Resampling Methods

Nguyen Hong Son, PhD HCM PTIT

Introduction

- When building models, we often face questions about how well a model will perform on new data or how accurate the estimates are.
- Resampling methods involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model. (Resampling involves reusing your one dataset many times.)
- For example, in order to estimate the variability of a linear regression fit, we can repeatedly draw different samples from the training data, fit a linear regression to each new sample, and then examine the extent to which the resulting fits differ.
- Resampling approaches can be computationally expensive, because they involve fitting the same statistical method multiple times using different subsets of the training data.

Typical resampling methods

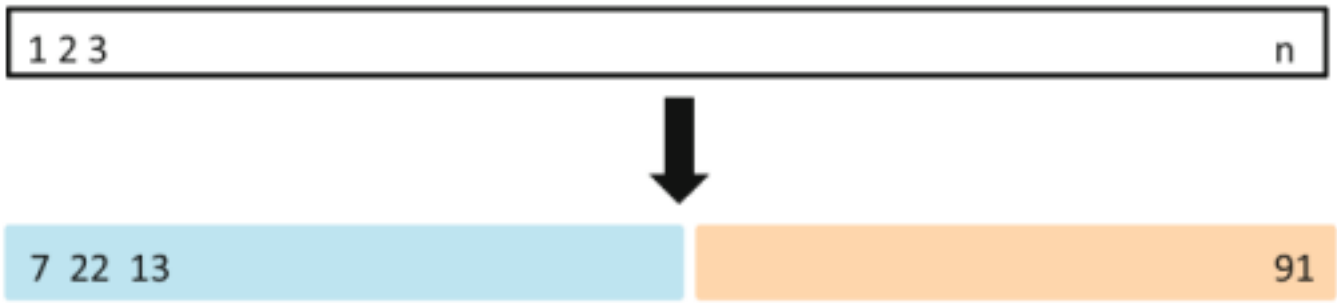
- Two of the most commonly used resampling methods:
  - **Cross-Validation**
  - **The Bootstrap**
- For example, cross-validation can be used to estimate the test error associated with a given statistical learning method in order to evaluate its performance, or to select the appropriate level of flexibility.
- The bootstrap is used in several contexts, most commonly to provide a measure of accuracy of a parameter estimate or of a given statistical learning method.

Cross-Validation (C-V)

- The training error rate vs the test error rate
- Given a data set, the use of a particular statistical learning method is warranted if it results in a low test error. The test error can be easily calculated if a designated test set is available. Unfortunately, this is usually not the case.
- In the absence of a very large designated test set that can be used to directly estimate the test error rate, a number of techniques can be used to estimate this quantity using the available training data.
- There is a class of methods that estimate the test error rate by holding out a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.

C-V: The Validation Set Approach

- Suppose that we would like to estimate the test error associated with fititng a particular statistical learning method on a set of observations.
- The validation set approach, displayed in figure, is a very simple strategy validation for this task.
  - The validation set approach involves randomly dividing the available set of observations into two parts, a training set and a validation set or hold-out set.
- The resulting validation set error rate provides an estimate of the test error rate

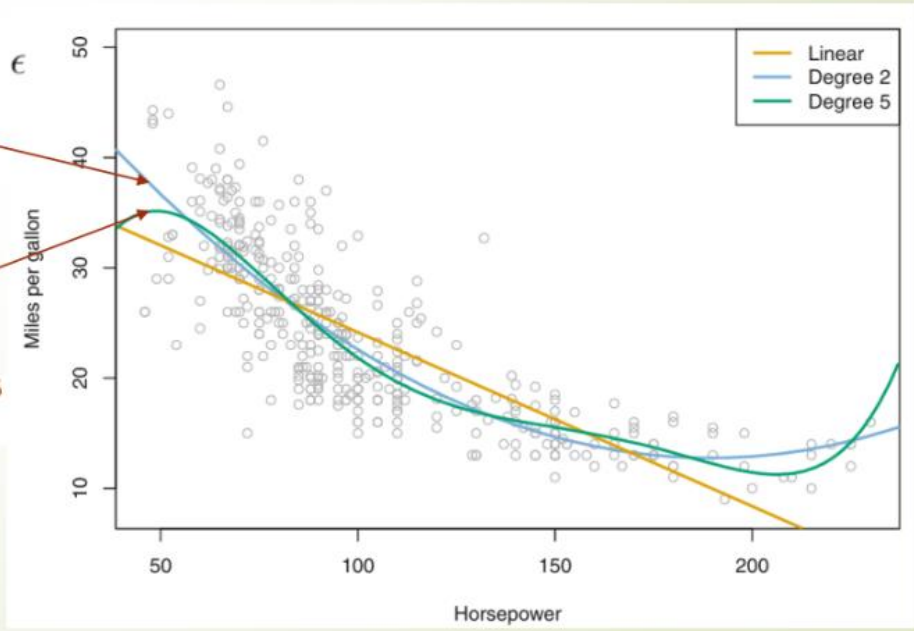


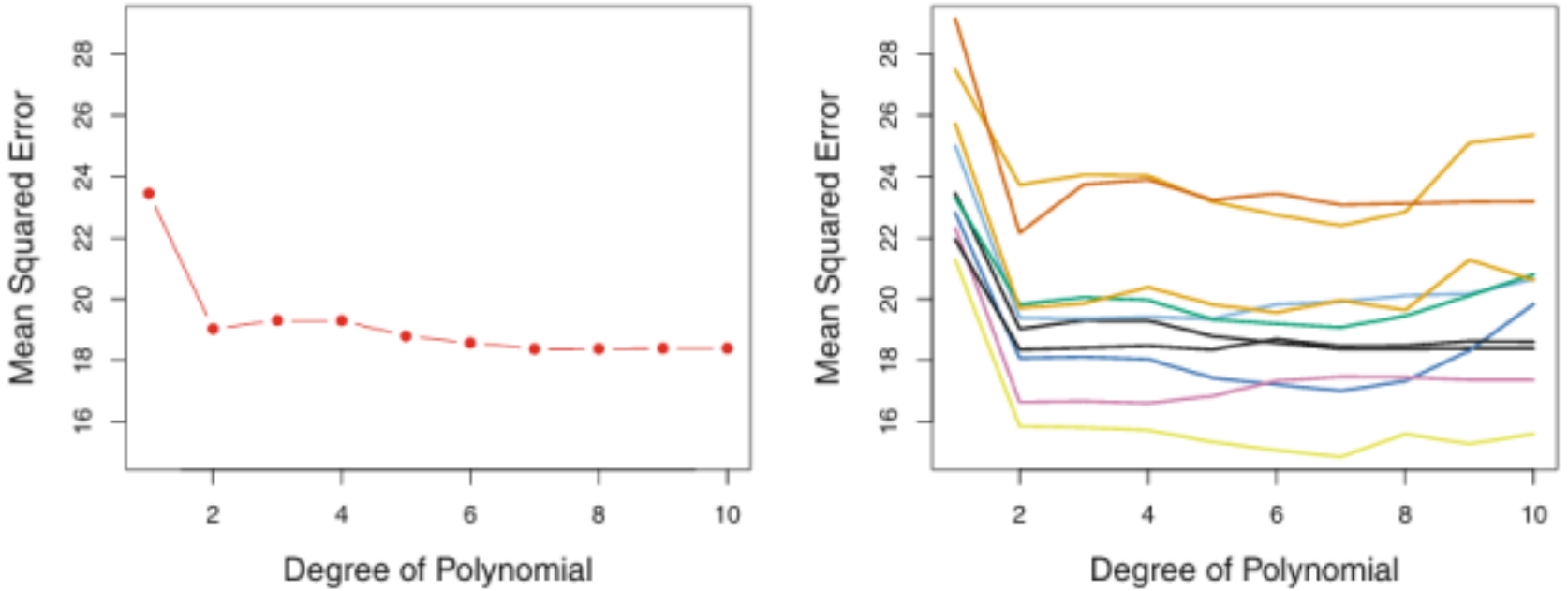
C-V: For Example:

- The validation set approach on the Auto data set
- Consider figure, in which the mpg (gas mileage in miles per gallon) versus horsepower is shown for a number of cars in the Auto data set (polynomial regression)

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

include $\texttt{horsepower}^3$, $\texttt{horsepower}^4$, or even $\texttt{horsepower}^5$





- The left-hand panel, we randomly divided the data set into two parts, a training set and a validation set.
- The right-hand panel displays ten different validation set MSE curves from the Auto data set, produced using ten different random splits of the observations into training and validation sets.
- Based on the variability among these curves, all we can conclude with any confidence is that the linear fit is inadequate for this data.

C-V: Drawbacks of The validation set approach

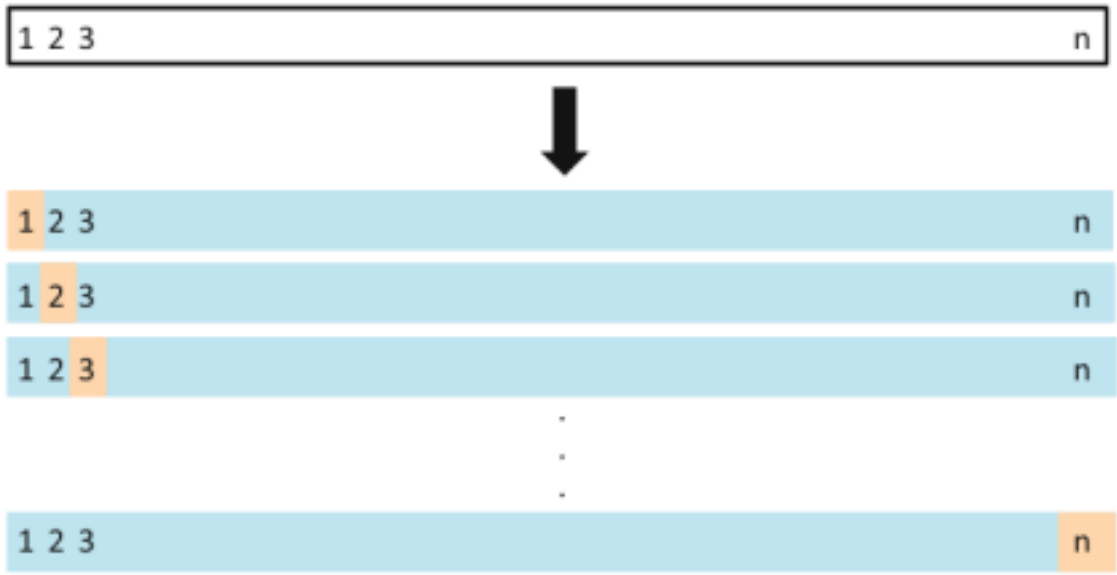
- The validation set approach is conceptually simple and is easy to implement. But it has two potential drawbacks:
  - The validation estimate of the test error rate can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.
  - Since 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.
- The cross-validation, a refinement of the validation set approach that addresses these two issues.

C-V: Leave-One-Out Cross-Validation

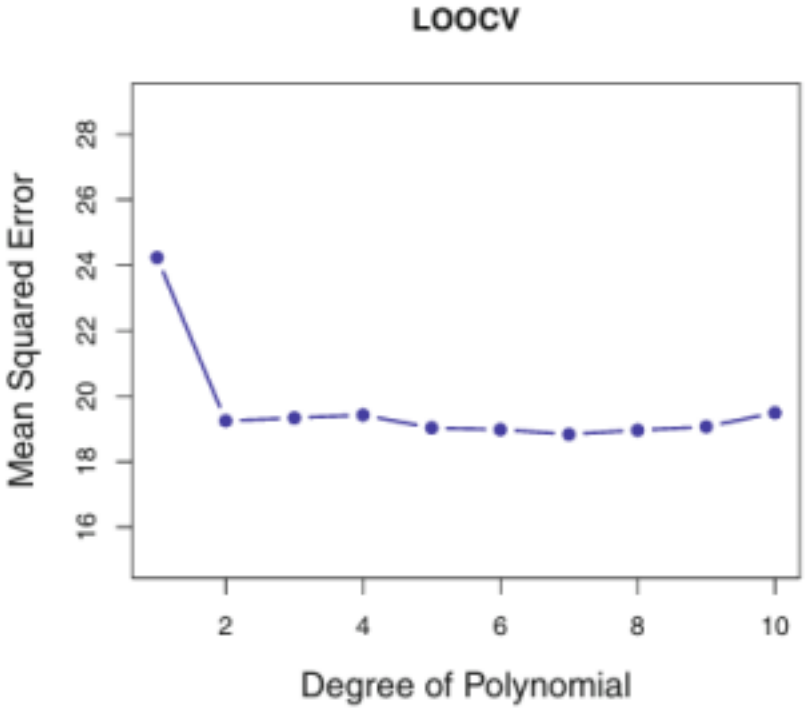
- LOOCV involves splitting the set of observations into two parts. However, instead of creating two subsets of comparable size, a single observation (x1,y1) is used for the validation set, and the remaining observations {(x2,y2),...,(xn,yn)} make up the training set

The LOOCV estimate for the test MSE is the average of these n test error estimates:

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



. . . .



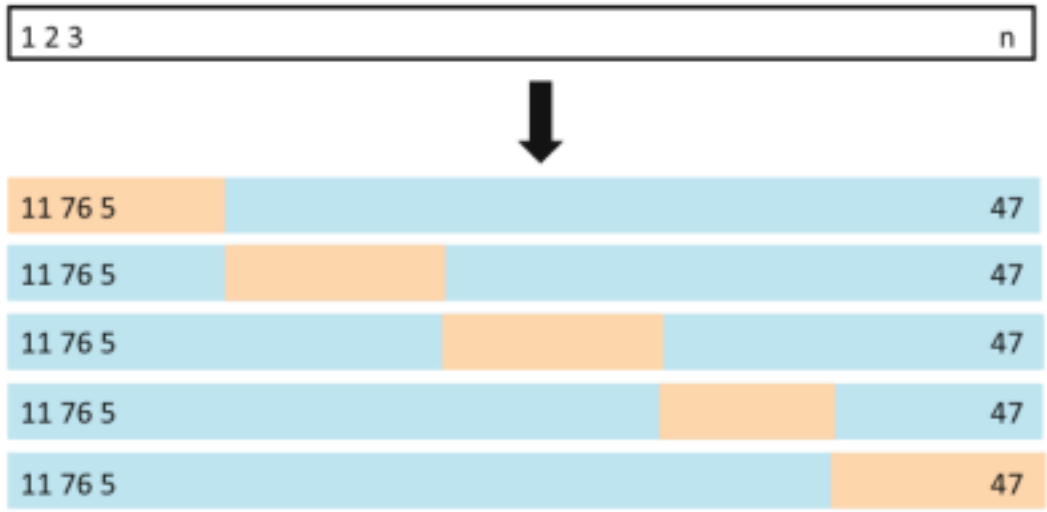
LOOCV on the Auto data set in order to obtain an estimate of the test set MSE that results from fitting a linear regression model to predict mpg using polynomial functions of horsepower

C-V: k-Fold Cross-Validation

 This approach involves randomly k-fold CV dividing the set of observations into k groups, or folds, of approximately equal size.

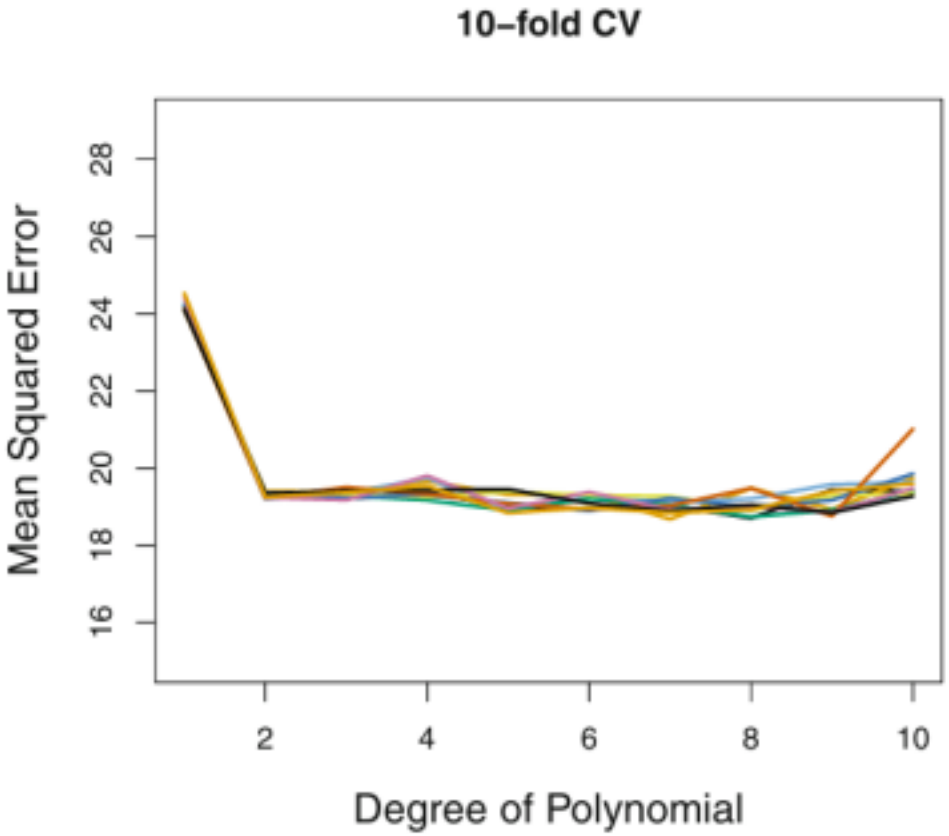
 The first fold is treated as a validation set, and the method is fit on the remaining k − 1 folds. The mean squared error, MSE_1, is then computed on the observations in the held-out fold. This procedure is repeated k times; The k-fold CV estimate is computed by averaging these values:

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



A schematic display of 5-fold CV

. . . .



Cross-validation was used on the Auto data set

10-fold CV was run nine separate times, each with a different random split of the data into ten parts.

Bias-Variance Trade-Off for k-Fold Cross Validation

- Performing k-fold CV will lead to an intermediate level of bias, since each training set contains (k−1)n/k observations—fewer than in the LOOCV approach, but substantially more than in the validation set approach. Therefore, from the perspective of bias reduction, it is clear that LOOCV is to be preferred to k-fold CV.
- Since the mean of many highly correlated quantities has higher variance than does the mean of many quantities that are not as highly correlated, the test error estimate resulting from LOOCV tends to have higher variance than does the test error estimate resulting from k-fold CV.
- There is a bias-variance trade-off associated with the choice of k in k-fold cross-validation. Typically, given these considerations, one performs k-fold cross-validation using k =5 or k = 10, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance.

C-V: An example in machine learning

- In this example, we use the **Iris dataset** and a **Support Vector Machine (SVM)** classifier. The data is split into 5 folds using **KFold**, and the classifier is trained and tested on each fold. The accuracy of the model is then averaged across all the folds to give a reliable estimate of its performance.
- The program in next slide

The example program

```python
# Step 1: Import necessary libraries
from sklearn.model_selection import KFold
from sklearn.datasets import load_iris
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score

# Step 2: Load the dataset
iris = load_iris()
X, y = iris.data, iris.target
# Step 3: Create a classifier (Support Vector Machine)
clf = SVC(kernel='linear')
# Step 4: Define the number of folds for cross-validation
kf = KFold(n_splits=5)
# Step 5: Perform K-Fold Cross-Validation
accuracies = []
for train_index, test_index in kf.split(X):
 X_train, X_test = X[train_index], X[test_index]
 y_train, y_test = y[train_index], y[test_index]
# Train the model
clf.fit(X_train, y_train)
# Predict and evaluate
y_pred = clf.predict(X_test)
accuracies.append(accuracy_score(y_test, y_pred))
# Step 6: Evaluate model performance
print(f"Cross-Validated Accuracy: {sum(accuracies) / len(accuracies) * 100:.2f}%")
```

⇨ `Cross-Validated Accuracy: 86.67%`

The Bootstrap

- The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- As a simple example, the bootstrap can be used to estimate the standard errors of the coefficients from a linear regression fit.
- Bootstrapping is an invaluable technique in machine learning for understanding model performance and addressing uncertainty.
- By generating multiple samples from the same dataset, bootstrapping helps estimate metrics like accuracy and feature importance, providing insights into model reliability and stability.

Bootstrap: Start by an example

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y, respectively, where X and Y are random quantities.
- We will invest a fraction α of our money in X, and will invest the remaining $1 - \alpha$ in Y
- Since there is variability associated with the returns on these two assets, we wish to choose a to minimize the total risk, or variance, of our investment. In other words, we want to minimize Var(α X +(1− α)Y).
- One can show that the value that minimizes the risk is given by

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

$$\sigma_X^2 = \mathrm{Var}(X), \sigma_Y^2 = \mathrm{Var}(Y), \text{ and } \sigma_{XY} = \mathrm{Cov}(X, Y)$$

Bootstrap: Start by an example:

estimating for α

- In reality, the quantities σ_Y^2, σ_X^2,and σ_{XY} are unknown. We can compute estimates for these quantities $\hat{\sigma}_Y^2$,,$\hat{\sigma}_X^2$ and $\hat{\sigma}_{XY}$, using a data set that contains past measurements for X and Y

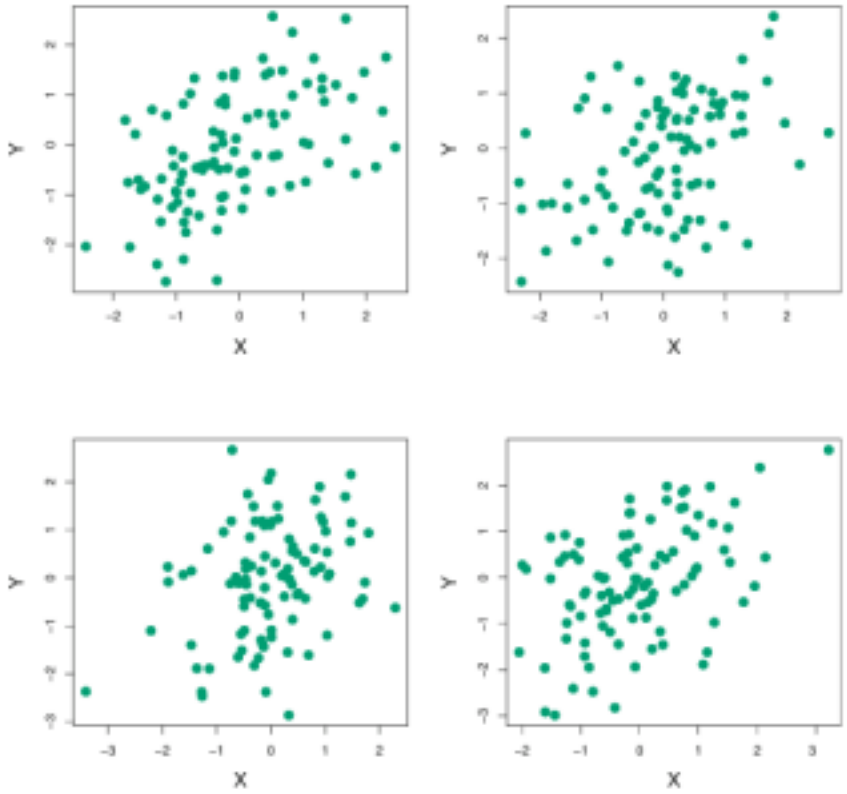
$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}.$$

Bootstrap: Start by an example:

simulated data set

- The figure illustrates this approach for estimating α on a simulated data set.

In each panel, we simulated 100 pairs of returns for the investments X and Y



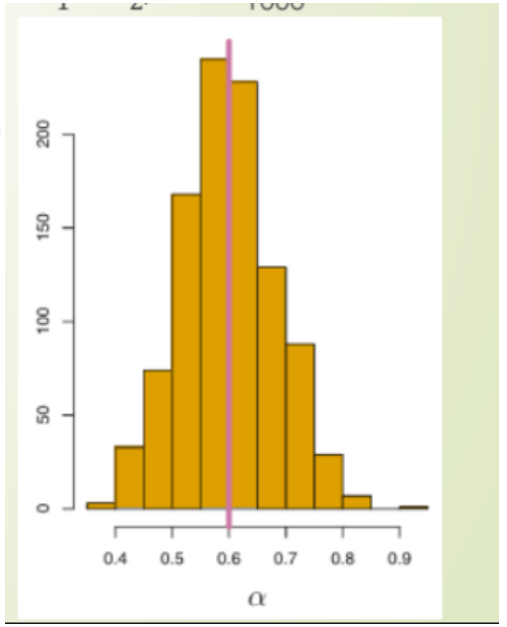
From left to right and top to bottom, the resulting estimates for α are 0.576, 0.532,0.657, and 0.651.

Bootstrap: Start by an example: the accuracy of α

- It is natural to wish to quantify the accuracy of our estimate of α. To estimate the standard deviation of α we repeated the process of simulating 100 paired observations of X and Y , and estimating α using the above formula 1000 times. We there by obtained 1,000 estimates for α, which we can call α_1, α_2, ..., α_{1000}. For these simulations the parameters were set to

$$\sigma_X^2 = 1, \sigma_Y^2 = 1.25, \text{ and } \sigma_{XY} = 0.5$$

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



A histogram of the estimates of a obtained by generating 1,000 simulated data sets from the true population.

Bootstrap: Start by an example: the accuracy of α^{**} (cont.)

- The standard deviation of the estimates is

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

- This gives us a very good idea of the accuracy of α^{**} :

$$\mathrm{SE}(\hat{\alpha}) \approx 0.083.$$

Bootstrap: Start by an example: the bootstrap approach

- In practice, however, the procedure for estimating SE(□□) outlined above cannot be applied, because for real data we cannot generate new samples from the original population.
- However, the bootstrap approach allows us to use a computer to emulate the process of obtaining new sample sets, so that we can estimate the variability of □□ 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.

Bootstrap: The sampling method

- This approach is illustrated in figure on a simple data set,

  which we call Z, that contains only n = 3 observations.

  We randomly select n observations from the data set in

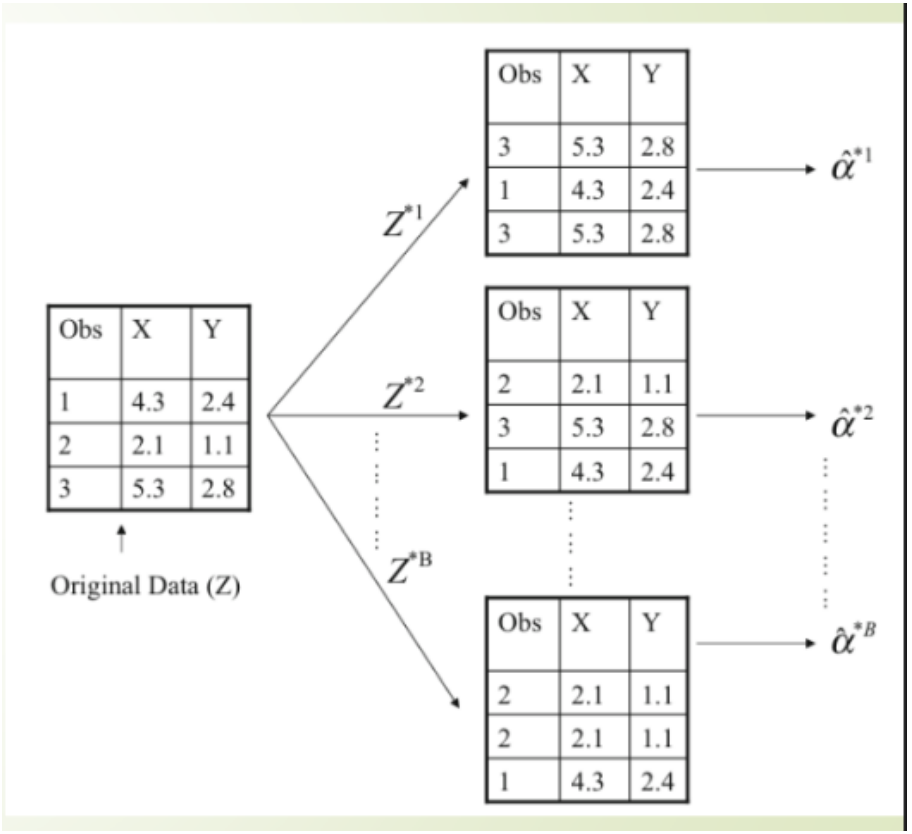
  order to produce a bootstrap data set, Z^{*1}.

- The sampling is performed with replacement, which

  means that the same observation can occur more than

  once in the bootstrap data set.

- This procedure is repeated B times for some large value of B,



$$Z^{*1}, Z^{*2}, \ldots, Z^{*B}$$

$$\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \ldots, \hat{\alpha}^{*B}$$

in order to produce B different bootstrap data sets:

- B corresponding a estimates:

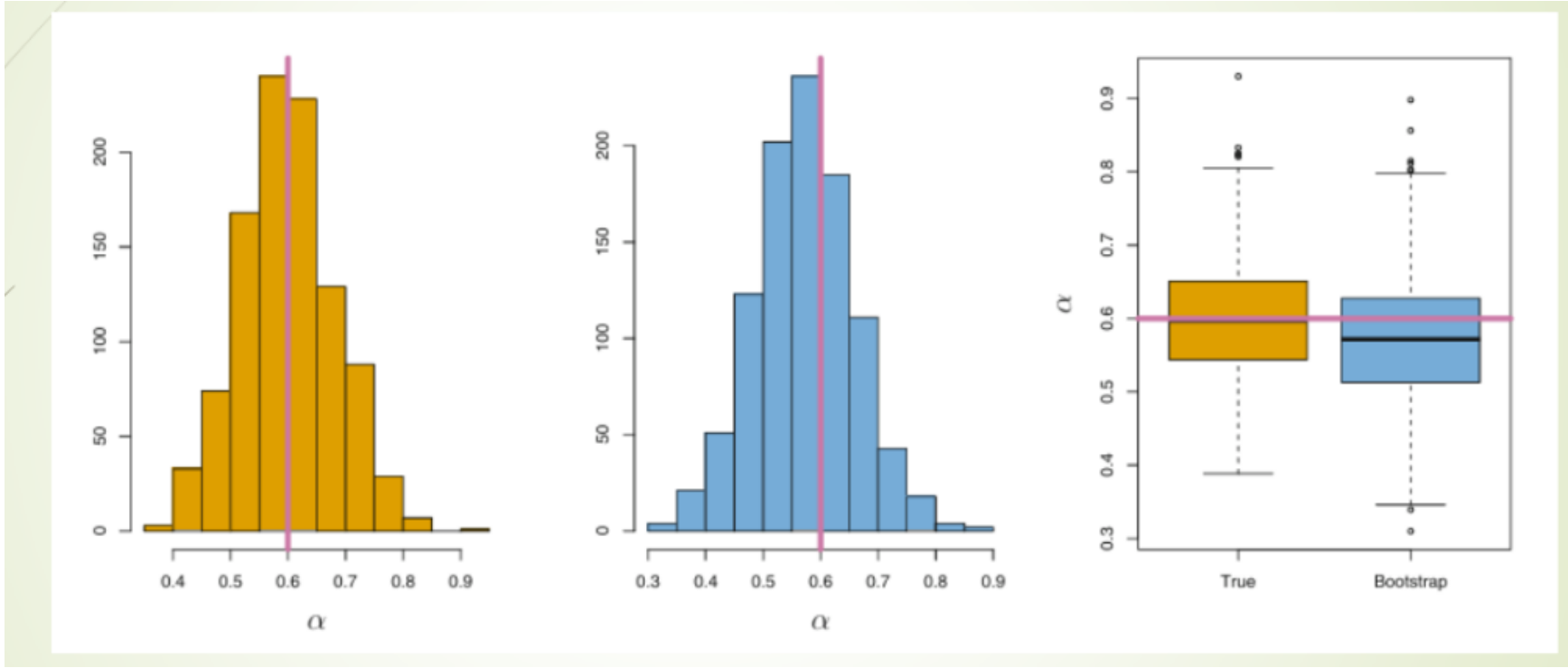
Bootstrap: The standard error of □□ estimated from the original data set

- We can compute the standard error of these bootstrap estimates using the formula

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

- This serves as an estimate of the standard error of □□ estimated from the original data set.

Bootstrap: The Histograms



Left: A histogram of the estimates of a obtained by generating 1,000 simulated data sets from the true population. Center: A histogram of the estimates of a obtained from 1,000 bootstrap samples from a single data set. Right: The estimates of a displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of a

Bootstrap: Discussion

- The center histogram looks very similar to the left-hand panel which displays the idealized histogram of the estimates of α obtained by generating 1,000 simulated data sets from the true population.
- The bootstrap estimate SE(□□) from SE_B(□□) is 0.087,very close to the estimate of 0.083 obtained using1,000 simulated data sets.
- The boxplots are quite similar to each other, indicating that the bootstrap approach can be used to effectively estimate the variability associated with □□ .
- **This center panel was constructed on the basis of a single data set, and hence could be created using real data.**
  - Keep in mind that bootstrapping does not create new data. Instead, it treats the original sample as a proxy for the real population and then draws random samples from it.

Homework

1. Implementing the example program in the slide #15 (all done)

2. Generating different bootstrapped samples of a dataset, the dataset at your disposal.

End of Lesson