Homework #2

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. Describe the null hypotheses to which the given p-values

	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

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.

For the model:$(sold) = \beta_0 + \beta_1(TV) + \beta_2(radio) + \beta_3(newspaper)$
the corresponding null hyphotheses are:

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_0 : \beta_1 = 0 \\ H_0 : \beta_2 = 0 \\ H_0 : \beta_3 = 0 \end{cases}$$

From the table, radio and TV advertising have p-value very low, so they significantly affect number of units sold. But the p-value of newspaper advertising is very high, so it has no significant effect on number of units sold.

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.

The third one is correct. Assume that GPA $= \infty$, IQ $= \epsilon$, where ϵ is a constant greater than 0.

When Gender = 1, in other words, a person is a female:

$$\begin{aligned}(\text{starting salary})_0 &= 50 + 20 \cdot \infty + 0.07 \cdot \epsilon + 35 \cdot 1 + 0.01 \cdot \infty \cdot \epsilon + (-10) \cdot \infty \cdot 1 \\ &= 85 + (20 + 0.01\epsilon - 10)\infty + 0.07\epsilon \\ &= 85 + (10 + 0.01\epsilon)\infty + 0.07\epsilon\end{aligned}$$

When Gender = 1, in other words, a person is a male:

$$\begin{aligned}(\text{starting salary})_1 &= 50 + 20 \cdot \infty + 0.07 \cdot \epsilon + 35 \cdot 0 + 0.01 \cdot \infty \cdot \epsilon + (-10) \cdot \infty \cdot 0 \\ &= 50 + (20 + 0.01\epsilon)\infty + 0.07\epsilon\end{aligned}$$

Because $(\text{starting salary})_1 > (\text{starting salary})_0$, so iii. is correct.

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

$$\begin{aligned}(\text{starting salary}) &= 50 + 20 \cdot 4 + 0.07 \cdot 110 + 35 \cdot 1 + 0.01 \cdot 4 \cdot 110 + (-10) \cdot 4 \cdot 1 \\ &= 50 + 80 + 7.7 + 35 + 4.4 - 40 \\ &= 137.1\end{aligned}$$

The predicted starting salary is 137,100 dollars.

(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.

False. We can not tell the significance of a term based on its coefficient.

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. $Y = Y = \beta_0 + \beta_1 X + \varepsilon$. 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.

We would expect the cubic regression has lower training RSS. The cubic regression model is more flexible than the linear regression model. So it makes sense to see lower training RSS in the cubic regression.

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

Since the true model is linear, so the cubic regression will have a larger 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.

We expect the cubic regression to have lower RSS. Adding predictors, quadratic term and cubic term in this case, into the model will not reduce its explaination to the total variation of the data. The linear model is the special case of the cubic model, in other words, the cubic model could only either explain the same as or more than the linear model. Thus the cubic model will be very likely to have lower RSS.

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

There is not enough information to tell. It depends on what true model is and the performance of these two models.

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.

```
library(ISLR)
attach(College)
```

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

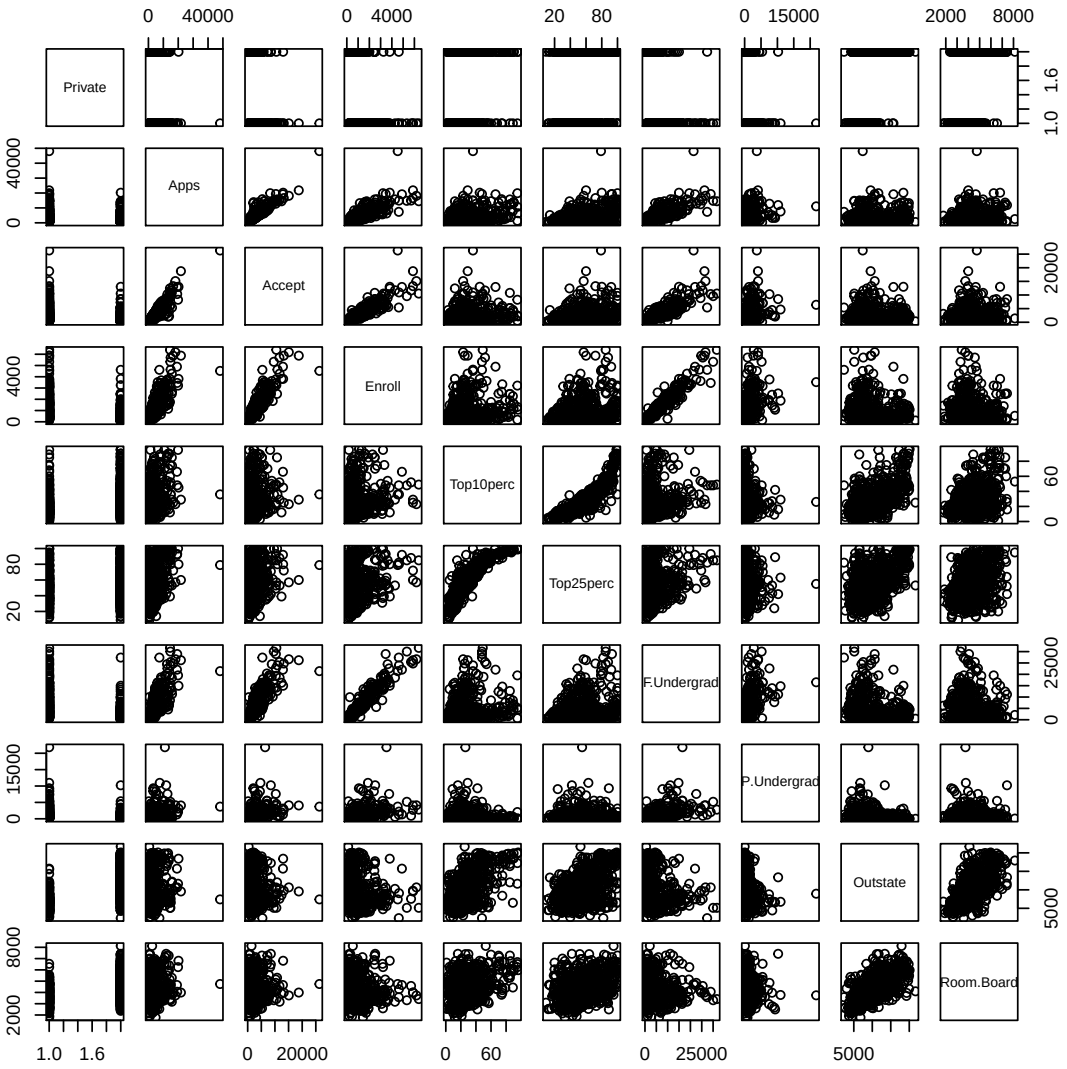
```
summary(College)
```

```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
```

```
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

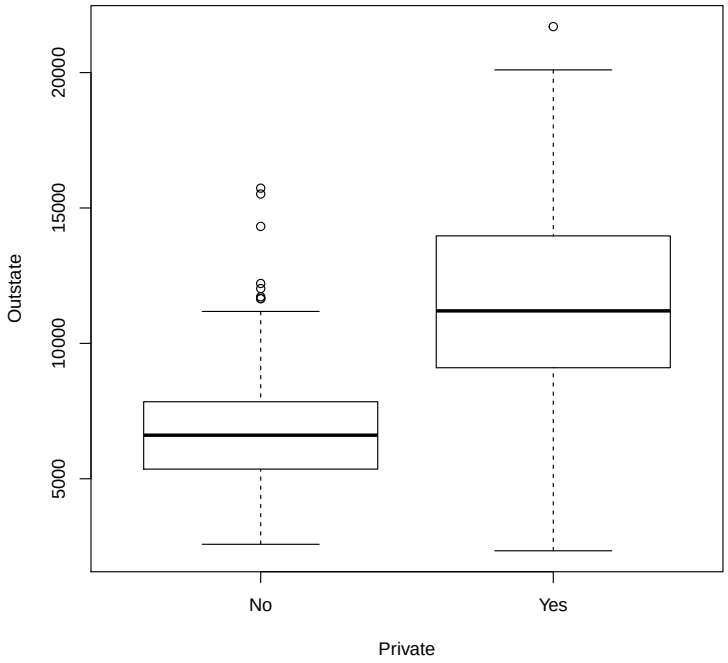
(c) Use the **pairs()** function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

```
pairs(College[,1:10])
```



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

```
plot(Outstate~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.

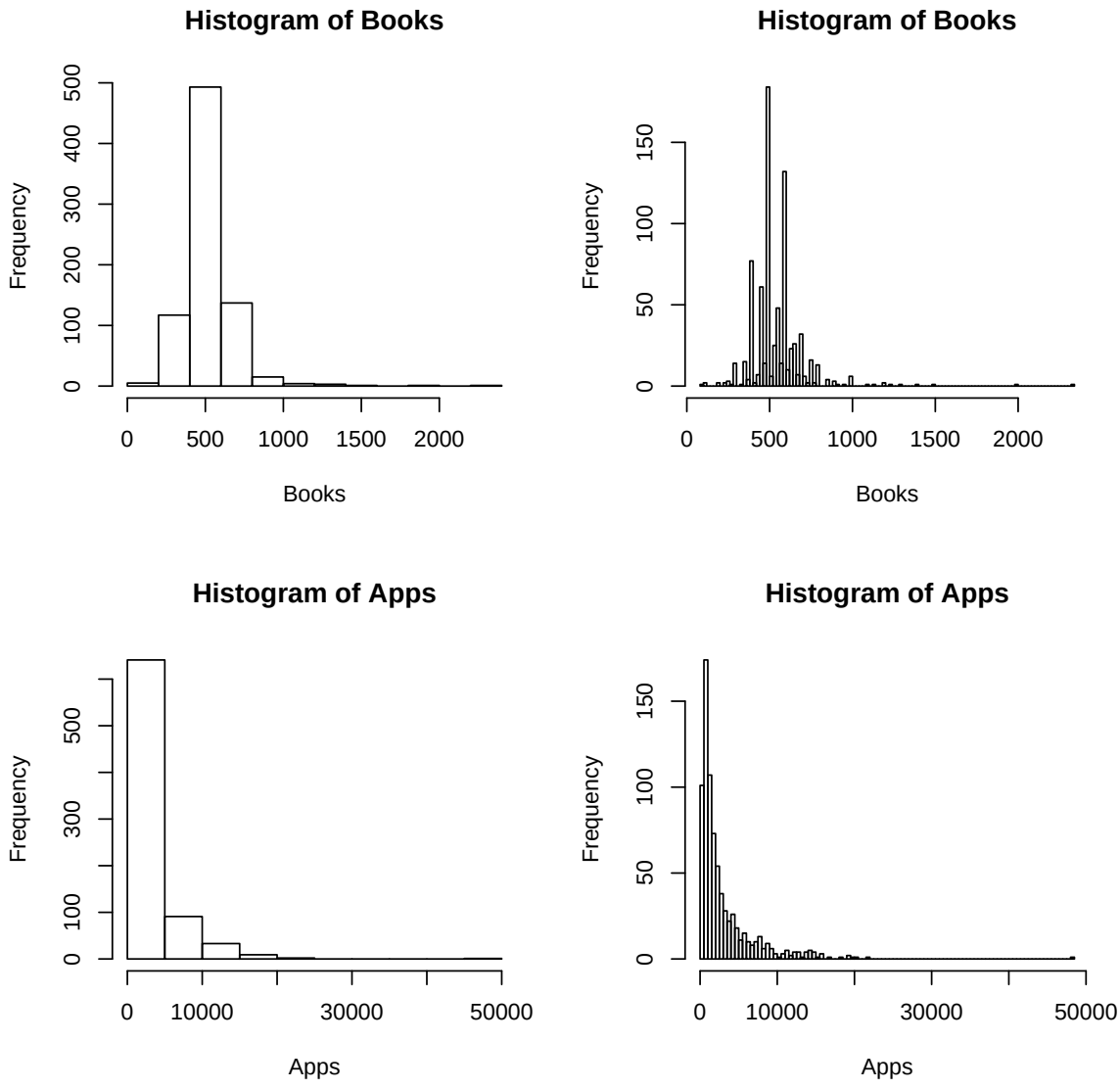
```
Elite = rep ("No",nrow(College))
Elite [College$ Top10perc > 50] = " Yes"
Elite = as.factor (Elite)
College = data.frame(College,Elite)
summary(College$Elite)
```

```
##  Yes   No
##   78  699
```

(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.

```
par(mfrow=c(2,2))
hist(Books, breaks = 10)
hist(Books, breaks = 100)
hist(Apps, breaks = 10)
hist(Apps, breaks = 100)
```



(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.

```
lm(Enroll ~ Grad.Rate + Terminal + PhD +
     Expend + Personal + Books + Room.Board + Outstate)
```

```
##
## Call:
## lm(formula = Enroll ~ Grad.Rate + Terminal + PhD + Expend + Personal +
##     Books + Room.Board + Outstate)
##
## Coefficients:
## (Intercept)    Grad.Rate     Terminal          PhD       Expend
##  -1.236e+03    4.282e+00    1.084e+01    1.476e+01    2.254e-02
##    Personal        Books   Room.Board     Outstate
##   2.657e-01    3.038e-01    8.804e-03   -9.386e-02
```

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 $(\bar{X}, \bar{Y})$.

SLR: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, we know that $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, when $X_i = \bar{X}$,

$$\hat{Y} = (\bar{Y} - \hat{\beta}_1 \bar{X}) + \hat{\beta}_1 \bar{X} = \bar{Y}$$

So, the SLR line passes throught $(\bar{X}, \bar{Y})$.