Homework 2

Due Date 10/02/2018

Q1. Chapter 5 Exercise 1

Q2. Chapter 5 Exercise 2

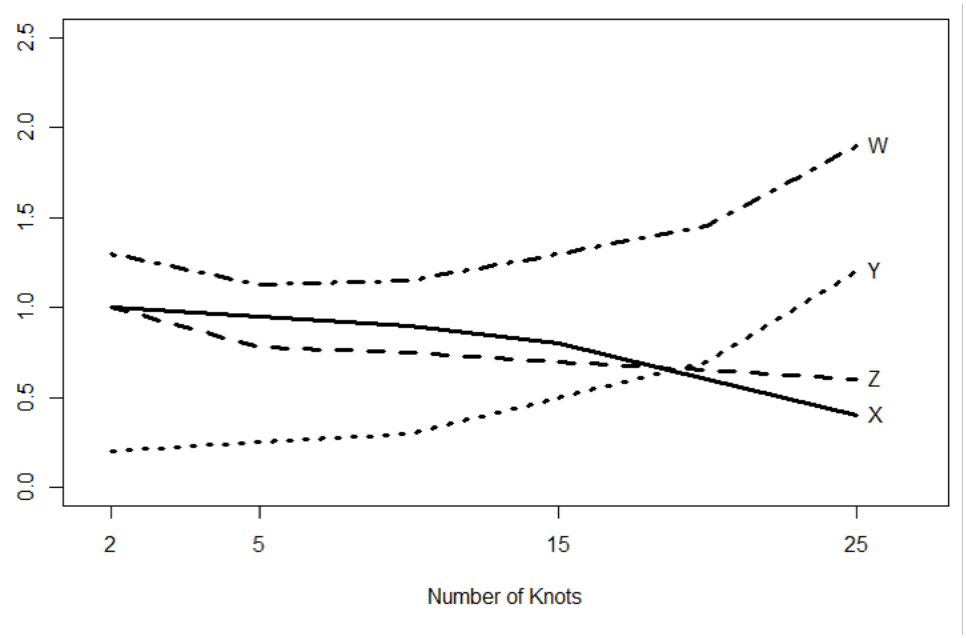
Q3. Chapter 5 Exercise 3

Q4. Chapter 5 Exercise 8

Q5. Suppose we want to compute 10-Fold Cross-Validation error on 100 training examples. We need to compute error n_1 times, and the Cross-Validation error is the average of the errors. To compute each error, we need to build a model with data of size n_2 and test the model on the data of size n_3. What are the appropriate numbers for n_1, n_2 and n_3?

We will need to compute the error n1=10 times for 10 folds. We will also need build a training model of n2= 9 observations and a test data observation size of n3 =1. Testing the data on this training model will result in a error estimate for 1 fold. Do this for all ten folds and we will estimate n1=10 errors.

Q6. You want to fit a cubic spline to a large dataset and need to determine the number of knots to use. Below is a chart of four statistics from this model valued for various numbers of knots:



Determine which set of statistics below best describes each line and why?

A. W is Test MSE; X is Variance; Y is Squared Bias; Z is Train MSE

B. W is Variance; X is Squared Bias; Y is Test MSE; Z is Train MSE

C. W is Train MSE; X is Test MSE; Y is Variance; Z is Squared Bias

D. W is Test MSE; X is Train MSE; Y is Variance; Z is Squared Bias

E. W is Variance; X is Train MSE; Y is Test MSE; Z is Squared Bias

This can be seen that of course the test MSE is W because the test error is always higher than the training. Additionally X is the Train MSE because we know that the more data the training has the more fit it will be and cause over fitting which also informs that the bias squared is Z, leaving Y as the variance which improves with flexibility.

CODE:

```
#Chapter 5
#exercise 2 part g


x = 1:100000
y = (1-(1-1/x)**x)
plot(x,y)

#it can be seen from the plot that it shows the curve reaching a limit
#which is approximately .6323


#exercise part h
store <- rep(NA, 10000)
for (i in 1:10000){
```

```
  store[i] <- sum(sample(1:100, rep=TRUE ) == 4) >0
}
mean(store)
#The result was .632
#this shows that the average value of the 4th observation was .632
#Also helps show that indeed we are hitting a limit of about .632



#exercise 8
#given
set.seed(1)
y<- rnorm(100)
x<- rnorm(100)
y <-x - 2*x^2+rnorm(100)

#part a
#the n is 100, p= 2
#model is Y = X -2X^2 + epsilon

#part b

plot(x,y)
#There seems to be a non-linear relationship
library(boot)
set.seed(1)
```

```
data <- data.frame(x,y)
model.1 <- glm(y~x)
cv.glm(data, model.1)$delta[1]

model.2 <- glm(y~poly(x,2))
cv.glm(data, model.2)$delta[1]

model.3 <- glm(y~poly(x,3))
cv.glm(data, model.3)$delta[1]

model.4 <- glm(y~poly(x,4))
cv.glm(data, model.4)$delta[1]
#part d
set.seed(333)


data <- data.frame(x,y)
model.1 <- glm(y~x)
cv.glm(data, model.1)$delta[1]

model.2 <- glm(y~poly(x,2))
cv.glm(data, model.2)$delta[1]

model.3 <- glm(y~poly(x,3))
cv.glm(data, model.3)$delta[1]

model.4 <- glm(y~poly(x,4))
```

```
cv.glm(data, model.4)$delta[1]
#the results are the same because LOOCV
#analyses the n folds of a single observation



#part e
#the smallest error was model 2 which make sense since the relationship looks quadratic
#this result confirms indeed that the realtion is quadratic

#part f

summary(model.4)
#based on the above results it can be seen that the Probabilites for linear and
#quadratic terms are stastically significant from the P-value and for the 3rd and 4th degree terms have no significance
#this backs up the claim that this relationship is indeed quadratic and agrees with the CV results
```