Homework #2

Justin Robinette

January 22, 2019

No collaborators for any problem

Problem 3.7.5, page 121: Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form

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

where

$$\hat{\beta} = \sum(x_i y_i)/(\sum(x_i^2)$$

Show that we can write

$$\hat{y}_i = \sum(a_(i')y_(i')$$

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.

Results:

$$a_i = (x_i x_j)/\sum_{i\prime=1}^{n} x_i\,'^2)$$

Problem 3.7.10, pg 123: This problem should be answered using the *Carseats* data set.

(a) Fit a multiple regression model to predict *Sales* using *Price*, *Urban* and *US*.

Results: Below, I've fit a multiple regression model using the variables requested in the question. I've printed the model code for convenience.

```
## [1] "Carseats_fit1 <- lm(Sales ~ Price + Urban + US, data = Carseats)"
```

(b) Provide an interpretation of each coefficient in the model. Be careful - some of the variables in the model are qualitative.

Results: The price variable coefficient shows that the average effect of the **Price** increase is one dollar decreases sales by 0.054459 units. The coefficient from the Urban variable shows that sales in **Urban** locations decrease by 0.021916 units. The coefficient from the US variable shows that sales in the **US** locations increases by 1.200573 units with all other predictors remaining fixed.

A summary is printed to reflect this information.

```
##
## 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
```

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

Results: Using the summary above (and commented out above), I derived the following equation form for the model:

$$Sales = 13.043469 + (-0.054459)Price + (-0.021916)Urban + (1.200573)US + \epsilon$$

(d) For which of the predictors can you reject the null hypothesis

$$H_0: \beta_j = 0$$

Results: We can reject the Null hypothesis for the *Price* and *US* variables because their p-values are below 0.05 indicating statistical significance of the variables.

P-Values of Predictors for Sales

	P-Value
(Intercept)	0.0000000
Price	0.0000000
UrbanYes	0.9357389
USYes	0.0000049

(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.

Results: Below I've fit a model based on the significant variables in the following portion of the exercise. This model is printed for convenience and a summary is shown.

```
## [1] "Carseats_fit2 <- lm(Sales ~ Price + US, data = Carseats)"

##
## 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?

Results: The R^2 for the models is nearly the same. Both models are able to explain approximately 23.93% of the variability in the *Sales* variable. Below is a table showing the R-Squared values from both models in parts **a** and **e**.

Comparison of Model R-Squared Values

Fit1 R-Squared (a)	Fit2 R-Squared (e)
0.23928	0.23926

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

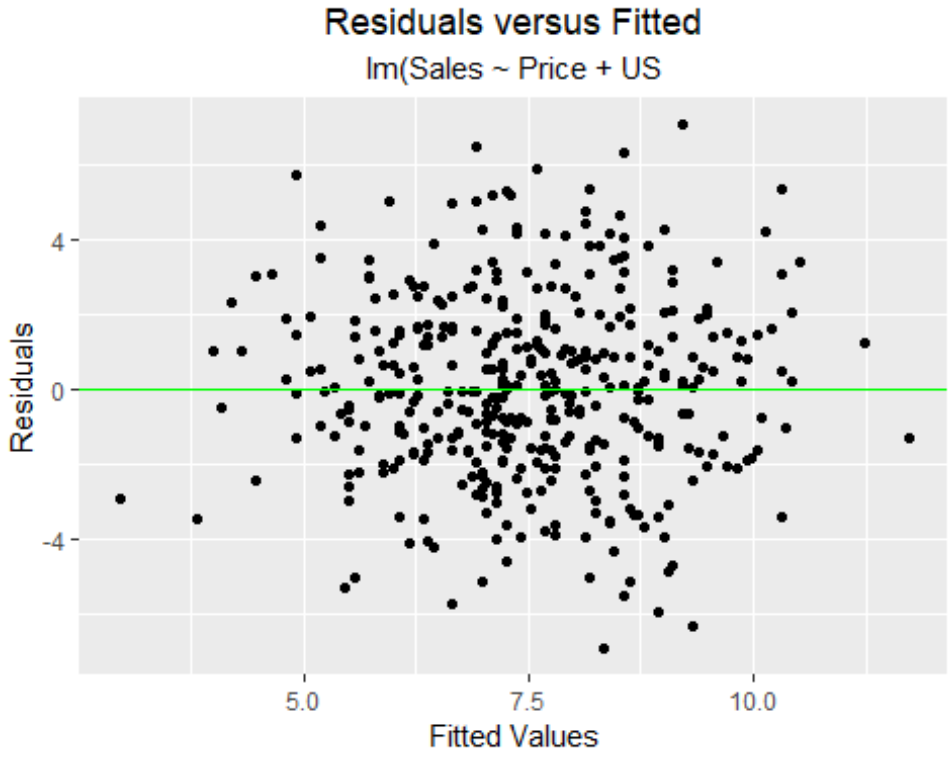
Results: The table below shows our 95% confidence interval for each coefficient from the 2nd model that was fit in exercise (e). Proportionally, we see that the largest confidence interval is see in the **US** variable.

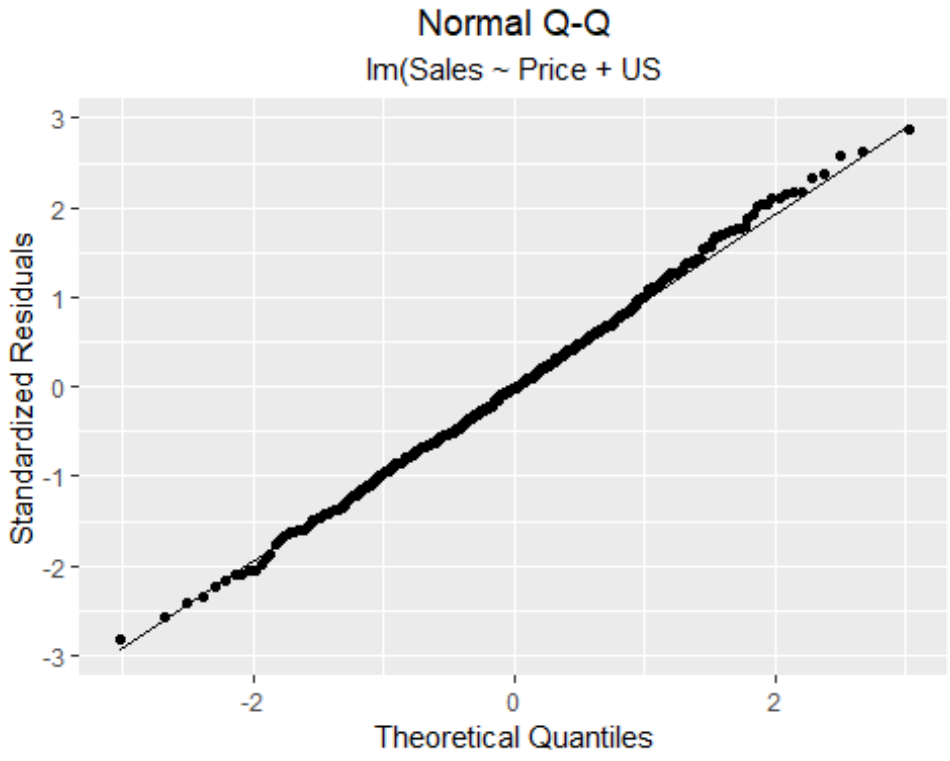
	2.5 %	97.5 %
(Intercept)	11.7903202	14.2712653
Price	-0.0647598	-0.0441954
USYes	0.6915196	1.7077663

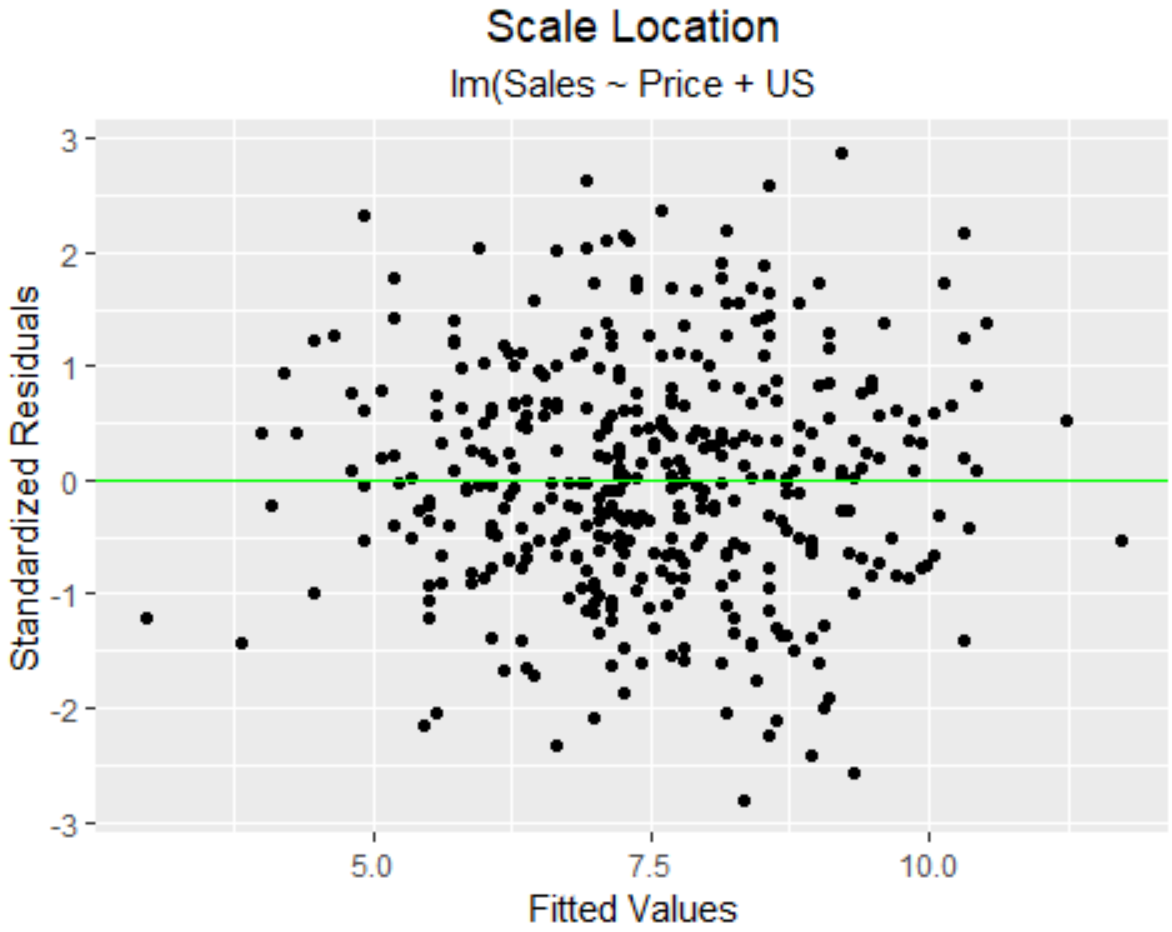
(h) Is there evidence of outliers or high leverarge observations in the model from (e)?

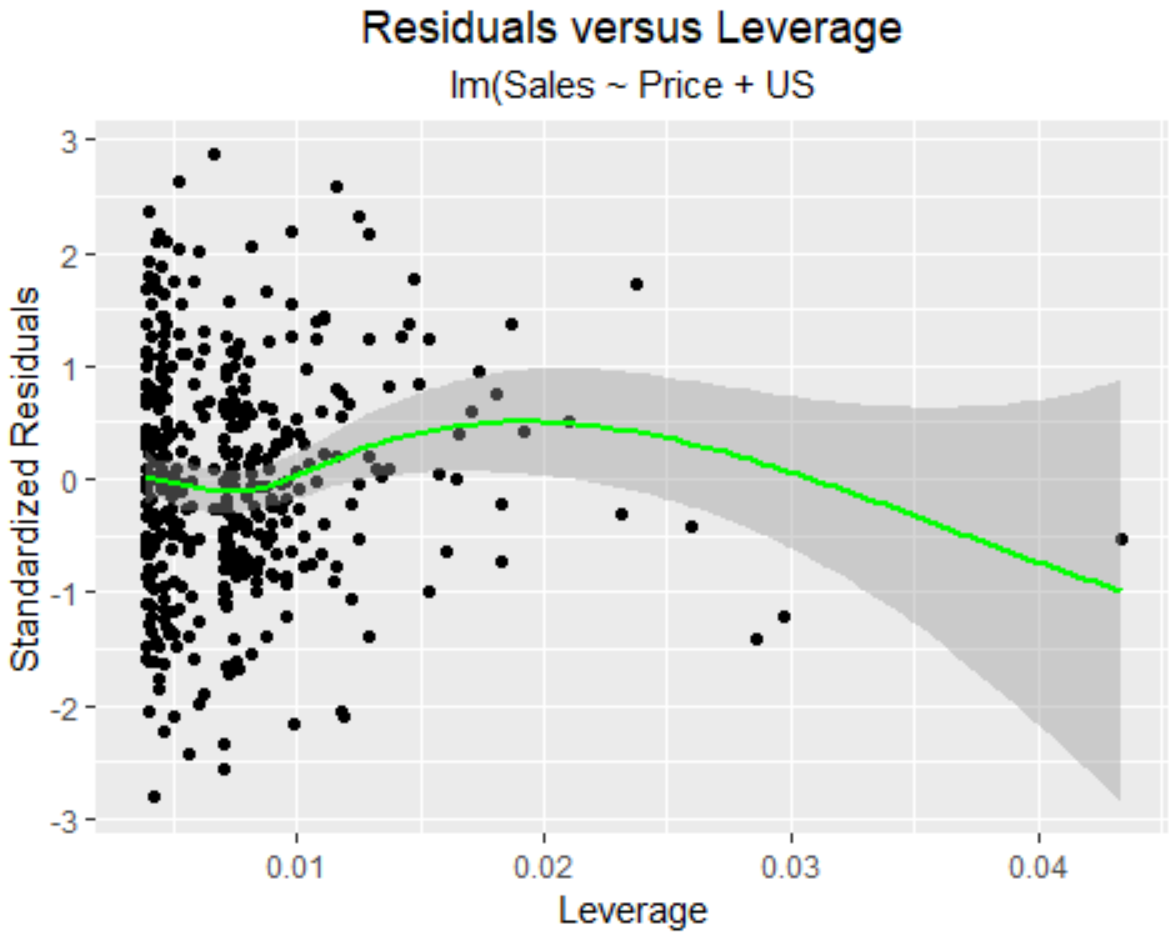
Results: Below I have the base R plots of the model from (e). I also used analogous ggplots, per homework instructions. In this application, I feel the base R plots are better / more informative. This is often not the case for me.

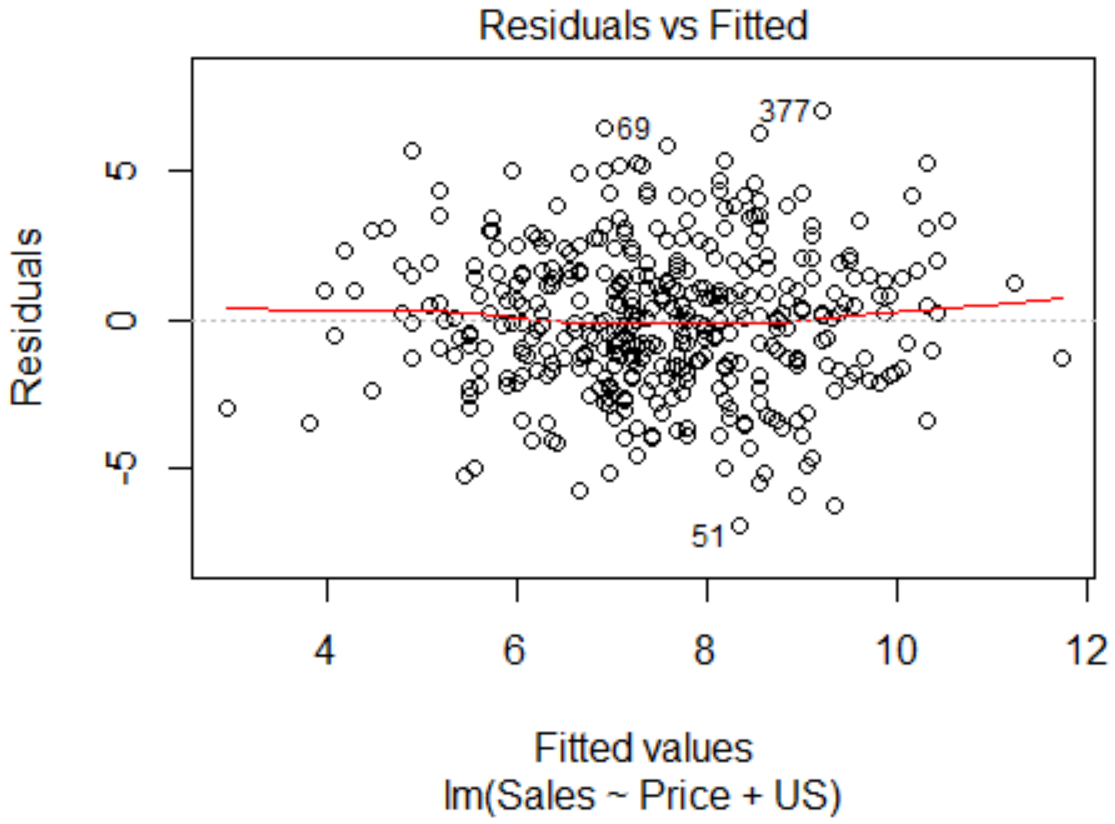
The plot of Residuals vs Leverage shows a few outliers and some leverage points.

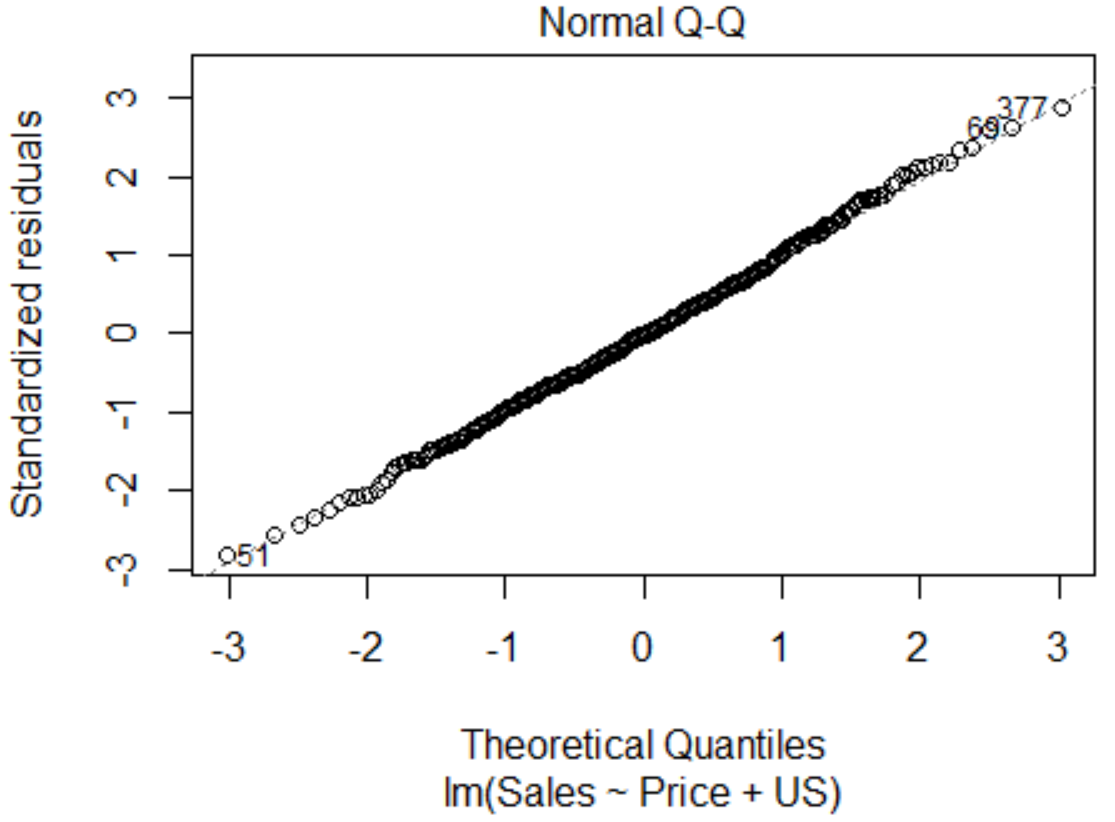


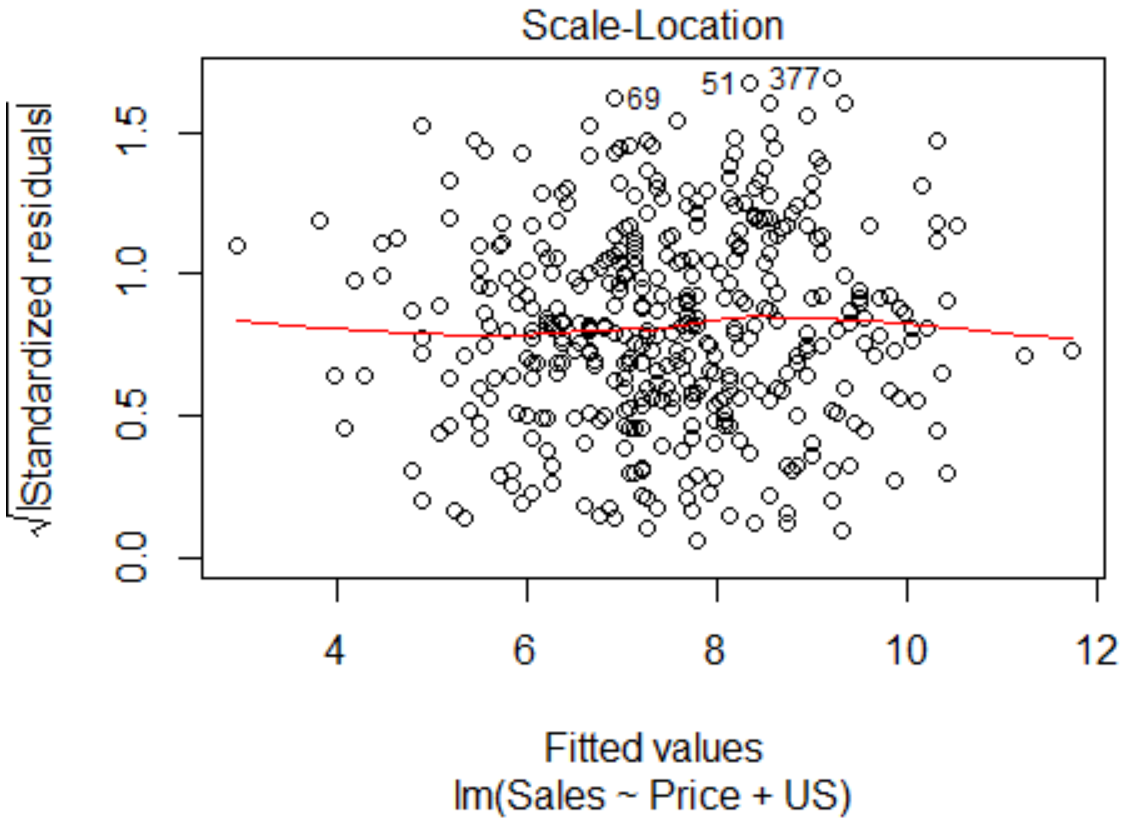


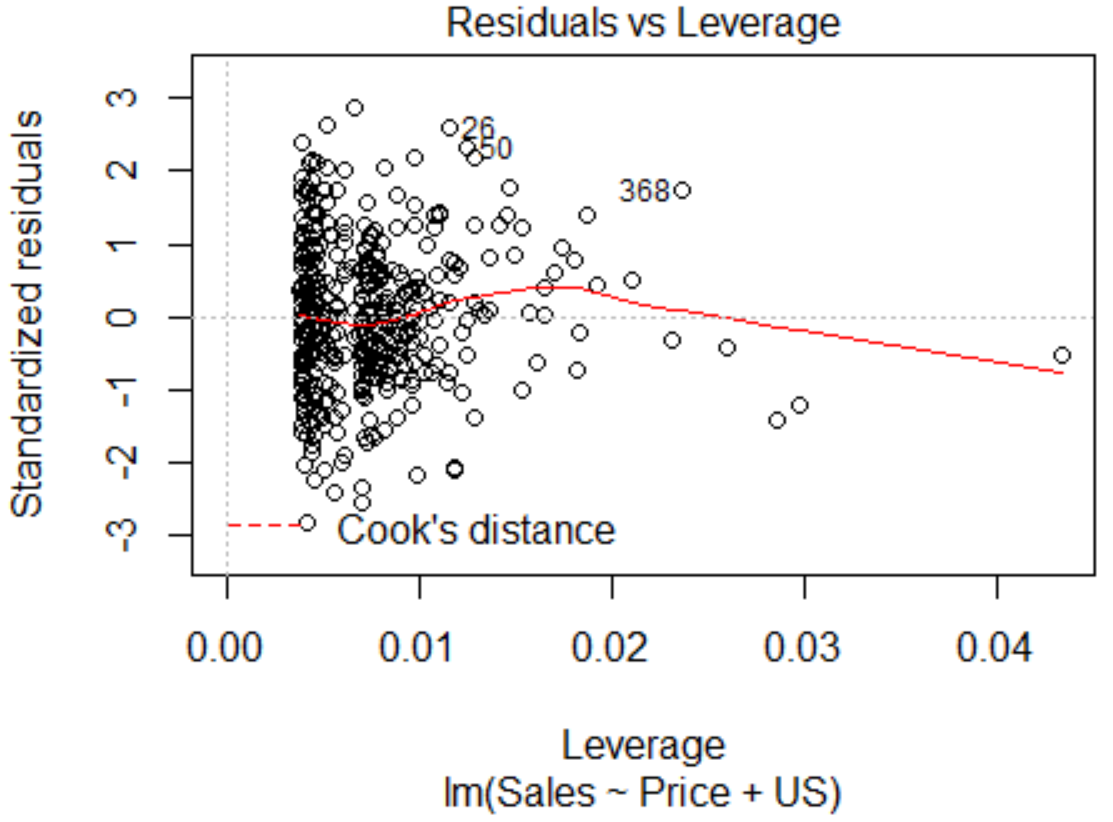

Scale Location
lm(Sales ~ Price + US
Standardized Residuals
3
2
1
0
-1
-2
-3
5.0
7.5
10.0
Fitted Values


Residuals versus Leverage
lm(Sales ~ Price + US
Standardized Residuals
3
2
1
0
-1
-2
-3
0.01
0.02
0.03
0.04
Leverage


Residuals vs Fitted
Residuals
Fitted values
lm(Sales ~ Price + US)
69
377
51
5
0
-5
4
6
8
10
12


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Sales ~ Price + US)
377
69
51
3
2
1
0
-1
-2
-3


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
69
51
377
4
6
8
10
12
Fitted values
lm(Sales ~ Price + US)


Residuals vs Leverage
Standardized residuals
-3
-2
-1
0
1
2
3
26
50
368
Cook's distance
0.00
0.01
0.02
0.03
0.04
Leverage
lm(Sales ~ Price + US)

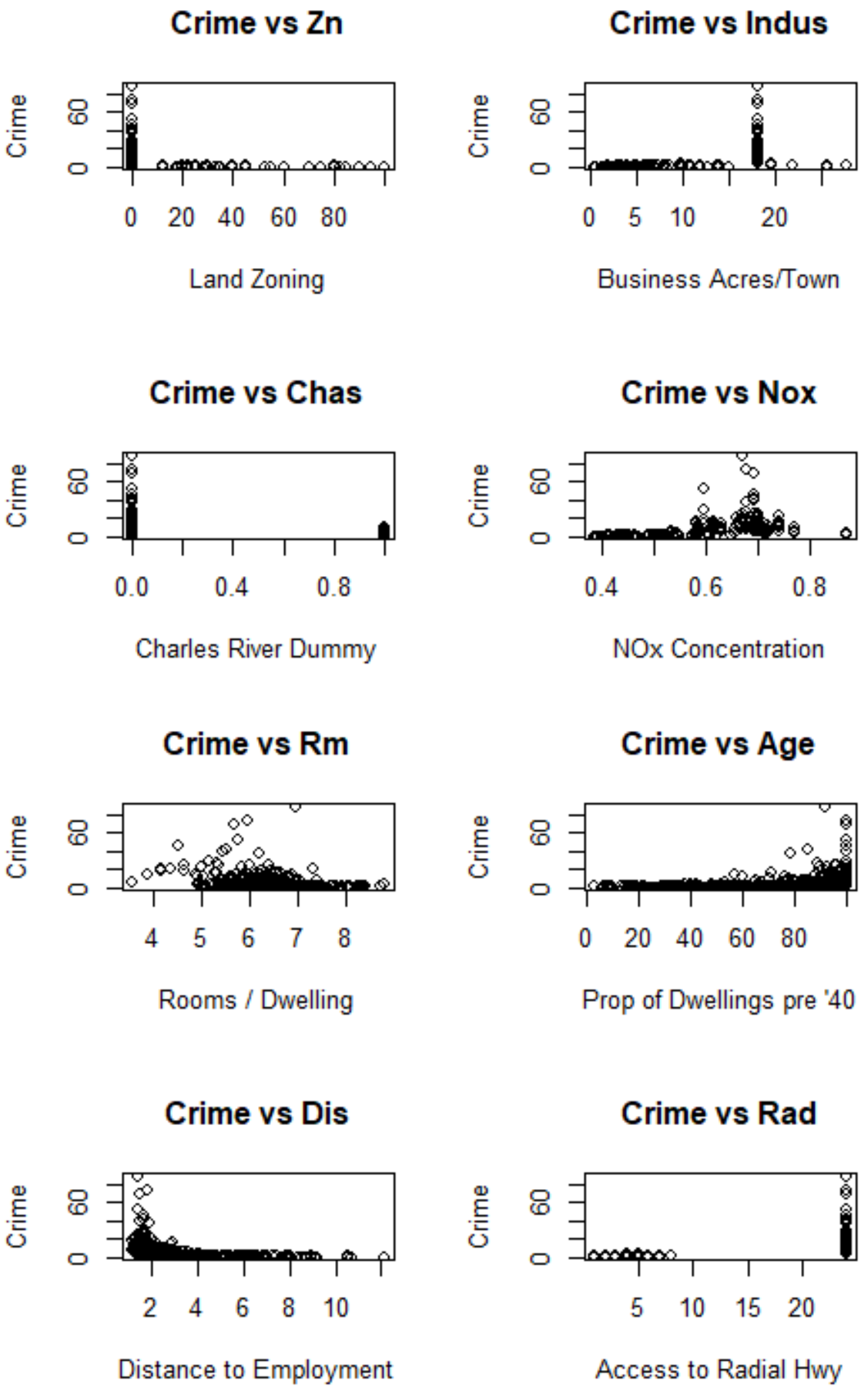
Problem 3.7.15, pg 126: This problem involves the *Boston* data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in the data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

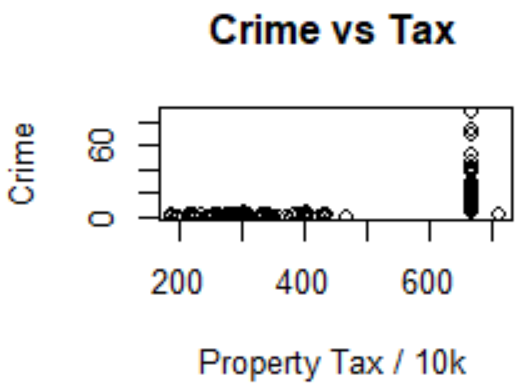
(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistcally significant association between the predictor and the response? Create some plots to back up your assertions.

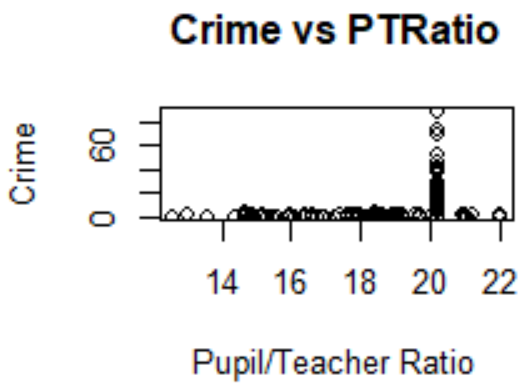
Results: First I fit simple linear regression models for each variable. I printed a table summarizing the p-values for each model (variable). As we see, all predictor variables besides *chas* are statistically significant at an alpha of 0.05.

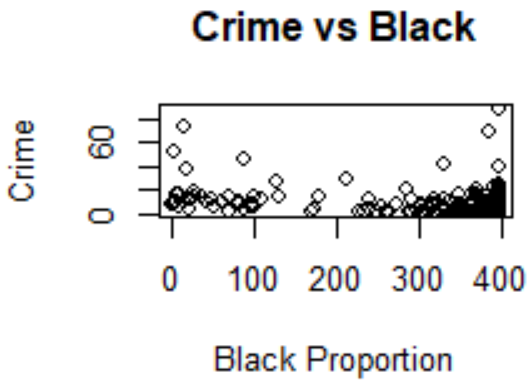
Next, I added scatterplots showing the relationship between each predictor and the response variable **crim**. These plots reinforce that the least statistically significant association exists between chas and crim. Analogous ggplots are added per homework instructions.

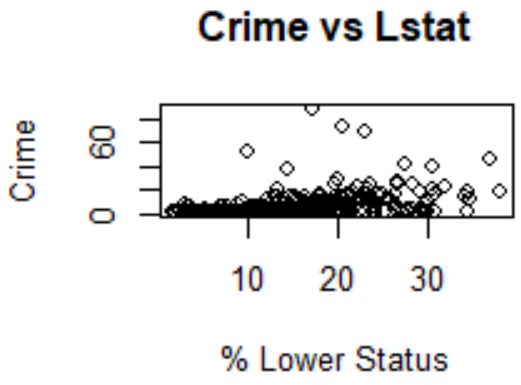
Variables	P-Value
zn	0.0000055
indus	0.0000000
chas	0.2094345
nox	0.0000000
rm	0.0000006
age	0.0000000
dis	0.0000000
rad	0.0000000
tax	0.0000000
ptratio	0.0000000
black	0.0000000
lstat	0.0000000
medv	0.0000000

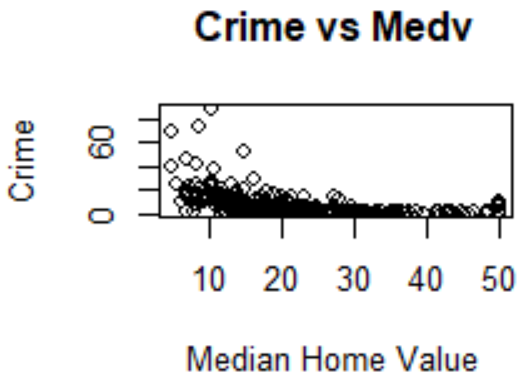

Crime vs Zn
Crime
0
60
0
20
40
60
80
Land Zoning
Crime vs Indus
Crime
0
60
0
5
10
20
Business Acres/Town
Crime vs Chas
Crime
0
60
0.0
0.4
0.8
Charles River Dummy
Crime vs Nox
Crime
0
60
0.4
0.6
0.8
NOx Concentration
Crime vs Rm
Crime
0
60
4
5
6
7
8
Rooms / Dwelling
Crime vs Age
Crime
0
60
0
20
40
60
80
Prop of Dwellings pre '40
Crime vs Dis
Crime
0
60
2
4
6
8
10
Distance to Employment
Crime vs Rad
Crime
0
60
5
10
15
20
Access to Radial Hwy

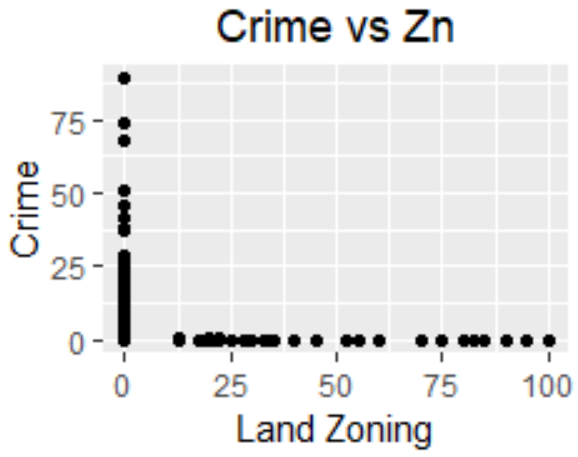

Crime vs Tax
Crime
0
60
200
400
600
Property Tax / 10k

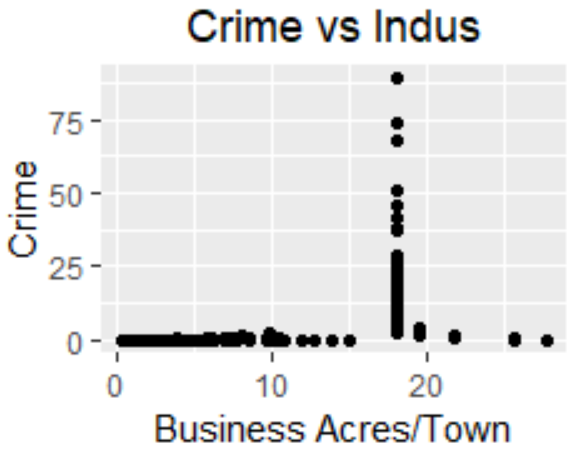

Crime vs PTRatio
Crime
0
60
14
16
18
20
22
Pupil/Teacher Ratio

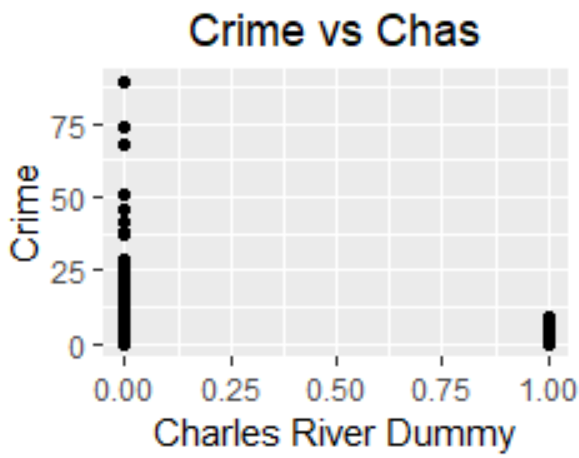

Crime vs Black
Crime
0
60
0
100
200
300
400
Black Proportion

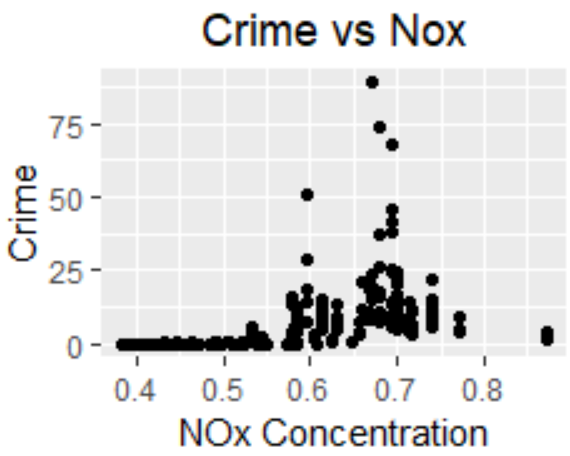

Crime vs Lstat
Crime
0
60
10
20
30
% Lower Status

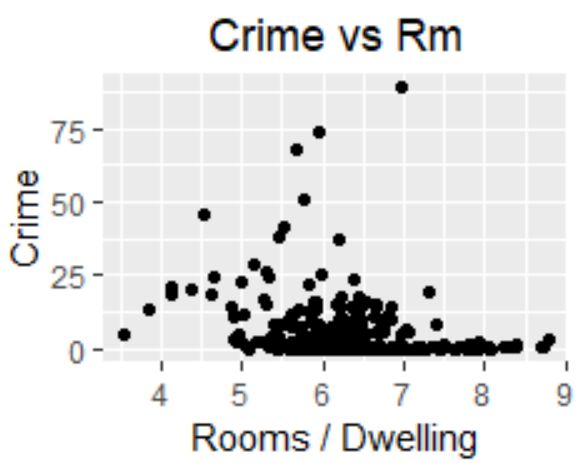

Crime vs Medv
Crime
0
60
10
20
30
40
50
Median Home Value

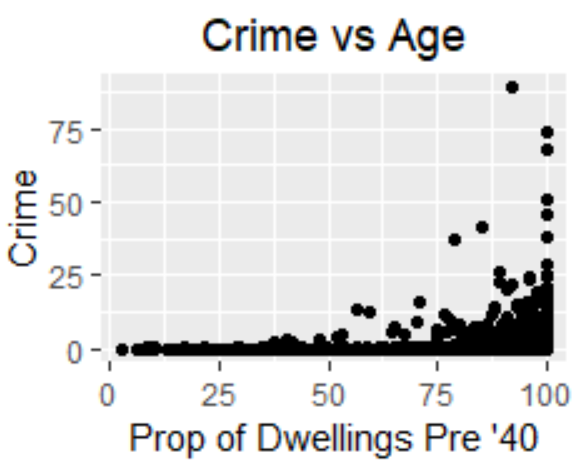

Crime vs Zn
Crime
0
25
50
75
0
25
50
75
100
Land Zoning

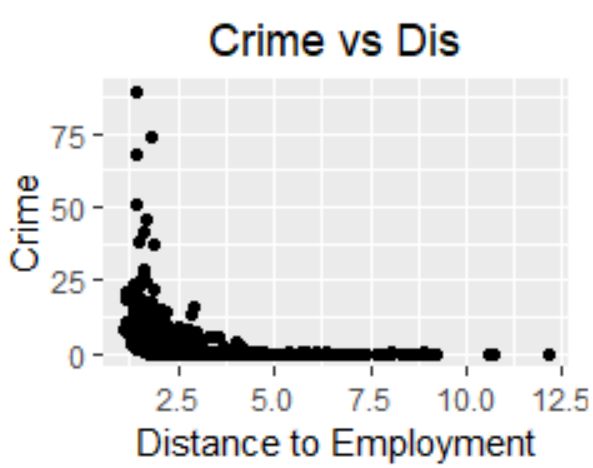

Crime vs Indus
Crime
0
25
50
75
0
10
20
Business Acres/Town

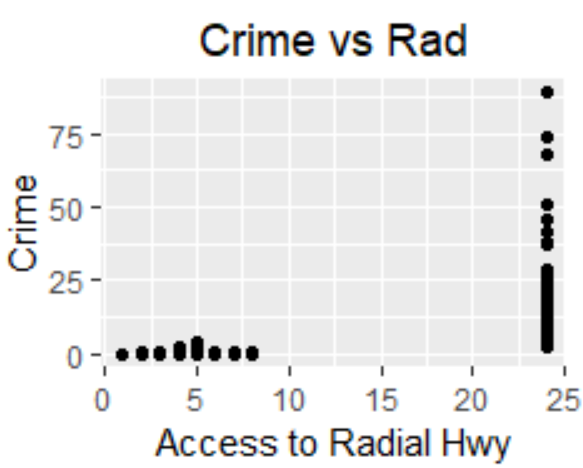

Crime vs Chas
Crime
0
25
50
75
0.00
0.25
0.50
0.75
1.00
Charles River Dummy

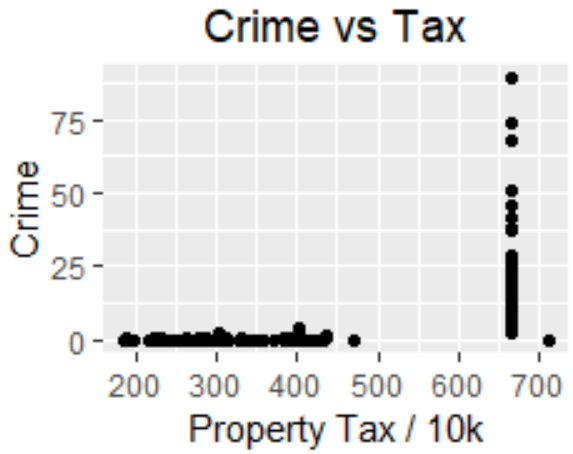

Crime vs Nox
Crime
0
25
50
75
0.4
0.5
0.6
0.7
0.8
NOx Concentration

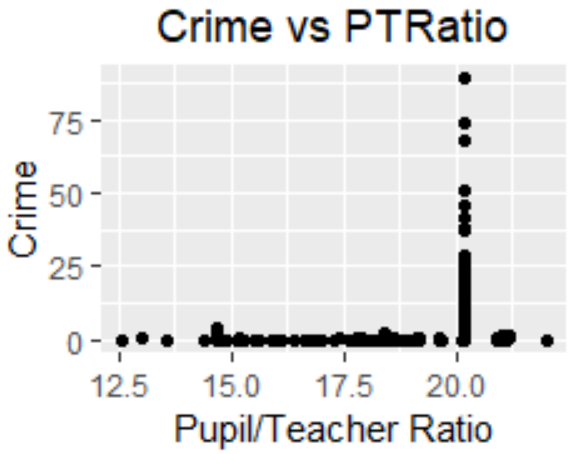

Crime vs Rm
Crime
0
25
50
75
4
5
6
7
8
9
Rooms / Dwelling

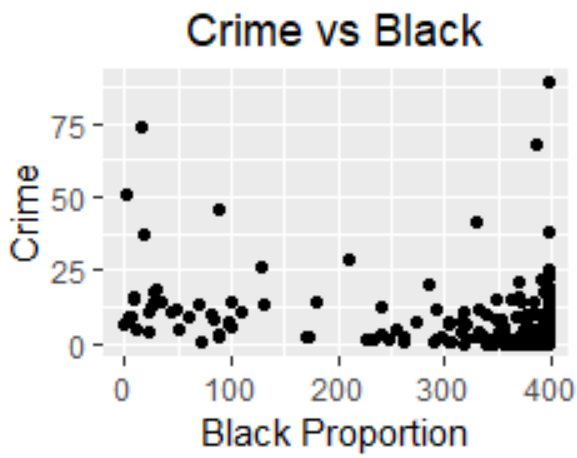

Crime vs Age
Crime
0
25
50
75
0
25
50
75
100
Prop of Dwellings Pre '40

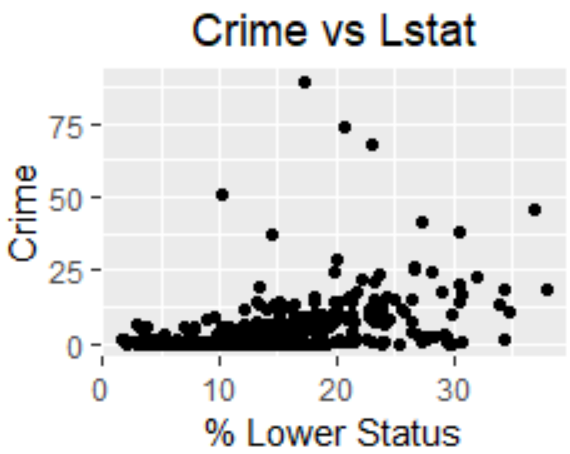

Crime vs Dis
Crime
0
25
50
75
2.5
5.0
7.5
10.0
12.5
Distance to Employment

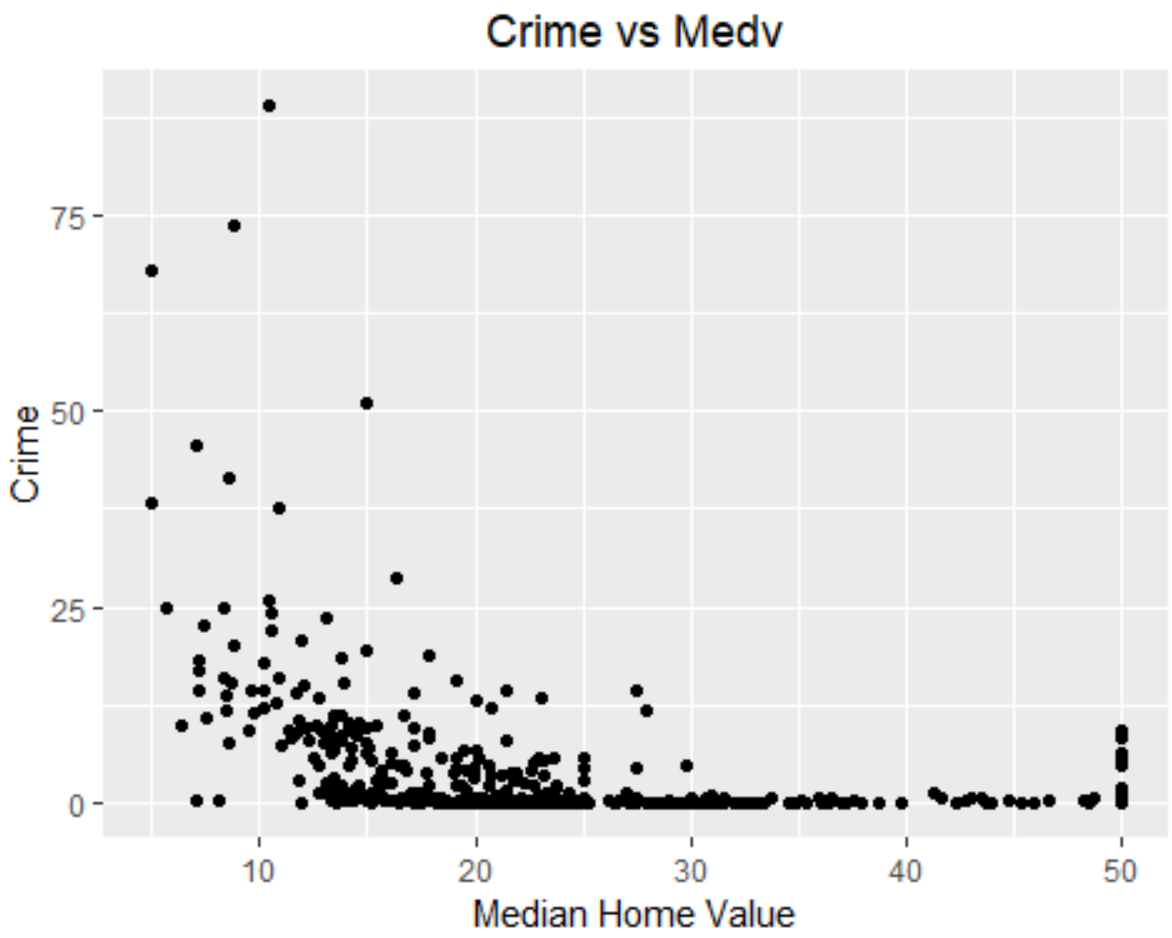

Crime vs Rad
Crime
0
25
50
75
0
5
10
15
20
25
Access to Radial Hwy


Crime vs Tax
Crime
75
50
25
0
200
300
400
500
600
700
Property Tax / 10k


Crime vs PTRatio
Crime
75
50
25
0
12.5
15.0
17.5
20.0
Pupil/Teacher Ratio


Crime vs Black
Crime
75
50
25
0
0
100
200
300
400
Black Proportion


Crime vs Lstat
Crime
75
50
25
0
0
10
20
30
% Lower Status


Crime vs Medv
Crime
75
50
25
0
10
20
30
40
50
Median Home Value

(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

$$H_0: \beta_j = 0$$

Results: Below is a table showing the p-values for each predictor of *crim* in the multiple regression model. From this table, we can reject the Null hypothesis for the following independent variables, assuming an alpha threshold of 0.05:

- zn (Proportion of residential land zoned for lots over 25k sq. ft.)

- dis (Weighted mean of distances to five Boston employment centres)

- rad (Index of accessibility to radial highways)

- black ($(1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town)

- medv (Median value of owner-occupied homes in $1000's)

P_Values of Predictors from Multiple Regression Model

	P-Values
(Intercept)	0.0189491
zn	0.0170249
indus	0.4442940
chas	0.5258670
nox	0.0511520
rm	0.4830888
age	0.9354878
dis	0.0005022
rad	0.0000000
tax	0.4637927
ptratio	0.1466113
black	0.0407023
lstat	0.0962084
medv	0.0010868

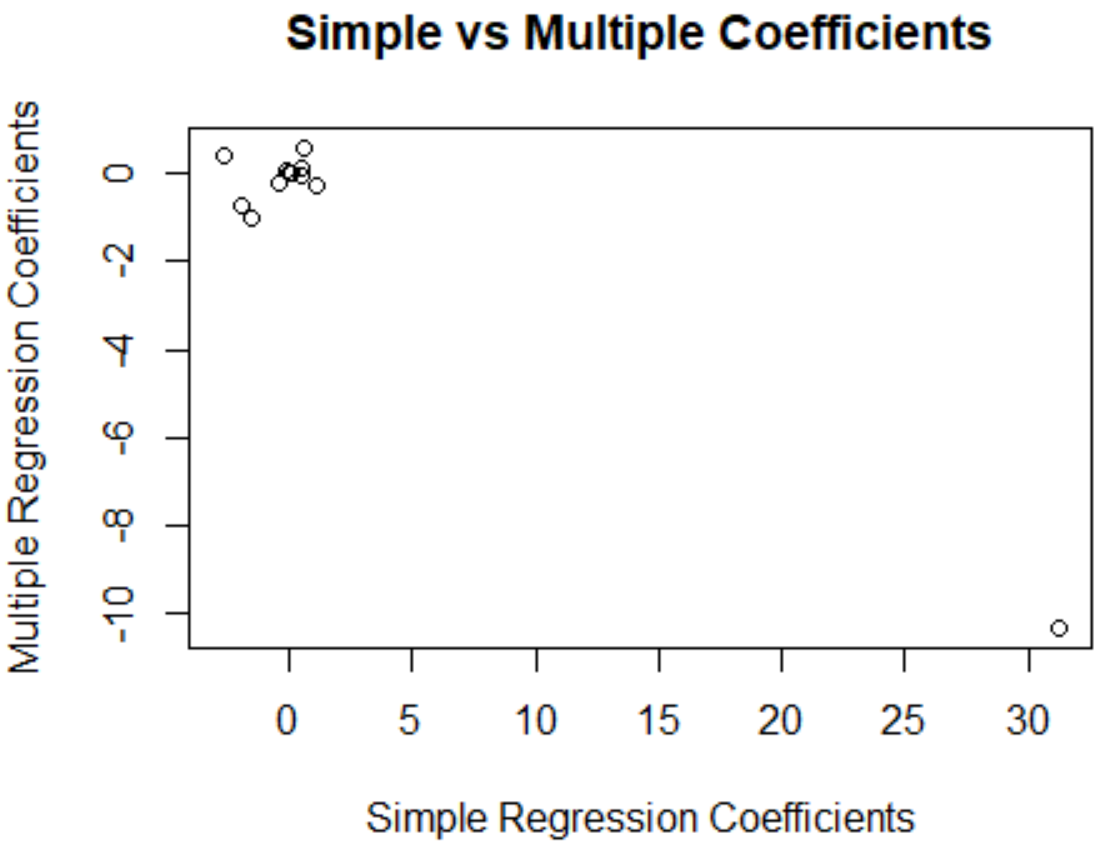
(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point on the plot. Its coefficient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

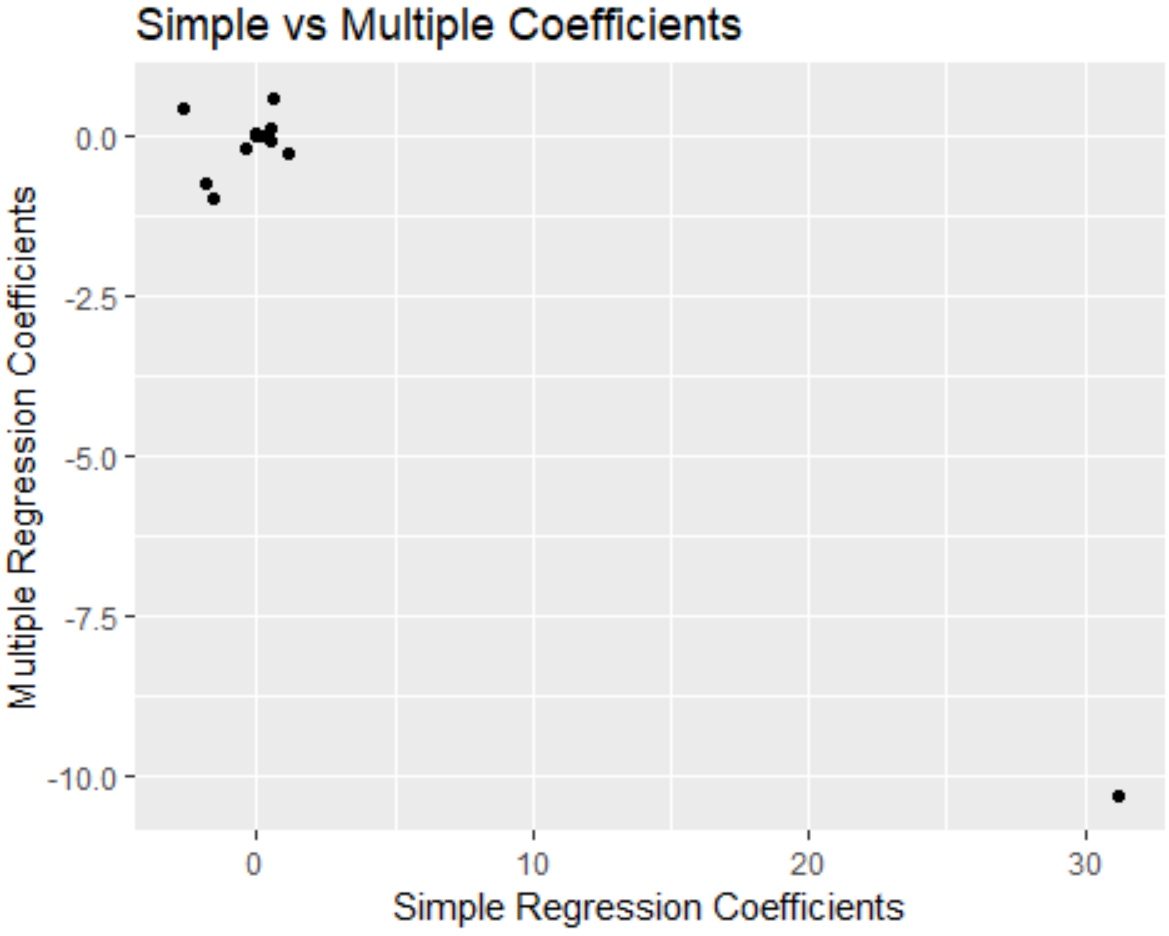
Results: First, I printed a table comparing the coefficients from each simple model and the coefficients for each predictor from the multiple regression model. As we can see, there are differences between the respective coefficients based on the model used.

Then, following the homework instructions, I printed a comparison scatterplot of the two sets of coefficients. The x axis shows the univariate regression coefficients and the y axis shows the multiple regression coefficients for each variable.

The reason for the difference is that, in the simple regression models, the coefficient doesn't take into account the other predictors. When moving to a multiple regression model, the coefficient takes into account the affect while keeping each of the other predictors fixed.

	Simple Regression Coefficients	Multiple Regression Coefficients
zn	-0.0739350	0.0448552
indus	0.5097763	-0.0638548
chas	-1.8927766	-0.7491336
nox	31.2485312	-10.3135349
rm	-2.6840512	0.4301305
age	0.1077862	0.0014516
dis	-1.5509017	-0.9871757
rad	0.6179109	0.5882086
tax	0.0297423	-0.0037800
ptratio	1.1519828	-0.2710806
black	-0.0362796	-0.0075375
lstat	0.5488048	0.1262114
medv	-0.3631599	-0.1988868


Simple vs Multiple Coefficients
Multiple Regression Coefficients
0
-2
-4
-6
-8
-10
0
5
10
15
20
25
30
Simple Regression Coefficients


Simple vs Multiple Coefficients
Multiple Regression Coefficients
0.0
-2.5
-5.0
-7.5
-10.0
0
10
20
30
Simple Regression Coefficients

(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

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

Results: First, I fit models including both a cubic and quadratic operator for each predictor. The table below summarizes the p-values from each predictor's association with the response.

The following variables show evidence of improved predictability using a quadratic or cubic version of the predictor (when going to 7 decimals):
- indus
- age
- rad
- lstat

The following show less predictability when using a quadratic or cubic version of the predictor (when going to 7 decimals):
- zn
- rm
- tax
- ptratio
- black

The following show no change (when going to 7 decimals):
- chas
- nox
- dis
- medv

P-Values for Linear and Non-Linear Associations

Variables	P-Value		
zn	0.0026123	I(dis^3)	0.0000000
I(zn^2)	0.0937505	rad	0.6234175
I(zn^3)	0.2295386	I(rad^2)	0.6130099
indus	0.0000530	I(rad^3)	0.4823138
I(indus^2)	0.0000000	tax	0.1097075
I(indus^3)	0.0000000	I(tax^2)	0.1374682
chas	0.2094345	I(tax^3)	0.2438507
nox	0.0000000	ptratio	0.0030287
I(nox^2)	0.0000000	I(ptratio^2)	0.0041196
I(nox^3)	0.0000000	I(ptratio^3)	0.0063005
rm	0.2117564	black	0.1385871
I(rm^2)	0.3641094	I(black^2)	0.4741751
I(rm^3)	0.5085751	I(black^3)	0.5436172
age	0.1426608	lstat	0.3345300
I(age^2)	0.0473773	I(lstat^2)	0.0645874
I(age^3)	0.0066799	I(lstat^3)	0.1298906
dis	0.0000000	medv	0.0000000
I(dis^2)	0.0000000	I(medv^2)	0.0000000
		I(medv^3)	0.0000000