Group assignment 1

1. Describe the null hypotheses to which the p-values given in Table 3.4 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.

The null hypothesis for the table is that the advertising budgets don't have an affect on sales. Each category has its own p value to be analyzed. Using the common significance level of .05 for the given p values in the table we can come to a few conclusions for the data. For tv the p value is .0001, since this value is less than the significance level, we reject the null hypothesis for this instance. Second is radio, which has the p value of .0001. This value is also less than the significance level, so we reject the hypothesis. Lastly we have newspapers, which have a p value of .8599. This value is larger than the significance level. So for this reason we do not reject the null hypothesis.

3. Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Level (1 for College and 0 for High School), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get β^0 = 50, β^1 = 20, β^2 = 0.07, β^3 = 35, β^4 = 0.01, β^5 = −10.

(a) Which answer is correct, and why?

$\hat{Y} = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + B_4 X_4 + B_5 X_5$

$\hat{B}_0 = 50$, $\hat{B}_1 = 20$, $\hat{B}_2 = 0.07$, $\hat{B}_3 = 35$, $\hat{B}_4 = .01$, $\hat{B}_5 = -10$

Level for college = 1, Level for highschool = 0

$X_1 = GPA$, $X_2 = IQ$, $X_3 = level$, $X_4 = (GPA)(IQ)$, $X_5 = (GPA)(level)$

→ College student

→ $50 + (20)(GPA) + (.07)(IQ) + (35)(1) + (.01)(GPA)(IQ) + (-10)(GPA)(1)$

→ $50 + 20(GPA) + (.07)(IQ) + 35 + (.01)(GPA)(IQ) - 10(GPA)$

→ $50 + 20x_1 + (.07)x_2 + 35 + (.01)x_4 - 10x_1$

→ $85 + 10x_1 + (.07)x_2 + (.01)x_4$

→ Highschool

→ $50 + (20)(GPA) + (.07)(IQ) + (35)(0) + (.01)(GPA)(IQ) + (-10)(GPA)(0)$

→ $50 + (20)(GPA) + (.07)(IQ) + (.01)(GPA)(IQ)$

→ $50 + 20x_1 + (.07)x_2 + (.01)x_4$

→ Highschool - College

→ $[50 + 20x_1 + .07x_2 + .01x_4] - [85 + 10x_1 + .07x_2 + .01x_4]$

→ $-35 + 10x_1 + 0 + 0$ → $-35 + 10x_1$ → $10x_1 - 35$

→ $10x_1 - 35 > 0$ → $10x_1 > 35$ → $x_1 > \frac{35}{10}$

→ $x_1 > 3.5$

→ When a highschool graduate earns a GPA higher than 3.5, they make more money

option iii is correct

(b) Predict the salary of a college graduate with an IQ of 110 and a GPA of 4.0.

IQ = 110 , GPA = 4.0

→ using previous equation

→ $50 + (20)GPA + .07(IQ) + 35(1) + (.01)(GPA)(IQ) - (10)(GPA)$

→ $50 + (20)(4.0) + .07(110) + 35 + (.01)(4.0)(110) - (10)(4.0)$

→ $50 + 80 + 7.7 + 35 + 4.4 - 40$

→ 137.1 → 137.1 x1000 → $137,100

(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.

This statement is false. The reason that it is false is because of the way the question is framed. Just because the interaction term is small, does not mean anything. To truly find out we would have to conduct a hypothesis test on β^4. It carries the interaction of GPA and IQ so we would have to obtain the p value, and from there analyze the result to see whether it is of interest or true.

8. This question involves the use of simple linear regression on the Auto data set.

(a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output. For example:

i. Is there a relationship between the predictor and the response?

Yes there is a relationship between the two and they can be analyzed in the summary data
The P value is less than 2e-16. Since the P value is less than 0.5

ii)How strong is the relationship between the predictor and the response?

Based on the R-squared values of .6059 and .6049. We can conclude that the relationship is strong

iii. Is the relationship between the predictor and the response positive or negative?

The regression coefficient is negative, hence the relationship is negative.

iv) What is the predicted mpg associated with a horsepower of 98? What are the associated 95 % confidence and prediction intervals?

The prediction for horsepower 98 is

```
> predict(lm.fit, data.frame(horsepower = 98), interval ="confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
```

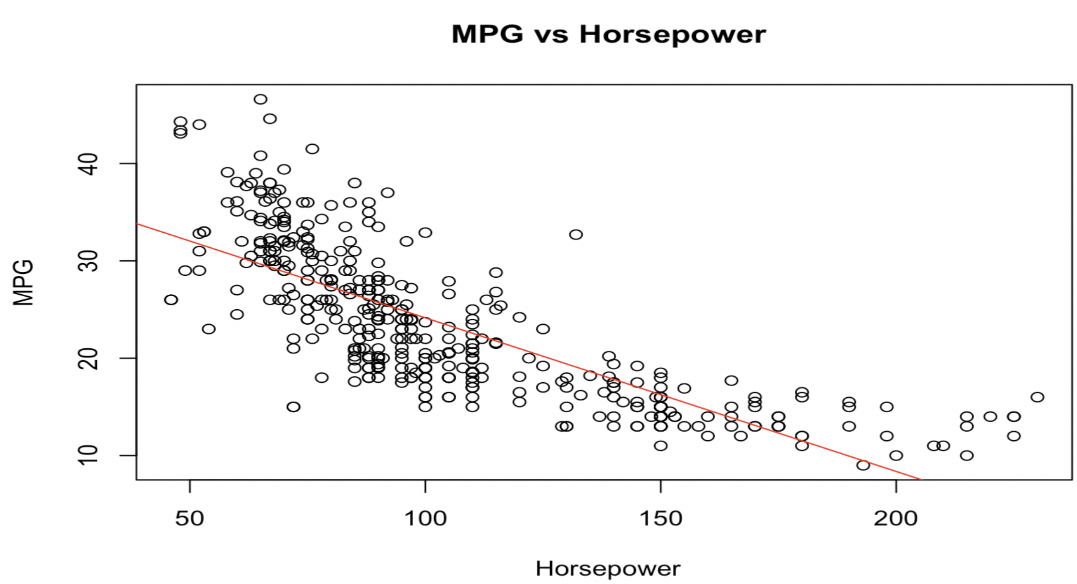
The confidence interval 95% is

```
> predict(lm.fit, data.frame(horsepower = c(95)), interval ="confidence")
       fit     lwr      upr
1 24.94061 24.4389 25.44232
```

The prediction interval 95% is

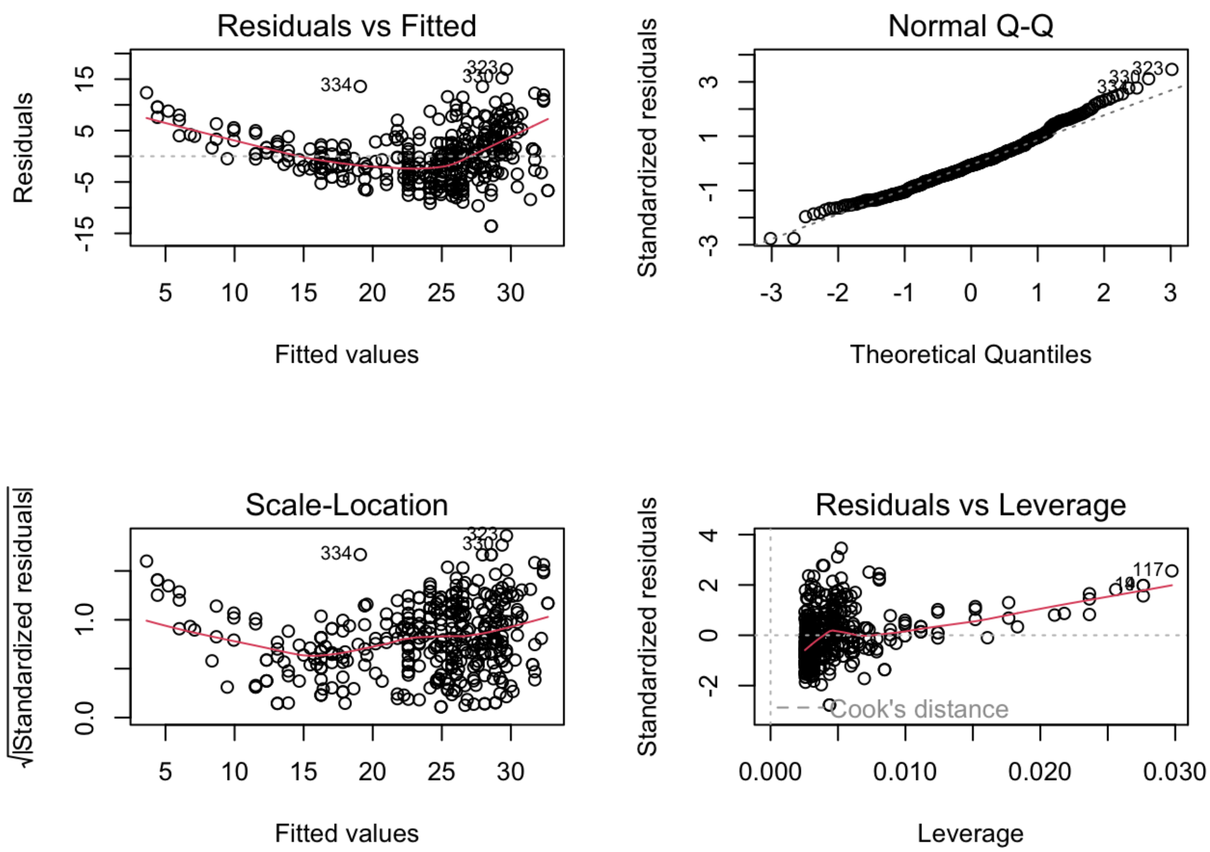
```
> predict(lm.fit, data.frame(horsepower = c(95)), interval ="prediction")
       fit      lwr      upr
1 24.94061 15.28253 34.59869
```

B) **Plot the response and the predictor. Use the abline() function to display the least squares regression line.**



(c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

One problem is that the plot shows an element of non linearity present within the variables.



Source code used for exercise 8:

```
library("ISLR")
lm.fit <- lm(mpg ~ horsepower, data = Auto)
summary(lm.fit)
predict(lm.fit, data.frame(horsepower = 98), interval ="confidence")
predict(lm.fit, data.frame(horsepower = c(95)), interval ="confidence")
predict(lm.fit, data.frame(horsepower = c(95)), interval ="prediction")
attach(Auto)
plot(mpg~horsepower, main =" MPG vs Horsepower", xlab = " Horsepower", ylab ="MPG")
abline(coef = coef(lm.fit), col ="red")
detach(Auto)
par(mfrow=c(2,2))
plot(lm.fit)
```