Rakesh Veetekat

Conceptual and theoretical questions

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 hypotheses that the p-values in Table 3.4 correspond to are that “TV”, “radio”, and “newspaper” should not have a significant relationship with sales. However, the p-values for “TV” and “radio” are very close to 0, while not at all close to 0 for “newspaper”, which means that the null hypotheses for “TV” and “radio” can be rejected, and the null hypothesis for “newspaper” should not been rejected.

2. Carefully explain the differences between the KNN classifier and KNN regression methods.

KNN classifier methods take a single observation as input, and output the average of the K nearest inputs, which is quite different from KNN regression methods. In KNN regression methods, the input is not an actual observation, but rather an prediction for an observation is used and the way the output is obtained does not change, but the use for the method is that it becomes a function that can predict values for any input.

3. Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Gender (1 for Female and 0 for Male), X4 = Interaction between GPA and IQ, and X5 = 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.

With the least squares values used to fit the model, the equation for the least squares line is:
$\hat{y}\text{-pred} = \hat{\beta}0 + \hat{\beta}1*X1 + \hat{\beta}2*X2 + \hat{\beta}3*X3 + \hat{\beta}4*X4 + \hat{\beta}5*X5$
y-pred = 50 + 20*GPA + 0.07*IQ + 35*Gender + 0.01*GPA*IQ + (-10)*GPA*Gender

The equation for males is:
y-pred_male = 50 + 20*GPA + 0.07*IQ + 0.01*GPA*IQ
The equation for females is:
y-pred_female = 85 + 10*GPA + 0.07*IQ + 0.01*GPA*IQ

By taking IQ and GPA out of the two equations since they are fixed for both, the two equations are reduced to y-pred_male = 50 + 20*GPA and y-pred_female = 85 + 10*GPA. After comparing the two, the inequality 50 + 20*GPA > 85 + 10*GPA can be solved to find that the model will predict that males with a GPA > 3.5 will on average earn more than females.

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

y-pred_female = 85 + 10*GPA + 0.07*IQ + 0.01*GPA*IQ

With an IQ of 110 and a GPA of 4.0, this individual will be predicted to earn 85 + 10*4 + 0.07*110 + 0.01*4*110 = 137.1, and this would be a salary of 137.1*1000 = $137100.

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. The coefficient does not give us evidence of an interaction effect. Instead, we have to test the null hypothesis for the GPA/IQ interaction term, and see if the p-value suggests that there is little evidence of an interaction event.

4. 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. $Y = \beta0 + \beta1X + \beta2X2 + \beta3X3 + \epsilon$.
    a. Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta0 + \beta1X + \epsilon$. 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 that the training RSS would be lower for the linear regression as opposed to the training RSS for the cubic regression because the relationship between X and Y appears to be linear.

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

Since we are now analyzing test RSS, we do not have enough information to accurately predict whether the linear or cubic test RSS is lower or higher. However, the cubic may have a higher test RSS because of overfitting.

c. Suppose that the true relationship between X and Y is not linear, but we don't 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.

The cubic regression training RSS would be lower than the linear regression training RSS because it will be flexible to the dataset.

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

In terms of the test RSS, we do not have enough information to determine whether the test RSS of the linear or cubic regression will be lower.

5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form

$$\hat{y}_i = x_i\hat{\beta},$$

where

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i y_i\right) / \left(\sum_{i'=1}^{n} x_{i'}^2\right). \tag{3.38}$$

Show that we can write

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}.$$

What is $a_{i'}$?

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations *of the response values.*

We're given:

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i y_i\right) / \left(\sum_{i'=1}^{n} x_{i'}^2\right)$$

a_i prime is the counter weight to y_i prime where it will offset the values of y_i prime when calculating y-pred.

6. Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point (x-bar, y-bar).

y = βˆ0 + βˆ1*x
If we substitute x-bar for x: y = βˆ0 + βˆ1*x-bar
Rearranging the first equation gives us: βˆ0 = y - βˆ1*x
So, we can write: y = (y - βˆ1*x) + βˆ1*x-bar
Which means that: y = y-bar - βˆ1*x-bar + βˆ1*x-bar = y-bar
This means that (x-bar, y-bar) should be on the least squares line.

Applied

7. 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:

```
Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

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

There is a significant relationship between the predictor (horsepower) and the response (mpg), as noted by the very small p-value for the F-statistic.

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

The R-squared value indicates how strong the relationship between the predictor and the response is, and it's moderately high with a value of 0.6059.

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

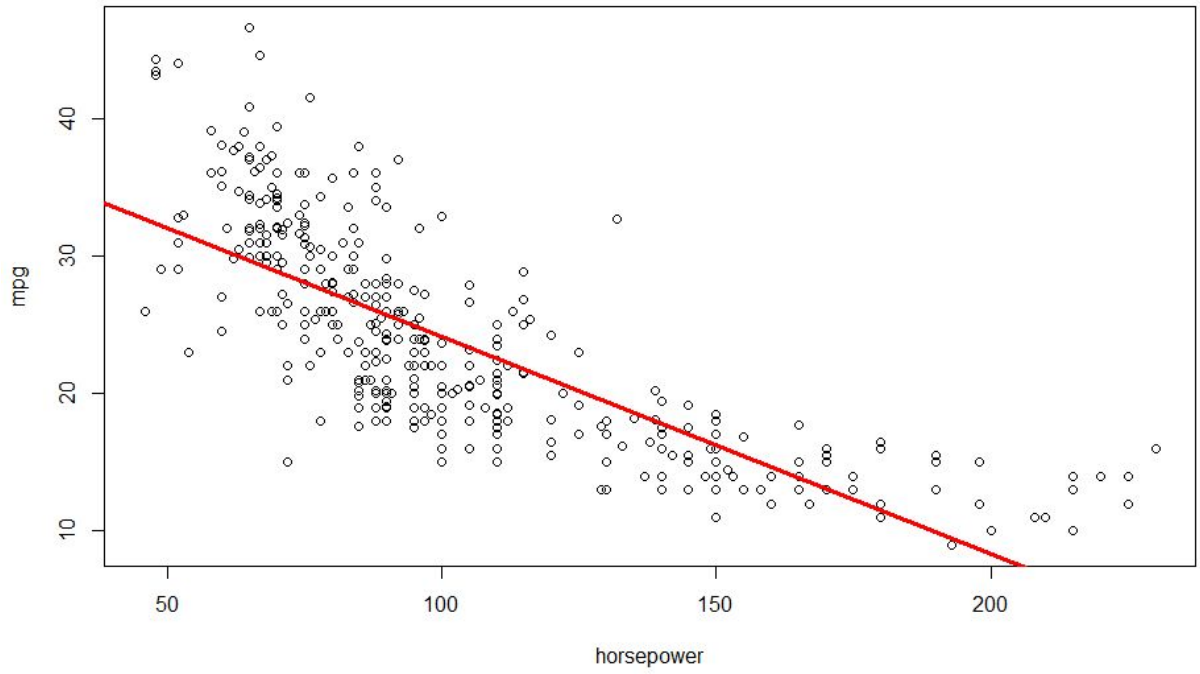
The relationship is negative because the coefficient for horsepower is negative in the summary.

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

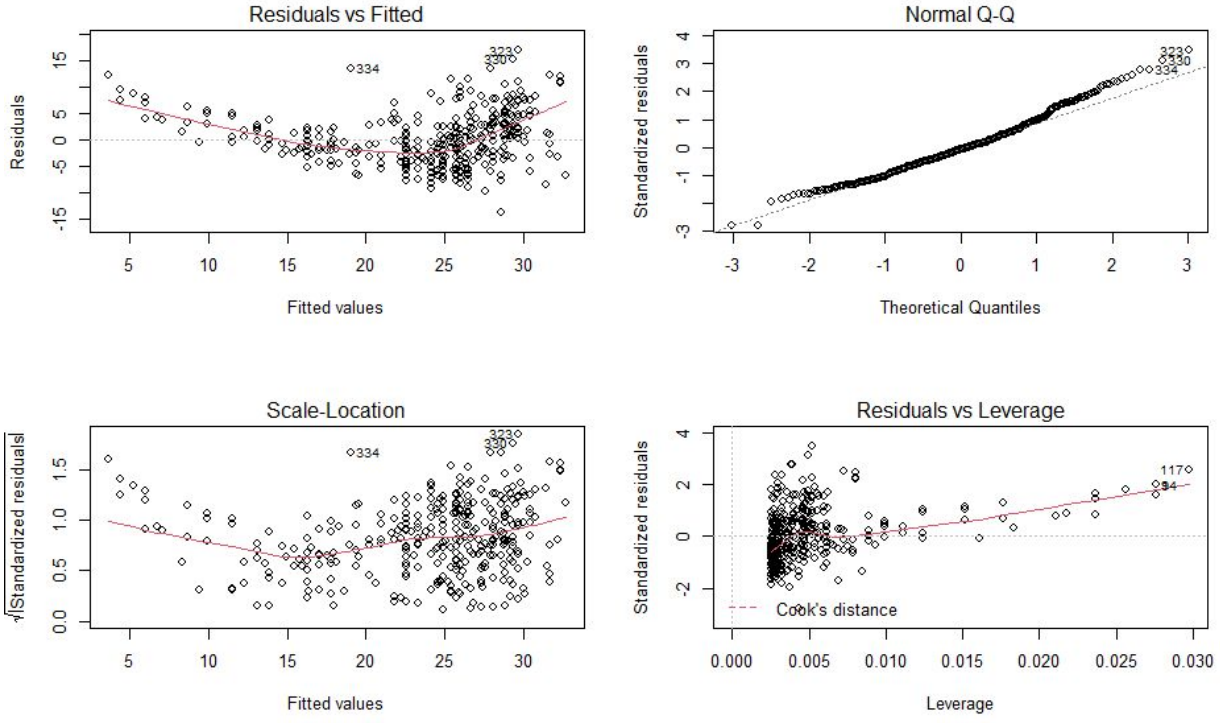
```
> predict(fit, data.frame(horsepower = 98), interval = "confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
> predict(fit, data.frame(horsepower = 98), interval = "prediction")
       fit     lwr      upr
1 24.46708 14.8094 34.12476
>
```

The predicted mpg associated with a horsepower of 98 is 24.46708. The associated 95% confidence and predictions intervals are shown in the image above.

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

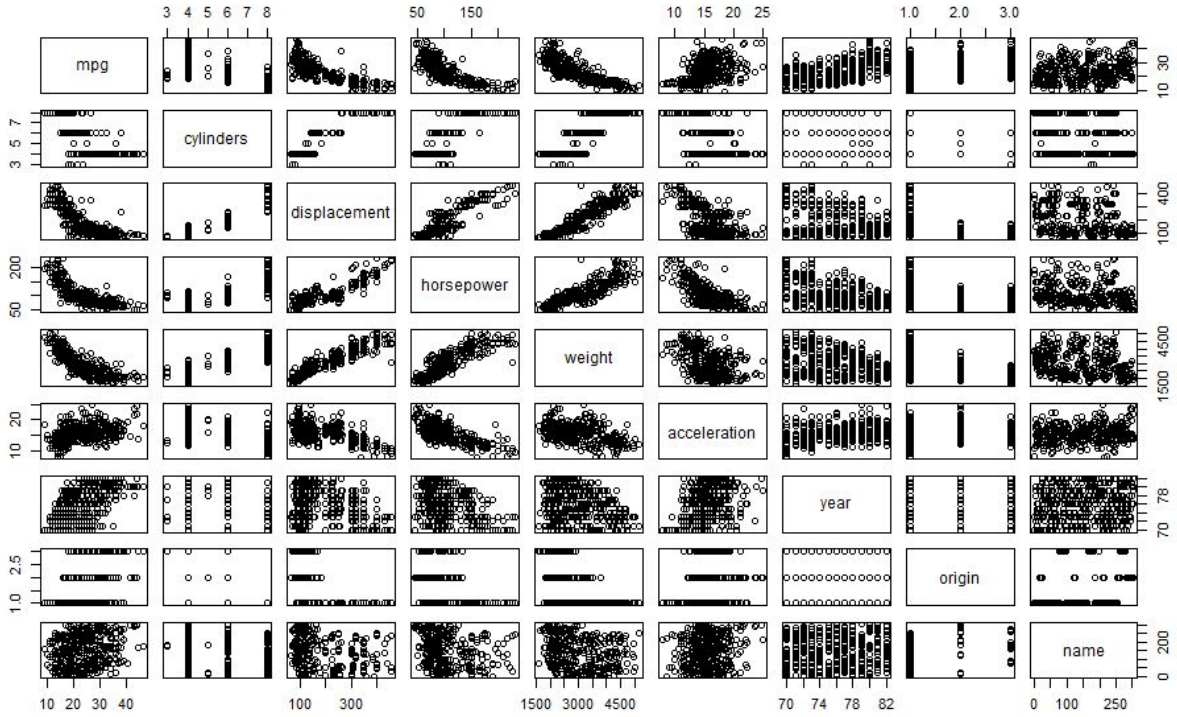


c. Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.



The three plots besides the “Normal Q-Q” display an abundance of non-linearity in the data. There also seems to be a strong bias for higher and lower values with less bias in the middle.

8. This question involves the use of multiple linear regression on the Auto data set.
    a. Produce a scatterplot matrix which includes all of the variables in the data set.



    b. Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, cor() which is qualitative.

```
                mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277  1.0000000
> |
```

    c. Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output. For instance:

```
Call:
lm(formula = mpg ~ . - name, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.218435   4.644294  -3.707  0.00024 ***
cylinders    -0.493376   0.323282  -1.526  0.12780
displacement  0.019896   0.007515   2.647  0.00844 **
horsepower   -0.016951   0.013787  -1.230  0.21963
weight       -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration  0.080576   0.098845   0.815  0.41548
year          0.750773   0.050973  14.729  < 2e-16 ***
origin        1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16

> |
```

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

Again, there is a significant relationship between the predictors and the response (mpg) because of the very small p-value for the F-statistic.

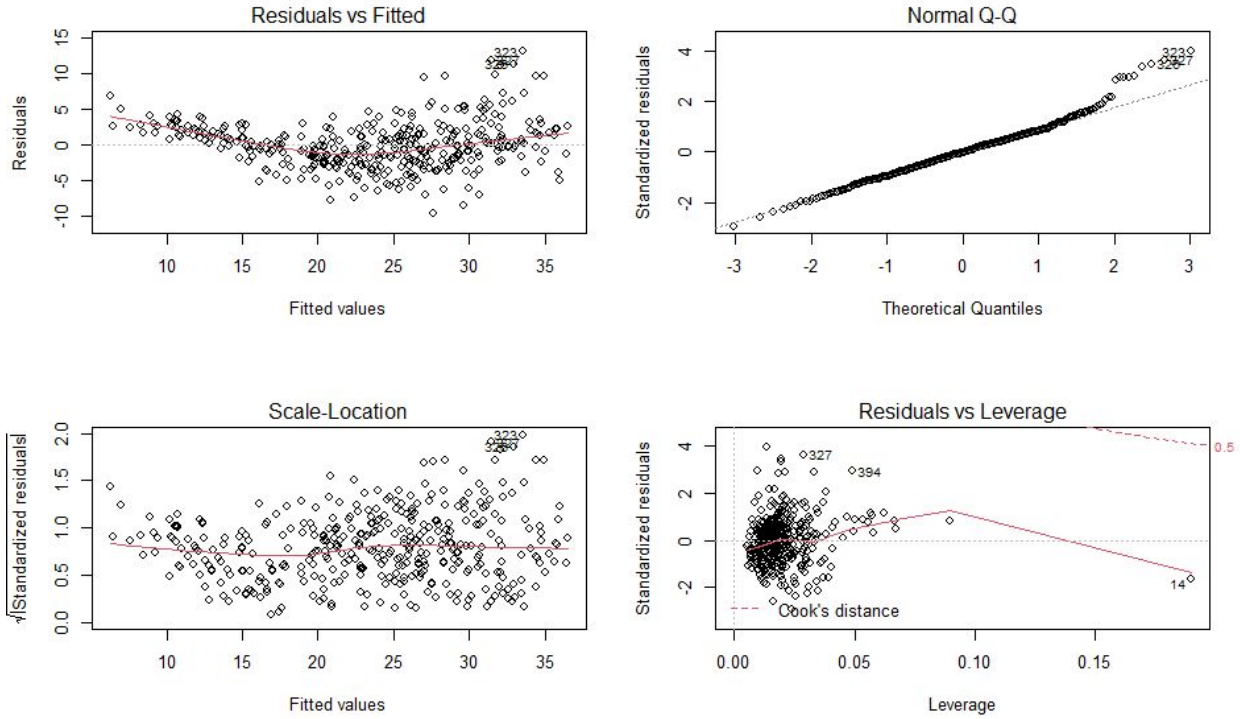
ii. Which predictors appear to have a statistically significant relationship to the response?

The predictors that appear to have the most statistically significant relationship to the response are all of them except for “cylinders”, “horsepower”, and “acceleration” because these 3 have p-values greater than 0.1.

iii. What does the coefficient for the year variable suggest?

The coefficient of the “year” variable suggests that there is a positive correlation that exists between the “year” and “mpg”. This can be interpreted as the mpg increasing as the manufacturing date of the automobiles increase.

d. Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?



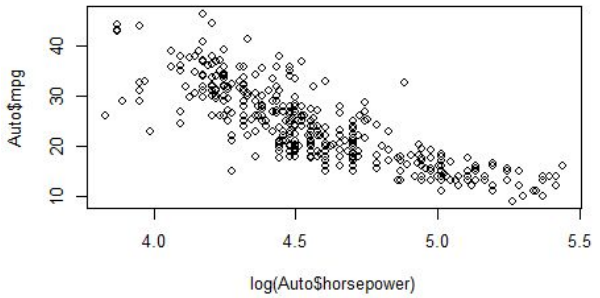
The residual plots do suggest that there are a few outliers. However, the leverage plot suggests that there aren't many observations with unusually high leverage.

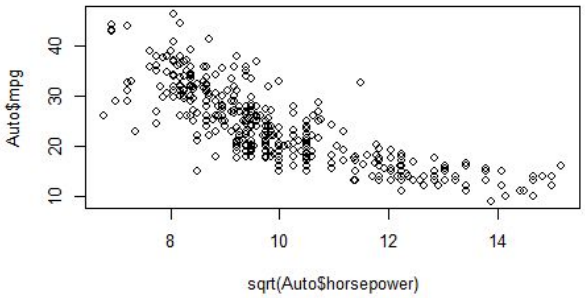
e. Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

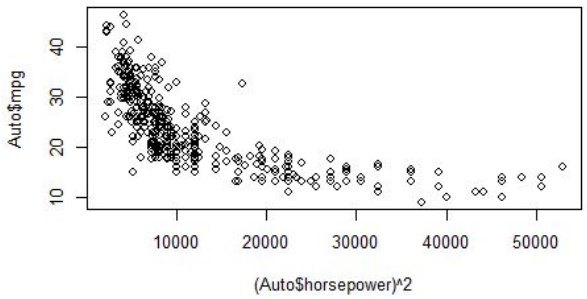
```
Call:
lm(formula = mpg ~ cylinders * displacement + displacement *
    weight, data = Auto[, 1:8])

Residuals:
     Min       1Q   Median       3Q      Max
-13.2934  -2.5184  -0.3476   1.8399  17.7723

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.262e+01  2.237e+00  23.519  < 2e-16 ***
cylinders            7.606e-01  7.669e-01   0.992    0.322
displacement        -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
weight              -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384
displacement:weight  2.128e-05  5.002e-06   4.254 2.64e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.103 on 386 degrees of freedom
Multiple R-squared:  0.7272,    Adjusted R-squared:  0.7237
F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16

> |
```

The interactions between “displacement” and “weight” still seem to be statistically significant because of the very low p-value. However, the interactions between “cylinders” and “displacement” do not appear to be as significant because of its higher p-value.

f. Try a few different transformations of the variables, such as log(X), √ X, X2. Comment on your findings.







After performing the "log(X), √ X, and X-squared" transformations on just the horsepower predictor, it appears that the log transformation makes it the most linear, with the square-root transformation second, and the square transformation coming in last in terms of making the plot the most linear.

9. This question should be answered using the Carseats data set.
    a. Fit a multiple regression model to predict Sales using Price, Urban, and US.

```
Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9206 -1.6220 -0.0564  1.5786  7.0581

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,    Adjusted R-squared:  0.2335
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16

> |
```

b. Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

The model shows that the “Price” variable is negatively correlated with “Sales”, and it can be interpreted as a 1 unit increase in “Price” will, on average, decrease the “Sales” by 54.459 units. In similar fashion, the “UrbanYes” variable is negatively correlated with “Sales”, and a 1 unit increase in “UrbanYes” will, on average, decrease “Sales” by 21.916 units. Lastly, the “USYes” variable is positively correlated with “Sales”, and a 1 unit increase in “USYes” will, on average, increase “Sales” by 1200.573 units.

c. Write out the model in equation form, being careful to handle the qualitative variables properly.

“Sales” = 13.043469 + (-0.054459 * “Price”) + (-0.021916 * “UrbanYes”) + (1.200573 * “USYes”)

The “UrbanYes” will either be 1 (if the store is in an urban location) or 0 (if the store is not in an urban location). The “USYes” will either be 1 (if the store is in the US) or 0 (if the store is not in the US).

d. For which of the predictors can you reject the null hypothesis H0 : βj = 0?

The null hypotheses for “Price” and “USYes” can be rejected because of their low p-values.

e. On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

```
Call:
lm(formula = Sales ~ Price + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9269 -1.6286 -0.0574  1.5766  7.0515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
Price       -0.05448    0.00523 -10.416  < 2e-16 ***
USYes        1.19964    0.25846   4.641 4.71e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared:  0.2393,    Adjusted R-squared:  0.2354
F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16

> |
```

f. How well do the models in (a) and (e) fit the data?

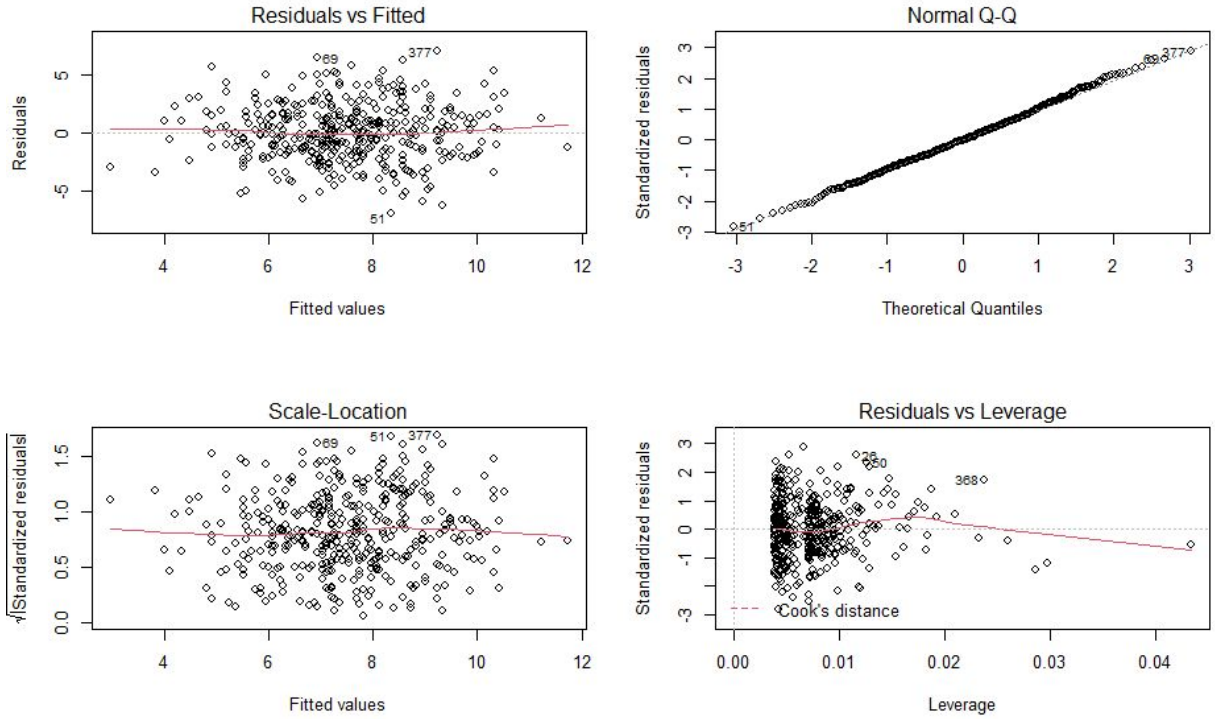
The "Adjusted R-squared" for the model from part (e) is a little bit higher than the same statistic for the model in part (a), which means that model in part (e) only fits the data a little bit better than the model in part (a).

g. Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).

```
> library(ISLR)
> lm.fit = lm(Sales ~ Price + US, data=Carseats)
> #summary(lm.fit)
> confint(lm.fit)
                  2.5 %      97.5 %
(Intercept) 11.79032020 14.27126531
Price       -0.06475984 -0.04419543
USYes        0.69151957  1.70776632
> |
```

The 95% confidence intervals for both variables can be seen in the image above.

h. Is there evidence of outliers or high leverage observations in the model from (e)?



The “Residuals vs Leverage” plot shows that there are a few outliers and also a few high leverage points in the model.

10. This problem involves the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.
    a. For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

```
Call:
lm(formula = crim ~ zn, data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-4.429 -4.222 -2.620  1.250 84.523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
zn          -0.07393    0.01609  -4.594 5.51e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.435 on 504 degrees of freedom
Multiple R-squared:  0.04019,   Adjusted R-squared:  0.03828
F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06

> lm.fit.indus=lm(crim~indus,data=Boston)
> summary(lm.fit.indus)

Call:
lm(formula = crim ~ indus, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-11.972  -2.698  -0.736   0.712  81.813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.06374    0.66723  -3.093  0.00209 **
indus        0.50978    0.05102   9.991  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.866 on 504 degrees of freedom
Multiple R-squared:  0.1653,    Adjusted R-squared:  0.1637
F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16

> lm.fit.chas=lm(crim~chas,data=Boston)
> summary(lm.fit.chas)
```

```
Call:
lm(formula = crim ~ chas, data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-3.738 -3.661 -3.435  0.018 85.232

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7444     0.3961   9.453   <2e-16 ***
chas         -1.8928     1.5061  -1.257    0.209
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.597 on 504 degrees of freedom
Multiple R-squared:  0.003124,  Adjusted R-squared:  0.001146
F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094

> lm.fit.nox=lm(crim~nox,data=Boston)
> summary(lm.fit.nox)

Call:
lm(formula = crim ~ nox, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-12.371  -2.738  -0.974   0.559  81.728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -13.720      1.699  -8.073 5.08e-15 ***
nox           31.249      2.999  10.419  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.81 on 504 degrees of freedom
Multiple R-squared:  0.1772,    Adjusted R-squared:  0.1756
F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = crim ~ rm, data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-6.604 -3.952 -2.654  0.989 87.197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.482      3.365   6.088 2.27e-09 ***
rm            -2.684      0.532  -5.045 6.35e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.401 on 504 degrees of freedom
Multiple R-squared:  0.04807,   Adjusted R-squared:  0.04618
F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07

> lm.fit.age=lm(crim~age,data=Boston)
> summary(lm.fit.age)

Call:
lm(formula = crim ~ age, data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-6.789 -4.257 -1.230  1.527 82.849

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.77791    0.94398  -4.002 7.22e-05 ***
age          0.10779    0.01274   8.463 2.85e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.057 on 504 degrees of freedom
Multiple R-squared:  0.1244,    Adjusted R-squared:  0.1227
F-statistic: 71.62 on 1 and 504 DF,  p-value: 2.855e-16
```

```
Call:
lm(formula = crim ~ dis, data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-6.708 -4.134 -1.527  1.516 81.674

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4993     0.7304  13.006   <2e-16 ***
dis          -1.5509     0.1683  -9.213   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.965 on 504 degrees of freedom
Multiple R-squared:  0.1441,    Adjusted R-squared:  0.1425
F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16

> lm.fit.rad=lm(crim~rad,data=Boston)
> summary(lm.fit.rad)

Call:
lm(formula = crim ~ rad, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-10.164  -1.381  -0.141   0.660  76.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.28716    0.44348  -5.157 3.61e-07 ***
rad          0.61791    0.03433  17.998  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.718 on 504 degrees of freedom
Multiple R-squared:  0.3913,    Adjusted R-squared:   0.39
F-statistic: 323.9 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = crim ~ tax, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-12.513  -2.738  -0.194   1.065  77.696

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.528369   0.815809  -10.45   <2e-16 ***
tax          0.029742   0.001847   16.10   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.997 on 504 degrees of freedom
Multiple R-squared:  0.3396,    Adjusted R-squared:  0.3383
F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16

> lm.fit.ptratio=lm(crim~ptratio,data=Boston)
> summary(lm.fit.ptratio)

Call:
lm(formula = crim ~ ptratio, data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-7.654 -3.985 -1.912  1.825 83.353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.6469     3.1473  -5.607 3.40e-08 ***
ptratio       1.1520     0.1694   6.801 2.94e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.24 on 504 degrees of freedom
Multiple R-squared:  0.08407,   Adjusted R-squared:  0.08225
F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11
```

```
Call:
lm(formula = crim ~ black, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-13.756  -2.299  -2.095  -1.296  86.822

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.553529   1.425903  11.609   <2e-16 ***
black       -0.036280   0.003873  -9.367   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.946 on 504 degrees of freedom
Multiple R-squared:  0.1483,    Adjusted R-squared:  0.1466
F-statistic: 87.74 on 1 and 504 DF,  p-value: < 2.2e-16

> lm.fit.lstat=lm(crim~lstat,data=Boston)
> summary(lm.fit.lstat)

Call:
lm(formula = crim ~ lstat, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-13.925  -2.822  -0.664   1.079  82.862

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
lstat        0.54880    0.04776  11.491  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.664 on 504 degrees of freedom
Multiple R-squared:  0.2076,    Adjusted R-squared:  0.206
F-statistic:   132 on 1 and 504 DF,  p-value: < 2.2e-16

Call:
lm(formula = crim ~ medv, data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-9.071 -4.022 -2.343  1.298 80.957

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.79654    0.93419   12.63   <2e-16 ***
medv        -0.36316    0.03839   -9.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.934 on 504 degrees of freedom
Multiple R-squared:  0.1508,    Adjusted R-squared:  0.1491
F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

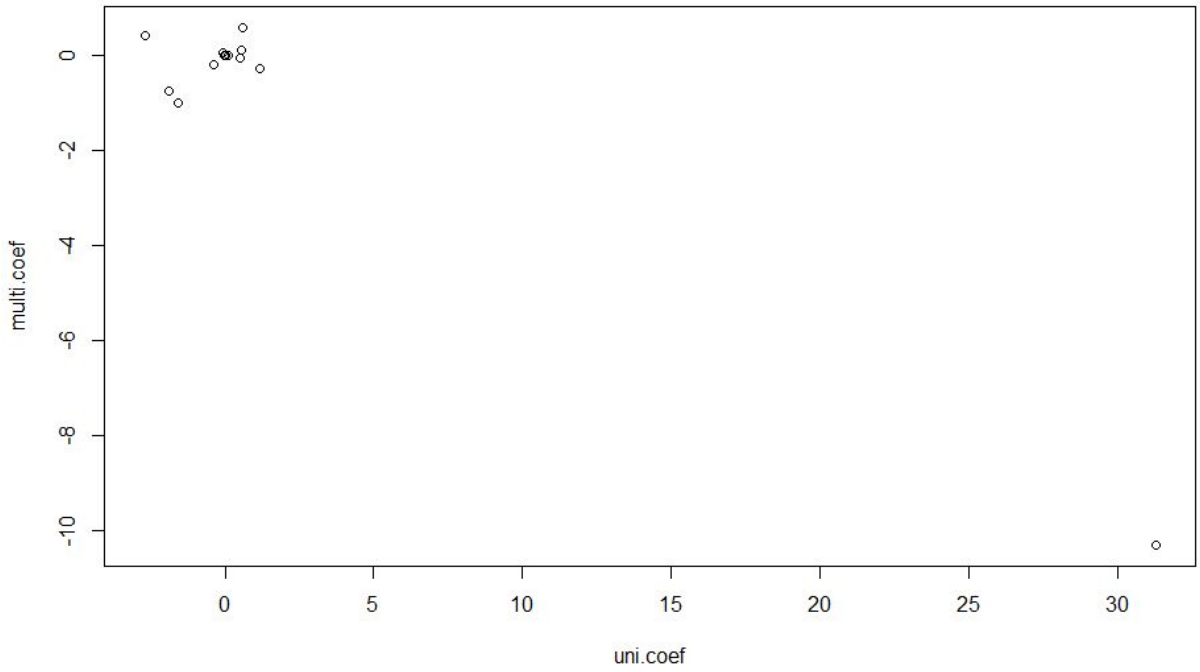
After creating a summary of each predictor's relationship with the response, sifting through the p-values for each uncovers that "chas" is the only predictor that does not have a significant enough relationship with "crim". In all of the other models, the predictors have a significant relationship with "crim" because of the low p-values for each.

b. Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 : βj = 0?

```
Call:
lm(formula = crim ~ ., data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-9.924 -2.120 -0.353  1.019 75.051

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.033228   7.234903   2.354 0.018949 *
zn            0.044855   0.018734   2.394 0.017025 *
indus        -0.063855   0.083407  -0.766 0.444294
chas         -0.749134   1.180147  -0.635 0.525867
nox         -10.313535   5.275536  -1.955 0.051152 .
rm            0.430131   0.612830   0.702 0.483089
age           0.001452   0.017925   0.081 0.935488
dis          -0.987176   0.281817  -3.503 0.000502 ***
rad           0.588209   0.088049   6.680 6.46e-11 ***
tax          -0.003780   0.005156  -0.733 0.463793
ptratio      -0.271081   0.186450  -1.454 0.146611
black        -0.007538   0.003673  -2.052 0.040702 *
lstat         0.126211   0.075725   1.667 0.096208 .
medv         -0.198887   0.060516  -3.287 0.001087 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454,     Adjusted R-squared:  0.4396
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16

> |
```

From the results of the multiple regression model shown above, the null hypotheses for the "zn", "dis", "rad", "black", and "medv" predictors can be rejected since they all have low p-values.

c. How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.



The results in part (a) show that a lot more variables have a significant relationship with “crim” than in part (b). This is because the univariate case makes it so that each predictor is measured against the response excluding the other variables entirely in the univariate case, while these variables exist in the multivariate case, just as fixed values.

d. Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$