Homework 1

NAME: Your Name
DUE DATE: February 15th, 5pm

It is important that you learn to write up your results carefully. For each problem assigned for homework, you should carefully describe and justify the analytic methods used and summarize key findings in carefully written English with reference to appropriate tables and figures as needed.

Problem 1 [ISL 2.1] (4 pts)

For each parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

a) The sample size n is extremely large, and the number of predictors p is small.

b) The number of predictors p is extremely large, and the number of observations n is small.

c) The relationship between the predictors and response is highly non-linear.

d) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

Problem 2 [ISL 2.2] (3 pts)

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry, and the CEO salary. We are interested in understanding which factors affect CEO salary.

b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

c) We are interested in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week, we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

Problem 3 (7 pts)

In this problem, you will learn how to analyze linear regression coefficients.

a) Assume that $Y = \beta^T X + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. What is the distribution of $\hat{\beta}|X$?

b) Given the distribution in part (a), write the formula for a 95% confidence interval for each β_i. Recall that σ^2 is unknown.

c) Describe how you would use hypothesis testing to test whether there is a relationship between the X_i and Y (i.e. $\beta_i \neq 0$).

Problem 4 (8 pts)

In this problem, you will explore type 1 and type 2 errors with a simple regression.

a) Simulate a random vector y of length 100 in which each element follows a normal distribution with mean 10 and variance 4. Then simulate a random vector x of length 100 with mean 3 and variance 1.

b) Regress y on x and find the p-value for the slope (Hint: use the **summary.lm** function or **summary(lmobject)** and use the coef element of the summary object if using **R**)

c) Now replicate this procedure 1000 times. Calculate the proportion of times the p-value is less than 0.05. How does this match with your intuition? [Hint: Should the slope be related to the outcome? What type of error are you simulating?]

d) Now simulate y and x together so that the conditional mean of y given x is 10+x and the conditional variance is 1. Repeat b-c above for this distribution. How does your answer differ from the first simulation? What are you simulating now?

Problem 5 (8 pts)

Now we will investigate overfitting which occurs when you are making a model too complex. First we will simulate some data in which y and multiple x's are unrelated.

a) Extend the function you have written to carry out exercise 1 so that you now generate the same y vector as in Problem 4 (i.e. 100 draws from $N(10, 4)$) but you now generate 100 draws from an 3-variate multivariate normal X matrix with means 1, 2, 3 and covariance matrix equal to the identity matrix of rank 3 (i.e., each X variable has variance 1 and they are uncorrelated). [Hint: use the **mvrnorm** function in the MASS library to generate multivariate normal draws].

b) Regress y on the X matrix and save the p-values from this regression.

c) Over 1000 iterations, determine the minimum p-value for each regression and calculate how many times the minimum p-value is less than 0.05. How does this match your intuition?

d) Now repeat this exercise setting the number of predictors p to every number from 1 to 99. Compute the minimum p-value in each simulated regression and plot the number of times you find at least one significant variable for each p. What pattern do you see? How do you explain this?

Problem 6 [Based on ISL 2.8] (10 pts)

This exercise relates to the *College* data set, which can be found in the file *College.csv* (http://www-bcf.usc.edu/~gareth/ISL/data.html). It contains a number of variables for 777 different universities and colleges in the US. The variables are

- **Private** : Public/private indicator
- **Apps** : Number of applications received
- **Accept** : Number of applicants accepted
- **Enroll** : Number of new students enrolled
- **Top10perc** : New students from the top 10% of high school class
- **Top25perc** : New students from the top 25% of high school class
- **F.Undergrad** : Number of full-time undergraduates
- **P.Undergrad** : Number of part-time undergraduates
- **Outstate** : Out-of-state tuition

- **Room.Board** : Room and board costs
- **Books** : Estimated book costs
- **Personal** : Estimated personal spending
- **PhD** : Percent of faculty with Ph.D.'s
- **Terminal** : Percent of faculty with terminal degree
- **S.F.Ratio** : Student/faculty ratio
- **perc.alumni** : Percent of alumni who donate
- **Expend** : Instructional expenditure per student
- **Grad.Rate** : Graduation rate

Use the following commands to read and look at the data. You may need to change the path to the csv file.

```
college = read.csv("College.csv")
fix(college)
```

You should notice that the first column is just the name of the university. We don't really want **R** to treat this as data. Instead, we will name the rows by these values.

```
rownames(college) = college[,1]
college = college[,-1]
fix(college)
```

a) Use the **summary()** function to produce a numerical summary of the variables in the data set.

b) Use the **pairs()** function to produce a scatterplot matrix of the first ten columns or variables of the data.

c) Use the **plot()** function to produce side-by-side boxplots of **Outstate** versus **Private**.

d) Create a new qualitative variable, called **Elite**, by binning the **Top10perc** variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

   ```
   Elite = rep("No",nrow(college))
   Elite[college$Top10perc >50] = "Yes"
   Elite=as.factor(Elite)
   college=data.frame(college,Elite)
   ```

   Use the **summary()** function to see how many elite universities there are. Now use the **plot()** function to produce side-by-side boxplots of **Outstate** versus **Elite**.

e) Continue exploring the data, and provide a summary of what you discover. Be sure to include at least a couple of figures using the **ggplot2** package, which is part of the **tidyverse** library.