Assignment 2
Shaheed Shihan
Modern Applied Statistics II

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

$$\hat{y}_i = x_i \frac{\sum_{j=1}^{n} x_i\, x_j}{\sum_{k=1}^{n} x_k^2}$$

$$\sum_{j=1}^{n} \frac{x_i x_j}{\sum_{k=1}^{n} x_k^2} y_j$$

$$\sum_{j=1}^{n} a_j y_j$$

10. 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!

Besides the store being in an Urban area, both the price and the store location in the US were significant variables that helped predict sales.
More descriptively, the price variable can be interpreted in this manner – an increase of a dollar caused a decrease of around 54 units in sales, with all other predictors being constant. Similar statements can be made about the Urban and the US variable.

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

$$Sales = 13.043469\,x + (-0.054459) * Price + (-0.021916) * Urban + (1.200573) * US + \varepsilon$$

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

We can reject the null hypothesis for the Price and US variable.

(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^2 value is slightly greater for the bigger model than the smaller one. It can be said

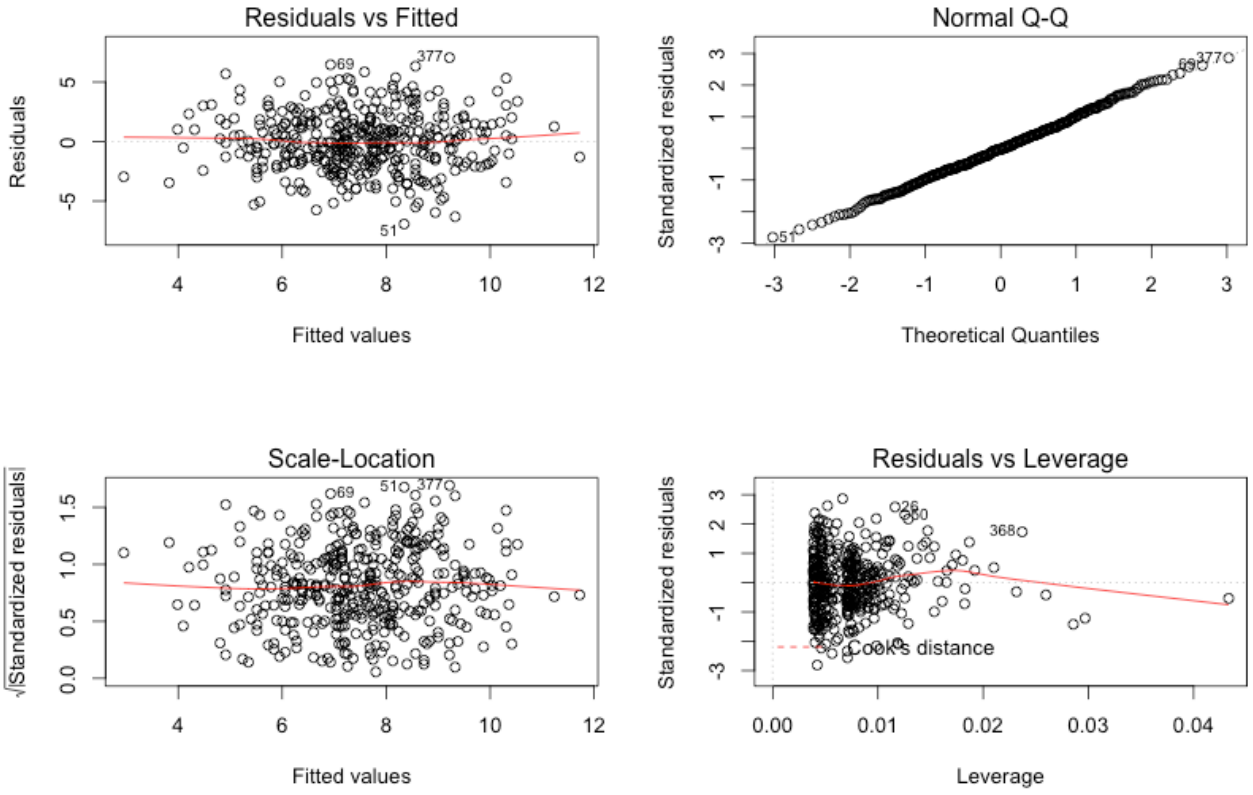
that approximately 23.54% of the variability in the model can be explained by the variance in the variables of Price and US.

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

	2.5 %	97.5 %
(Intercept)	11.79032020	14.27126531
Price	-0.06475984	-0.04419543
USYes	0.69151957	1.70776632

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

The first plot seems to indicate the possibility of a few outliers but none of which are substantial. There seems to be some leverage points as well.



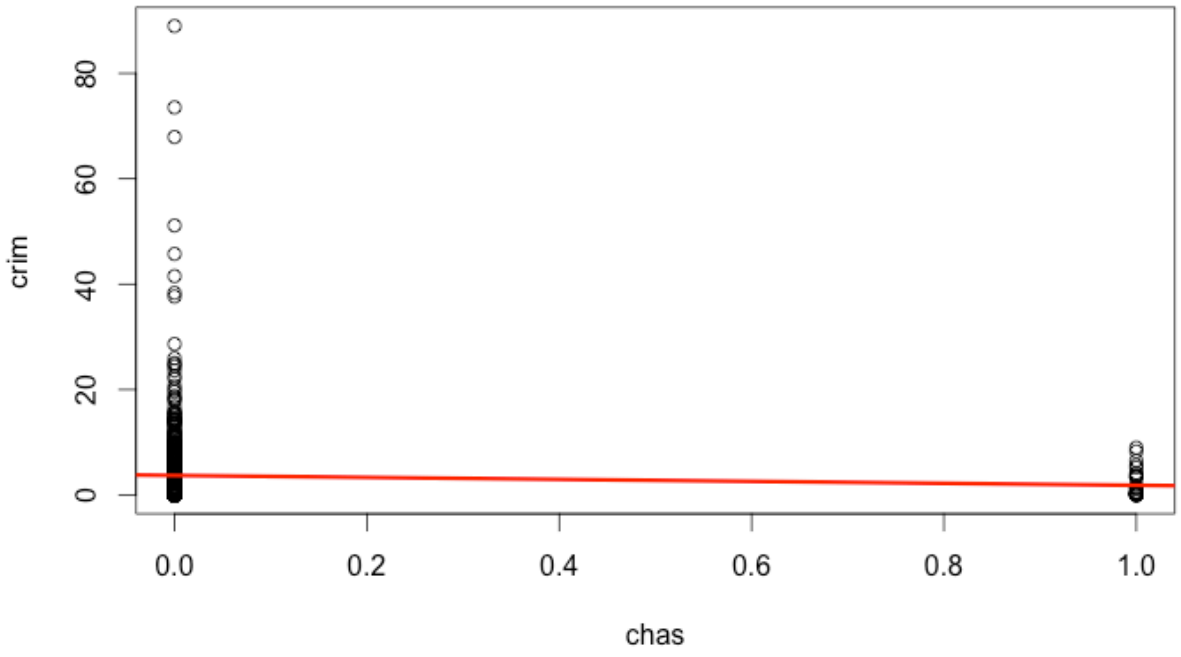
15. This problem involves the Boston data set, which we saw in the lab for this chapter. We

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.

For space and repetition sake, I chose not to attach any of the linear regression here. Please find it in the attached R code.

It was found that all predictors had a p-value less than 0.05 except chas. Therefore only chas was plotted to see if this relationship held.



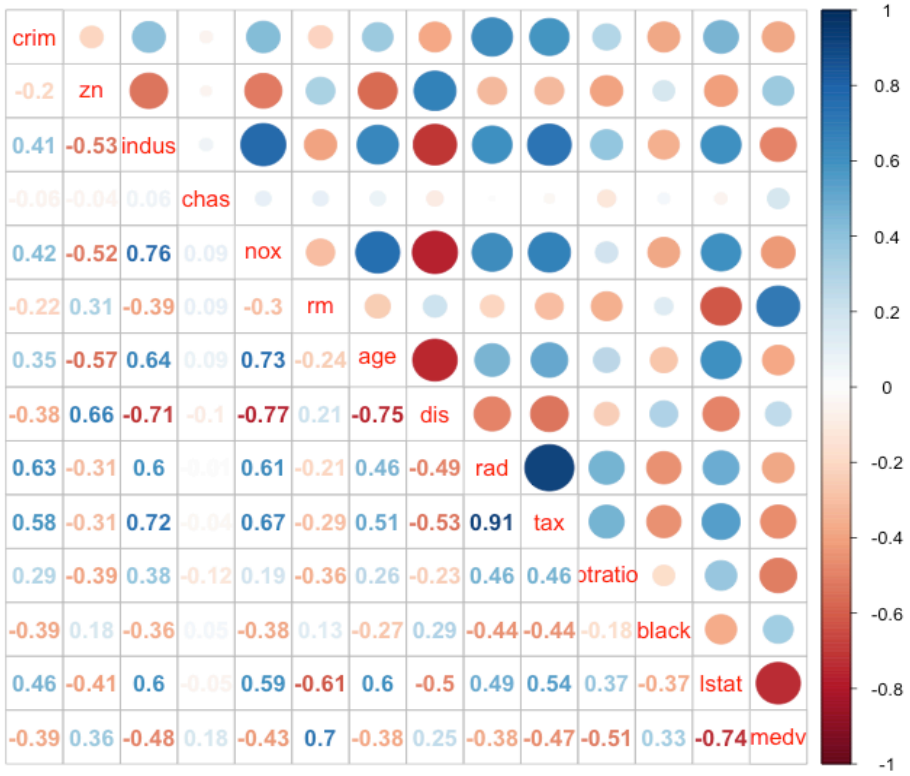
(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H0: βj = 0?

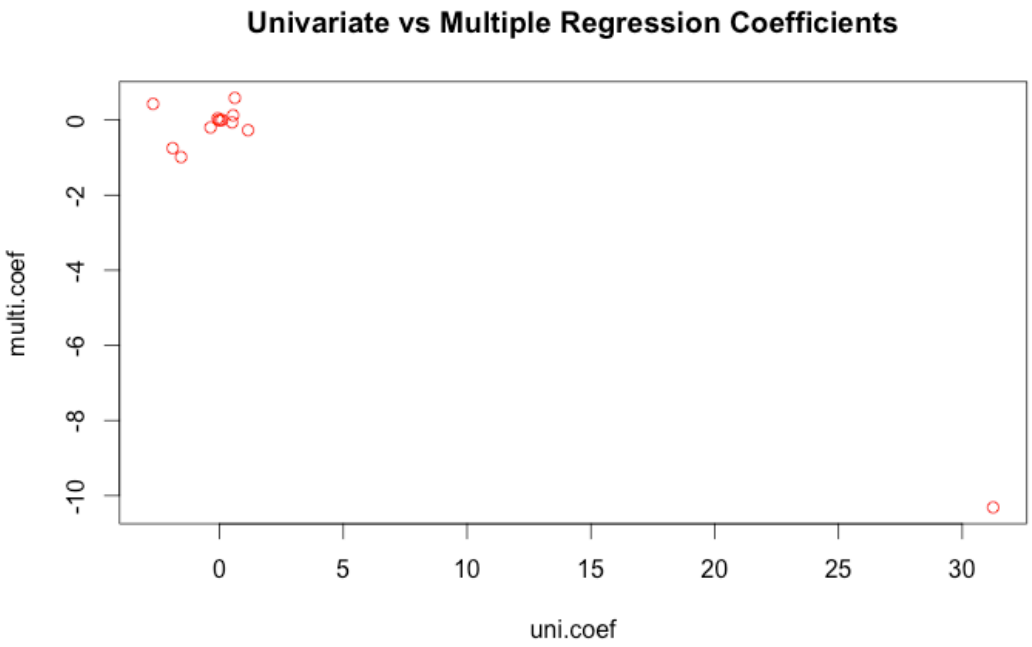
```
Call:
lm(formula = crim ~ ., data = Boston)

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
```

We can reject the null hypothesis for "zn", "dis", "rad" and "medv". The variable "black" can be included in this list as well but further investigation is required.

(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 plot shows that there is a significant difference between the univariate and multiple regression coefficients. In the multiple regression, the slope is the average effect of an increase in the predictor, while other predictors are held constant. This isn't the case in the univariate regression case where the effect of the other predictors is completely ignored.
The multiple regression coefficient proves that there is no relationship between some of the predictor variables and the response variable.





(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

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

Once again, for space and repetition sake, the results from modelling each individual predictor variable wasn't plotted.

It was found that for "zn", "rm", "rad", "tax" and "lstat" as predictor, the p-values for the cubic coefficient weren't significant. For "indus", "nox", "age", "dis", "ptratio" and "medv" as predictor, the p-values for the cubic coefficient were adequate.