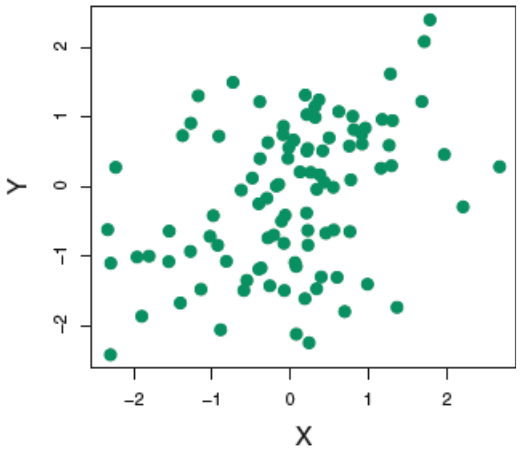
$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

Example: Simulation Study

- Each panel contains 100 pairs of observations for X and Y

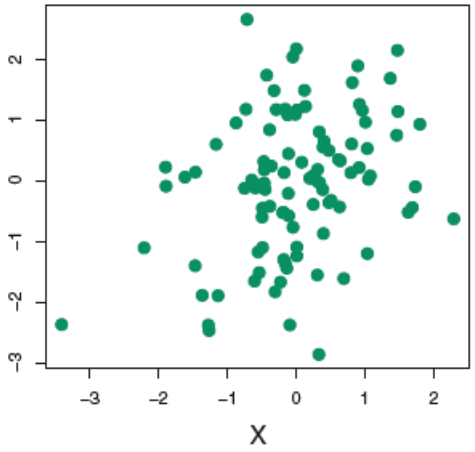
$\hat{\alpha} = 0.576$

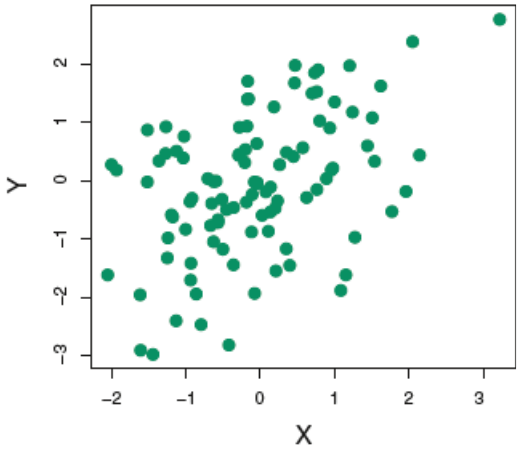




$\hat{\alpha} = 0.532$

$\hat{\alpha} = 0.657$





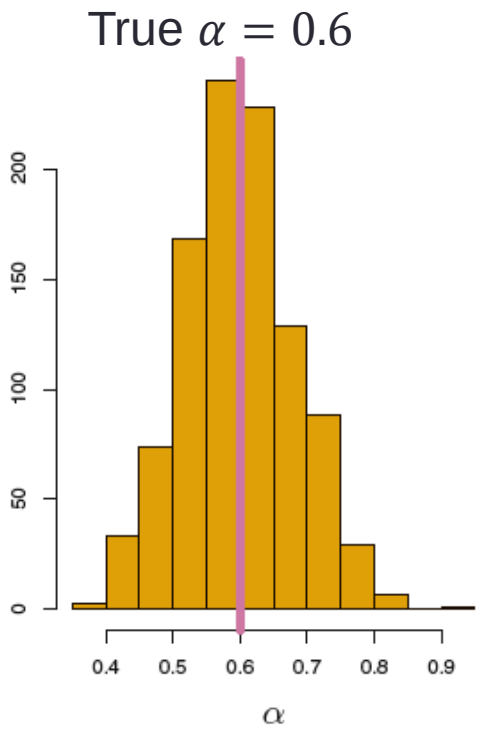
$\hat{\alpha} = 0.651$

Example: Quantifying Accuracy of Estimates

- Generate 1000 simulations, each of which yields one estimate of α. Calculate mean and standard deviation of these estimates.

$$\bar{\alpha} = \frac{1}{1,000} \sum_{r=1}^{1,000} \hat{\alpha}_r = 0.5996$$

$$\sqrt{\frac{1}{1,000-1} \sum_{r=1}^{1,000} (\hat{\alpha}_r - \bar{\alpha})^2} = 0.083$$



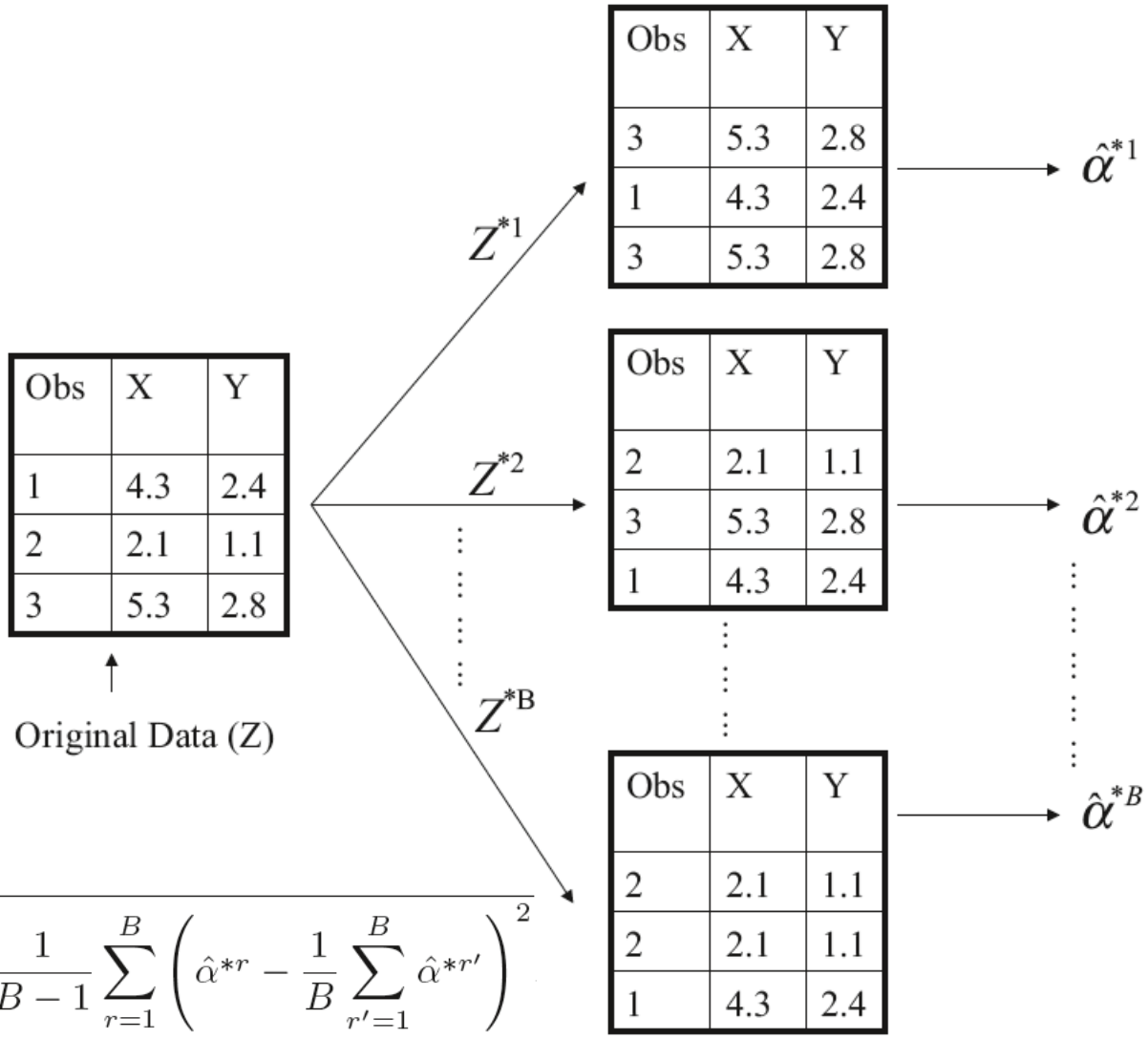
However…

- This procedure needs 1000 data sets from the true population to quantify the accuracy (or uncertainty) of the estimate of α.
- However, in practice, it is usually not possible to generate new samples from the true population.
- Consider the case where only one data set is available, how can we quantify the accuracy of $\hat{\alpha}$?

Idea of The Bootstrap

- The bootstrap approach uses a computer to emulate the process of obtaining new sample sets, so that we can estimate the variability of $\hat{\alpha}$ 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.

Illustration on A Dataset with $n = 3$

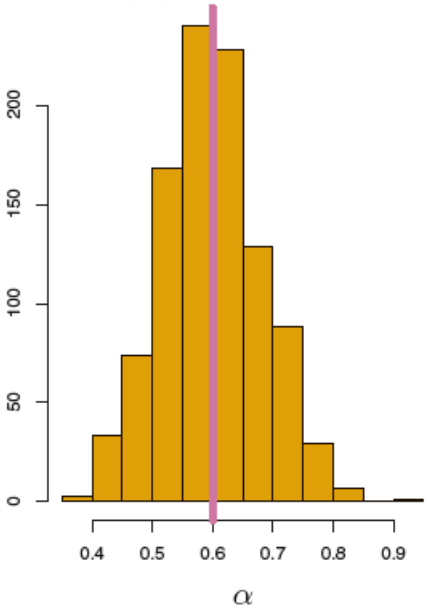


$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}$$

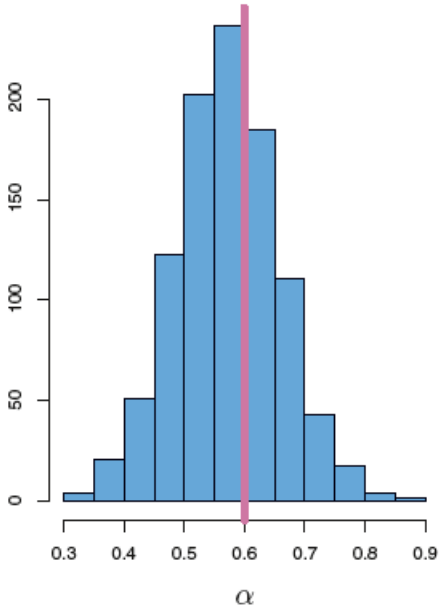
Example: Results from the Bootstrap

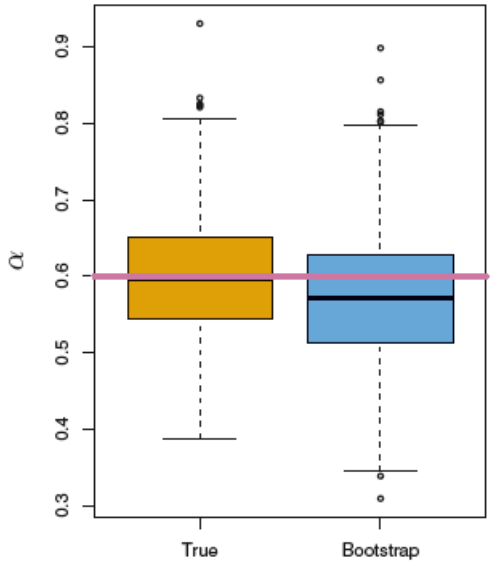
- Left: histogram of estimates obtained by generating 1000 simulated data sets from the true population.
- Center: histogram obtained from 1000 bootstrap samples from a single data set.

$SE(\hat{\alpha}) = 0.083$



$SE(\hat{\alpha}) = 0.087$





Example: Summary of Results

- The bootstrap estimate of $SE(\hat{\alpha})$ is very similar to the estimate based on 1000 data sets.
- The left is based on 1000 simulated data sets from the true population, while the right is based on only a single data set.
- The left represents the idealized situation, and the right is for real data.

Advantages of The Bootstrap

- One of the great advantages of the bootstrap approach is that it can be applied in almost all situations (for various statistical learning methods and various types of data).
- No complicated mathematical calculations are required.

Questions (1)

Read slides of Topic 4. Then answer the following questions.

- What are the two scenarios where resampling methods are often used?
- What are the three cross validation methods?
- What are the pros and cons of each cross validation method? Which one of the three is the most popular?
- Does the short-cut for calculating LOOCV apply for all methods?
- What are the popular choices of K in K-fold CV?
- What is the main difference of CV for classification vs. CV for regression?
- Is CV estimates accurate in model assessment?
- Is CV estimates accurate in model comparison?

Questions (2)

- For what purpose the Bootstrap is often used?
- Is the Bootstrap needed to find accuracy of coefficient estimates in linear regression? When is it needed?
- In Bootstrap, is the sampling from the original data set with or without replacement?
- What are the major advantages of the Bootstrap?