Stat 627/427 (Statistical Machine Learning)

Homework # 2: Regression and R. Chap. 3 (3.1-3.2) and Chap. 2 (Lab sec. 2.3)

1. (#3-1, page 120). The following table contains results of linear regression analysis of *Advertising* data. It was used to model number of units sold as a function of radio, TV, and newspaper advertising budgets.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Describe the null hypotheses to which the given p-values correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

2. (#3-3, page 120). Suppose we have a data set with five predictors, X_1 =GPA, X_2 =IQ, X_3 =Gender (1 for Female and 0 for Male), X_4 =Interaction between GPA and IQ, and X_5 =Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.

   (a) Which answer is correct, and why?

      i. For a fixed value of IQ and GPA, males earn more on average than females.

      ii. For a fixed value of IQ and GPA, females earn more on average than males.

      iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

      iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

   (b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

   (c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

3. (#3-4, pages 120-121). I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$.

   (a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \varepsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

   (b) Answer (a) using test rather than training RSS.

   (c) Suppose that the true relationship between X and Y is not linear, but we dont know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) Answer (c) using test rather than training RSS.

4. (R project, #2-8, p.54-55, let me know if you need more time for it) This exercise relates to the **College** data set from our textbook. 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 top 10% of high school class
Top25perc	New students from top 20% 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

(a) Read the data into R, for example, by the **load(”College.rda”)** and **attach(”College.rda”)** command. Make sure you have the directory set to the correct location for the data.

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

(c) Use the **pairs()** function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

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

(e) 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.

(f) Use the **hist()** function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command **par(mfrow=c(2,2))** useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

(g) Use the **lm** function to find a regression equation predicting the number of new students based on the graduation rate, qualifications of the faculty, and various expenses.

Only for Stat-627:

5. (#3-6, page 121). Argue that in the case of simple linear regression, the least squares line always passes through the point of averages $(\bar{X}, \bar{Y})$.