Assignment 10 (25 points)

12/05/17

Notes:

- This homework assignment is due December 11th 2017.
- It has two parts which count for 10 and 30 points
- The home work is marked out of 25 points - therefore you can get up to an additional 15 bonus points

Part 1: (10 Points)

1. (2 points) Problem 9.7 #12, in Chihara/Hesterberg.
2. (2 points) Problem 9.7 #14, in Chihara/Hesterberg.
3. (2 points) Problem 9.7 #15, in Chihara/Hesterberg.
4. (2 points) Problem 9.7 #17, in Chihara/Hesterberg.
5. (2 points) Problem 9.7 #20, in Chihara/Hesterberg.

Part 2: (30 Points)

In this exercise set you will be going over the steps of building and interpreting a simple and multiple regression model in `R`. You will be analysing the Boston Housing Datset, the schema for the data set can be found at: https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Boston.html

To start first load the data set in `R` using the command:

```
library(MASS)
data(Boston)
head(Boston, n=5)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
##   lstat medv
## 1  4.98 24.0
```

```
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
```

Take a minute to explore the data.

Simple linear model (12 Points)

Next we will be building a simple linear model to compare the median cost of a house (in $1000s) to the average room size.`{R, eval=F} medv_model<- lm(medv~rm+lstat+dis, data=medv_subset)`

a) (1 Point) Plot a 2d scatterplot of `medv`(dependent variable) vs `rm` (independent variable)
b) (1 Point) What do you notice about the slope, is it positive or negative? Do you think it will pass through 0?
c) (3 Points) Using the function `lm`:
    i) Find the slope and intercept for the model.(*Remember that the dependent variable in the formula is on the write side of the tilde:*`y x`)
    ii) Plot the linear model on your scatterplot (you can do this using the function `abline(your model in here)`)
    iii) What is the interpretation of the slope? How about the intercept?
d) (3 Points) Using the function `residuals`
    i) Find the residuals of the fitted model
    ii) Plot a histogram and q-q plot for the residuals
    iii) Based on ii) do the residuals look normally distirbuted?
        - If not, what are some of the things we could do to identify points that don't fit our normal assumptions?
        - If yes, what does that imply about the model
e) (4 Points) Using the `summary` command you can pull-out additional data about your linear model.
    i) Use `summary` to identify the p-values for the intercept and and slope constants, are they statistically significant?
    ii) What is the `Mulitple R Squared` for the model? What does it mean?
    iii) What is the F-Statistic for the model? Does it contradict the `Mulitple R Squared`?

Polynomial Regression (10 Points)

We will next assess a polynomial fit.

a) (2 Points) Plot Nox vs Dis. Is it a linear fit? If not, what kind of fit does it look like.
b) (5 Points) We are going to assess a whether a quadratic fit is appropriate, using the code below fit a quadratic model and plot the resulting curve on the scatterplot.

i) Does the fit look quadratic?
ii) What is the interpretation of the parameters
iii) What are the p-values for the different parameters, are they all important?
iv) What are the multiple R^2 and F statistic?
v) Looking at the plot what might be one of the risks if we go beyond a Dis of 12?

```
# Scatterplot
plot(Boston$dis,Boston$nox, main='Dis vs. Nox',
     xlab='Dis', ylab='Nox', pch=16, cex=0.5)
# Setting up model
m2 <- lm(nox ~ poly(dis,2), data=Boston)
# New data for prediction + plotting
newdata <- data.frame(dis=seq(0,12,by=0.1))
predicted_value <- predict(m2, newdata = newdata)
# Plotting the cruve
points(unlist(newdata),predicted_value, col='red', type='l')
```

c) (3 Points) Alter the code to above to fit a cubic (polynomial of degree 3)
   i) Based on the multiple R^2 is it a better fit than a quadratic?
   ii) Check to see that the residuals are normally distributed using a q-q plot and histogram

Multiple regression (8 Points)

Let's next tackle multiple regression.

a) (2 Points) Use the function `pairs` to plot the pair-pair plot. What variables look correlated with `age`?

```
pairs(Boston, cex=0.4, pch=15)
```

b) (3 Points) Suppose we wanted to predict `medv` using the variables `lstat`,`rm` and `dis`. Use the following code to subset to generate pair-pair plots for the variables
   i) From the previous question are `rm` and `lstat` correlated? If yes, explain what kind of issues this make cause when modeling. If no, show using a simple linear model that there is not enough evidence for a linear relation (i.e. you need to show that β_1's p-value is not significant).
   ii) Fit a linear model with all three variables to `medv`. What is it's adjusted R^2 squared.
   iii) Next fit a model where we take the `log(dis)`
   - How does it's adjusted R^2 compare, is it better than the simple linear transformed model.
   - Plot the histogram of residuals, does it look normal?

```
# --- Part b.i ---
medv_subset <- Boston[,c('medv','rm','dis','lstat')]
```

```r
pairs(medv_subset, cex=0.4, pch=15)
```

```r
# --- Part b.ii ---
medv_model<- lm(medv~rm+lstat+dis, data=medv_subset)
```

```r
# --- Part b.iii ---
medv_model<- lm(medv~rm+lstat+log(dis), data=medv_subset)
```

c) (3 Points) Finally we'll fit a model without an intercept
   i) In the previous models was the intercept statistically significant?
   ii) Fit a linear model without an intercept, what is it's resulting adjusted R^2
   iii) Plot the histogram of residuals. How does it compare to the histogram in part b.iii)

```r
medv_model<- lm(medv~rm+lstat+dis - 1, data=medv_subset)
```