Claremont Graduate University, Mathematics M454, Statistical Learning
Lab 2
Instructor: Qidi Peng
TA: Nan Rao
Feb. 27, 2018

CH 3: Resampling Methods and CH 4: Forecasting Strategies

1. **(LOOCV for Linear or polynomial regressions)** Consider a multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon.$$

Suppose x ($n \times (p+1)$ matrix of real values) is the training predictors and y ($n \times 1$ matrix of real values) is the training response. The least squares estimates of $\beta = (\beta_0\ \beta_1 \ldots\ \beta_p)^T$ is given as

$$\hat{\beta} = (x^T x)^{-1} x^T y,$$

where x^T denotes the transpose of x. Therefore the estimates of y is given as

$$\hat{y} = x\hat{\beta} = \underbrace{(x(x^T x)^{-1} x^T)}_{\text{denoted by } S}\ y.$$

Then it is known that the LOOCV prediction MSE (also called test MSE) is

$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - S_{ii}}\right)^2$, the leverage S_{ii} is the ith diagonal element of the square matrix S.

Questions.

(a) Create a function to calculate $CV_{(n)}$, with input arguments (x, y). Here x is an $n \times (p+1)$ matrix (its first column must be 1s) $y = (y_1, \ldots, y_n)$ is a vector.
(Hint: In R, $S[i,i]$ is the i-th diagonal element of the square matrix S. Matrix product of A and B: $A\% * \%B$; inverse matrix A: $solve(A)$; transpose of matrix A: $t(A)$.)

(b) (Real world project: polynomial regression) An experiment is designed to relate three variables (X_1 =temperature, X_2 =ratio, and X_3 =height) to a measure of odor in a chemical process. Each variable has three levels, but the design was not constructed as a full factorial design (i.e., it is not a 3^3 design). Nonetheless, we can still analyze the data using a response surface regression routine, which is essentially polynomial regression with multiple predictors. The data obtained **(odor.txt)** is available in **Canvas→Files→Labs**.

We consider 2 polynomial regression models to fit the relationship between X and Y:

Model 1: $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_1^2 + \alpha_5 X_2^2 + \alpha_6 X_3^2 + \epsilon$;
Model 2: $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_1^2 + \beta_4 X_2^2 + \epsilon$.

Please download **"odor.txt"** from **Canvas → Files → Labs** and do the following:

1 Read this data in R using for example (replace my file path below with yours).

```
> odor_data<-read.table("C:/Users/pengq/Desktop/odor.txt",header=T);
```

2 Visualize the data set.

```
> fix(odor_data);
```

3 Convert data frame to numeric matrix.

```
> odor_data<-data.matrix(odor_data);
```

Use the function you created in (a) to calculate the LOOCV prediction MSE of Model 1 and Model 2. Which model would you prefer?
(Hint: the key is to create the right matrix X for each model.)

2. **(Bootstrap) In this section, please use** $set.seed(1)$ **before you generate random numbers.** Consider the "**Boston**" housing data set, from the package "**MASS**".

Questions.

(a) Find the definition (or meaning) of the vector **tax** in the housing data set "**Boston**".
(Hint: use $help(Boston)$.)

(b) Calculate the median value of the data **tax** in the data frame "**Boston**".
(Hint: use *median()* function in **R** to calculate sample median.)

(c) We denote by $\hat{\mu}_{tax}$ the estimate of μ_{tax}, the median of the entire Boston area **tax** (while the tax in "**Boston**" is only one sample). Estimate the standard error $SE(\hat{\mu}_{tax})$ using 1000 bootstrap samples.
(Hint: first generate a function $Se(X, B)$ to return the bootstrap estimator of the standard error of the sample X, using B bootstrap samples. You will use $sample(X, n, replace = TRUE)$ to generate ONE bootstrap sample.)

(d) Suppose that $\hat{\mu}_{tax}$ is asymptotically a normal random variable with mean μ_{tax}. Provide the 95% confidence interval of the true median.
(We remark that without bootstrapping it is hopeless to find $SE(\hat{\mu}_{tax})$ in this case.)

3. **(Generalized linear regression)** Load the data set "crab.txt" from Canvas files and name every column as in classnotes LAB2. Use Poisson model to fit the relationship between Sa and (W, Wt) (here are 2 predictors). Provide the estimates of all coefficients.

4. **(Real world project: Bass model)** In this project we model the U.S. sales figures of 2 Toyota series "Camry" and "FJ Cruiser" by Bass curve. The result shows that Camry has a much better sales record than FJ Cruiser does. In fact, FJ Cruiser has been discontinued, making the 2014 FJ Cruiser the last model year. However, excitement and capability live on with adventure-ready Toyota vehicles like 4Runner and the off-road-ready TRD Pro Series (this is not an Ad, but telling you that market strategy should be and has been adjusted). Before running Base model, get your data set prepared. The data set[1] is named by **"ToyotaSales.txt"**. Please download it from **Canvas** → **Files** → **Labs** and do the following:

1 Read this data in R using for example (replace my file path below with yours).

```
> Sales<-read.table("C:/Users/pengq/Desktop/ToyotaSales.txt",header=T);
```

2 Visualize the data set.

```
> fix(Sales);
```

3 Extract the second column from **Sales** by using

```
> Camry<-Sales$CamrySales
```

4 Extract the third column from **Sales** and remove missing data N/A.

```
> Cruiser<-Sales$FJCruiserSales;
> Cruiser<-Cruiser[-c(1,2,3,4)];
```

[1] All real data is obtained at $http : //www.goodcarbadcar.net/2011/01/toyota - camry - sales - figures.html$ and $http : //www.goodcarbadcar.net/2011/01/toyota - fj - cruiser - sales - figures.html$.

5 **Cruiser** is a set of factors, because it contained N/A. We convert it to numeric.

> Cruiser<-as.numeric(paste(Cruiser));

Project exercises.

(a) Use Bass curve to fit **Toyota**. Give estimates of its m, p, q and provide its Bass curve. Does Bass model fit **Camry**?
(Hint: it is super difficult to guess the initial values for nonlinear least squares, so please use the function $nlsLM()$ in package "minpack.lm", in lieu of the function $nls()$. Put initial values $P = 0.5$, $Q = 0.65$. Of course you can try other initial values.)

(b) Use Bass curve to fit **Cruiser**. Give estimates of its m, p, q and provide its Bass curve. Compare the 2 Bass curves and conclude: which series has better sales performance?