Chapter 5 Resampling Methods

Resampling methods: 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.

1. Resampling approaches can be computationally expensive, because they involve fitting the same statistical model multiple times using different subsets of the training data.

Cross-validation:

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

1. Model Assessment: Evaluate a model’s performance
2. Model Selection: The process of selecting the proper level of flexibility

5.1 Cross-Validation

Test error: The average error that results from using a statistical learning method to predict the response on a new observation

\*A measurement that was not used in training the method

\*Training Error rate can dramatically underestimate the latter

Cross Validation Approach: 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

5.1.1 The Validation Set Approach

Validation set approach: Randomly dividing the available set of observations into two parts, a training set and a validation set/hold-out set

1. The model is fit on the training set, and the fitted model is used to predict response for the observations in the validation set.
2. Split the observations into two sets-a training set and a validation set.
3. The validation set error rates(using MSE as a measure of validation set error): Resulting from fitting various regression models on the training sample and evaluate their performance on the validation sample

Drawbacks:

1. The validation estimate of the test error rate can be highly variable, depending on precisely which h observations are included in the training set and which observations are included in the test set.
2. Only a subset of the observations-those that are included in the training set rather than in the validation set- are used to fit the model

\*Validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set.

5.1.2 Leave-One-Out Cross-Validation(LOOCV)

LOOCV involves splitting the set of observations into two parts:

1. A single observation ( is used for the validation set,
2. The remaining observations make up the training set.
3. The statistical learning method is fit on the n-1 training observations, and a prediction is made for the excluded
4. Since was not used in the fitting process, provides an approximately unbiased estimate for the test error.

\*Even though is unbiased for the test error, it is a poor estimate because it is highly variable(based upon a single observation)

A close up of a clock

Description automatically generated

\*LOOCV estimate for the test MSE is the average of the n test error estimates

Advantages of LOOCV over the validation set approach:

1. Has far less bias(Repeatedly fit the statistical learning method using training sets that contain n-1 observations, almost as many as re in the entire data sets)🡪Tends not to overestimate the test error rate as much as the validation set approach
2. Performing LOOCV multiple times will always yield the same results: there is no randomness in the training/validation set splits
3. With least squares regression, linear and polynomial regression, a short cut that makes he cots of LOOCV the same as that of a single model fit:

A close up of a clock

Description automatically generated

5.1.3 k-Fold Cross Validation

1. Randomly dividing the set of observations into k groups, or folds, of approximately equal size.
2. The first fold is treated as a validation set, and the method is fitted on the remaining k-1 folds.
3. The mean squared error is calculated for the held-out fold.
4. The procedure is repeated k times; each time, a different group of observations is treated as a validation set.

A close up of a clock

Description automatically generated

1. LOOCV is a special case of k-fold CV in which k is set to equal to n.
2. In practice, one typically performs k-fold cv using k=5 or k=10

Advantages of k fold CV:

1. Less computationally intensive
2. CV is a very general approach that can be applied to almost any statistical learning method
3. The variability is typically much lower than the variability in the test error estimates that results from the validation set approach

Applying k-fold Cross Validation:

1. How well a given statistical learning procedure can be expected to perform on independent data🡪 Actual estimate of the test MSE is of interest
2. The location of the minimum point in the test MSE curve🡪Performing cross-validation on a number of statistical learning methods, or on a single method using different levels of flexibility, in order to identify the method that results in the lowest test error.

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

K-Fold CV often gives more accurate estimates of the test error rate than does LOOCV🡨 Bias Variance Trade-Off

1. Using k-fold will give intermediate level of bias between LOOCV and Validation Ser Approach
   1. Each training set contains fewer observations than validation set but more than LOOCV
2. The test error estimates resulting from LOOCV tends to have higher variance than does the test error estimate resulting from k-fold CV

5.1.5 Cross-Validation on Classification Problem

LOOCV error rates under classification setting:

A picture containing object, clock

Description automatically generated



In practice, when the Bayes decision boundary and the test error rates are unknow, we can use cross-validation in order to make the decision.

1. The training error tends to decrease as the flexibility of the fit increases
2. The test error displays a characteristic U-shape.

5.3 The Bootstrap

Used to quantify the uncertainty associated with a given estimator or statistical learning method.

\*The bootstrap can be sued to estimate the standard errors of the coefficients from a linear regression fit.

\*Can be easily applied to a wide range of statistical learning methods

Example: Investing in two financial assets that yield returns of , fraction of money invested in . Fraction of money invested in .

1. We choose to minimize the total risk, or variance, of our investment.

A close up of a logo

Description automatically generated



We can compute estimates for the population metrics using using a data set that contains past measurements for and

A close up of a logo

Description automatically generated

To estimate the standard deviation of , repeat the process of simulating 100 paired observations of X and Y, and estimating 1000 times

Bootstrap approach allows us to use computer to emulate the process of obtaining new sample sets:

1. We can estimate the variability of without generating additional samples.
2. Randomly select n observations from the data set in order to produce a bootstrap data set
   * Performed with replacement🡪 Same observation can occur more than once in the data set
   * The standard error of the bootstrap estimate is

A close up of a clock

Description automatically generated

\*This standard error serves as the standard error of estimated from the original data set

5.3 Lab: Cross-Validation and the Bootstrap

5.3.1 The Validation Set Approach

1. Generally a good idea to set a random seed when performing an analysis such as cross-validation that contains an element of randomness

🡪Results obtained can be reproduced precisely at a later time

Sample(x,size): random sample and permutations, sample takes a sample of the specified size from the elements of x using either with or without replacement.

A picture containing table

Description automatically generated





1. Use predict to estimate the response of all observations
2. Mpg-predicted response
3. -train index selects response that are not in the training set
4. Squared and use the mean function to calculate the MSE

5.3.2 Leave-One-Out Cross Validation

LOOCV estimate can be automatically computed for any generalized linear model using glm() and cv.glm() functions

Glm() without passing in the family argument performs linear regression, just like the lm() function.

A picture containing knife, table

Description automatically generated

cv.glm() produces a list with several components:

1. The two numbers in the delta vector contain the cross-validation results

\*First is the standard k-fold CV estimate

\*Bias Corrected Version

\*Essentially the same of LOOCV, but may be different for k-fold CV

A picture containing knife, table

Description automatically generated

5.3.3 k-Fold Cross-Validation

cv.glm() contains an additional argument K= that can be used in k-fold cross validation

1. Computation time is much shorter than LOOCV

A picture containing bird

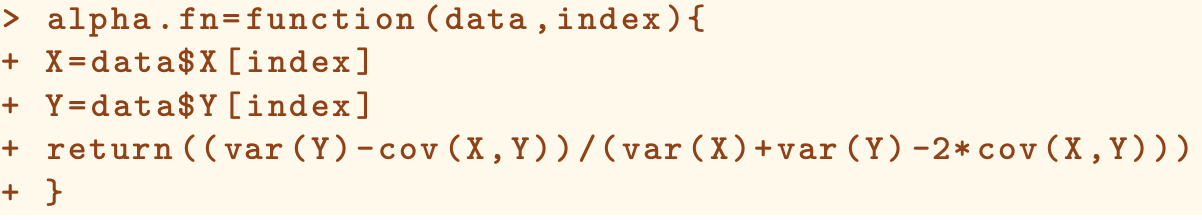
Description automatically generated

5.3.4 The Bootstrap

Bootstrap approach can be applied in almost all situations

Perform Bootstrap In R:

1. Must choose a function that computes the statistic of interest
2. Use the boot() function, part of the boot library, to perform the bootstrap by repeatedly sampling observations from the data set with replacement.



1. Create alpha.fn() which takes as input the (X,Y) data as well as a vector indicating which observations should be used to estimate
2. The function returns an output, an estimate for based on the observations indexed by the argument index.

A picture containing knife, table

Description automatically generated

1. Use sample() to randomly sample 100 observations from size 100 with replacement
2. Equivalent to construct a new bootstrap data set

Boot()L function automates the sample approach

A screenshot of a cell phone

Description automatically generated

1. R=1000, number of bootstrap samples produced
2. Bootstrap statistics shows both outputs and the standard errors

Estimating the Accuracy of a Linear Regression ModA screenshot of a cell phone

Description automatically generated

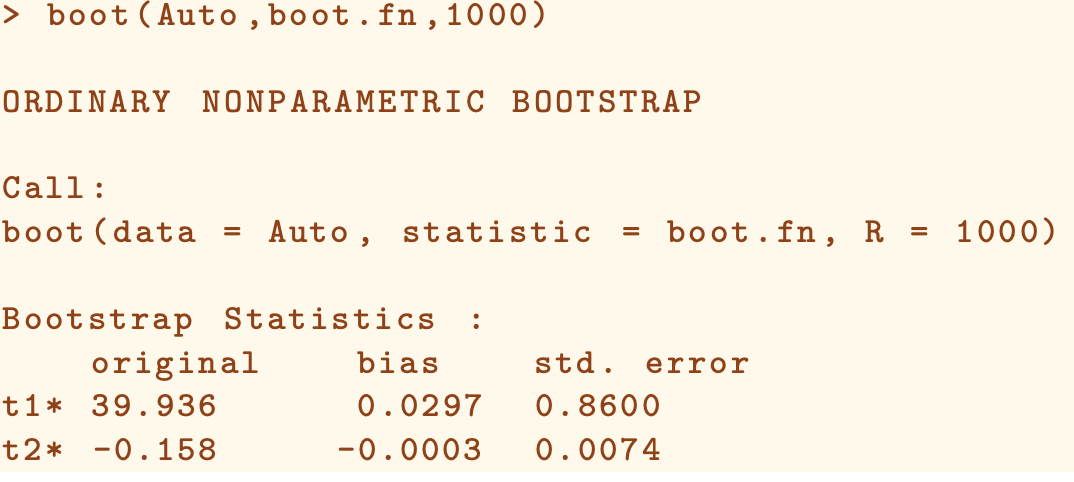
\*Create boot.fn(data,index) that takes the data for prediction and response and index for specifying the training set

\*Returns a vector of the coefficients of the fitted model

A picture containing bird

Description automatically generated

1. Create bootstrap estimates for the intercept and the slope terms by randomly sampling from among the observations with replacement



The approach does not rely on any of the assumptions with standard error estimate calculation, it is more likely giving a more accurate estimate of the standard errors of in summary function