Homework 2

Andrew Boschee

No Outside Resources

1. Question 3.7.5 pg 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?

$$a_(i')$$

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

Carseats Multiple Regression

2. Question 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.

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

Results: For negative correlation with price. For every unit increase of price there is a decrease of .054 in sales units. Regarding Urban variable being binary, if it is urban, there is -.0219 unit decrease in sales units. Similar to the Urban variable, using binary (0/1) for X there will be in increase of 1.2 units of sales.

	Estimate	Std. Error	t value	Pr(>|t|)
Price	-0.0544588	0.0052419	-10.3892320	0.0000000
UrbanYes	-0.0219162	0.2716503	-0.0806778	0.9357389
USYes	1.2005727	0.2590415	4.6346731	0.0000049

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

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

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

$$H_0 : \beta_j = 0$$

From output below we can reject the null hypothesis for the Price and Us variables at a signficance level of .05

Table 2: P-Values of Predictors to Reject Null Hypothesis

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

Multiple Regression with Signficant Variables

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.

A summary of the model is given showing the two significant predictors.

```
##
## 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
```

R-Squared Model Comparison

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

The two significant variables make up roughly 24% of the variability of the dependent variable(Sales). The output below compares the r-squared output when using both 2 predictors and 3 predictors. There is very little difference when including the Urban dependent variable and reinforces our assumption that it is not significant regarding sales.

Table 3: Comparison of R-Squared Between Models

Model 1 (3 Predictors)	Model 2 (2 Predictors)
0.2392754	0.2392629

Confidence Intervals

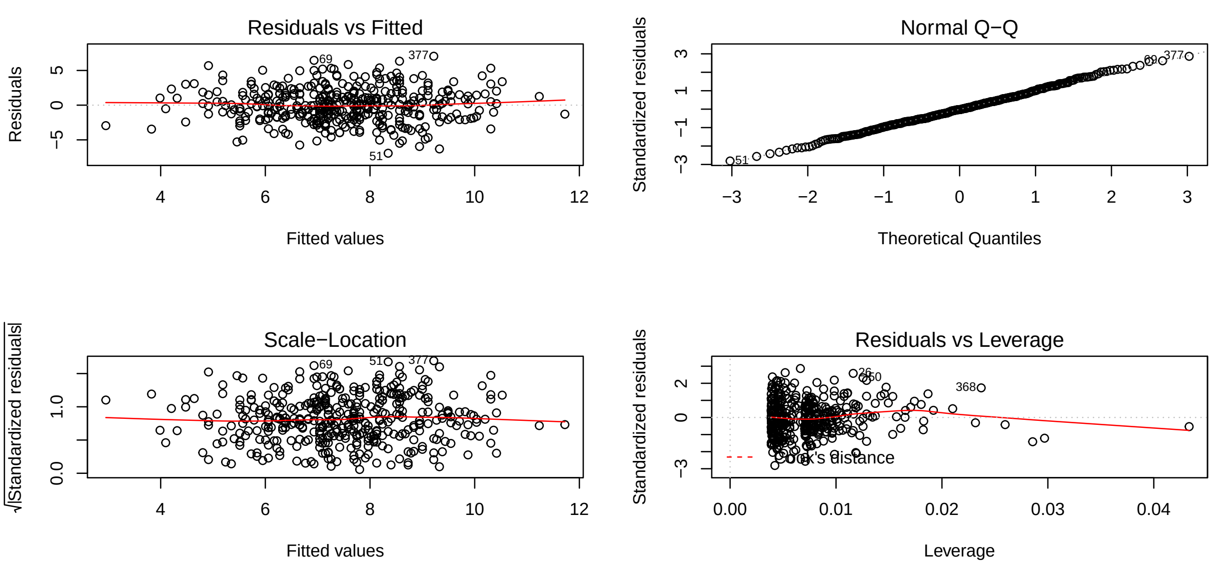
g. Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

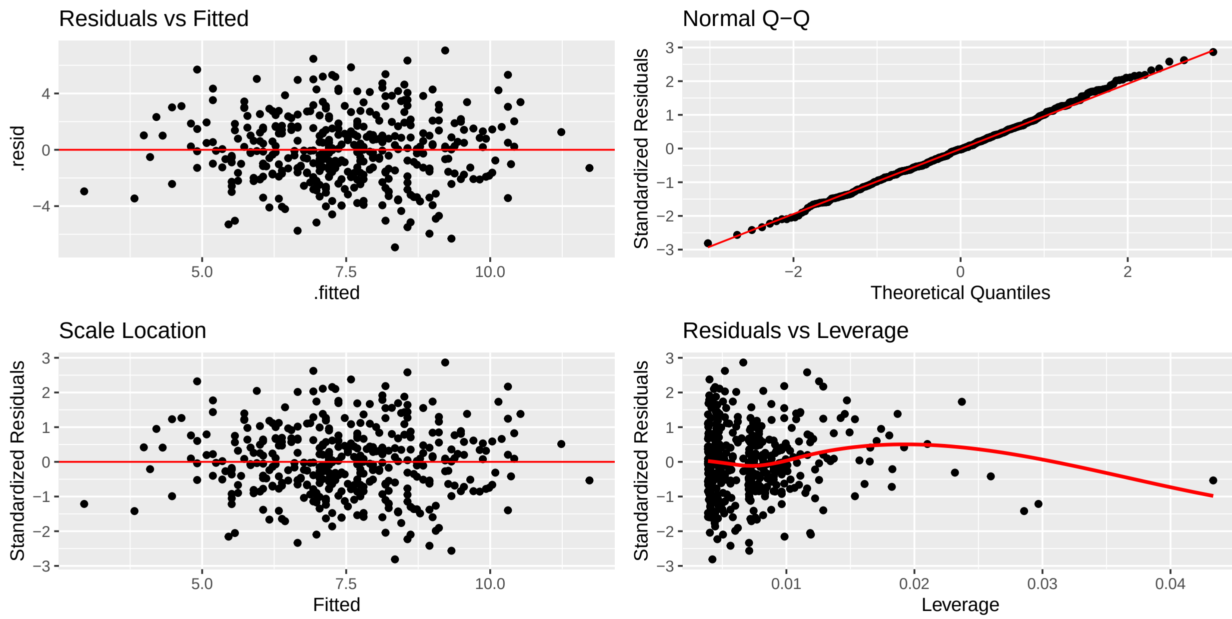
Comparing the outcome of the confidence intervals, we can immediately see a much broader interval for the US variable than the Price variable.

	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)?

There are not many outliers when looking at the residuals and QQ-Plot. The only outliers that stand out come from the Residuals vs Leverage plot on the high end of x-axis(Leverage)





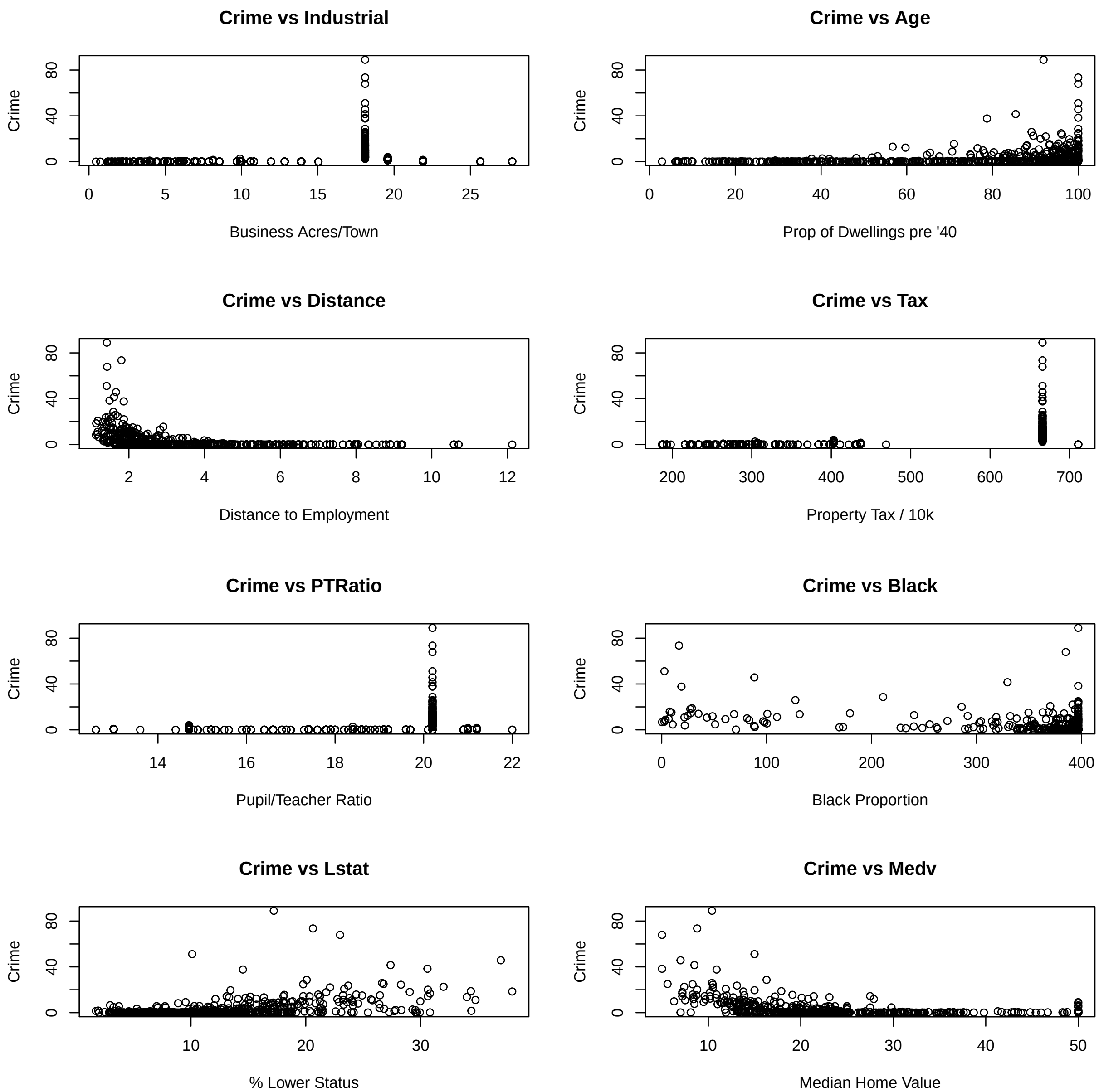
Boston - Simple Linear Regression

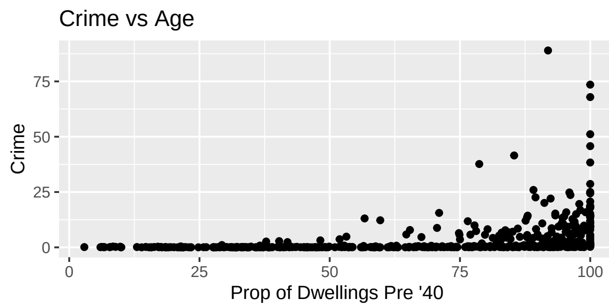
3. Question 3.7.15 pg 126 - This problem involves the *Boston* data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in the data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

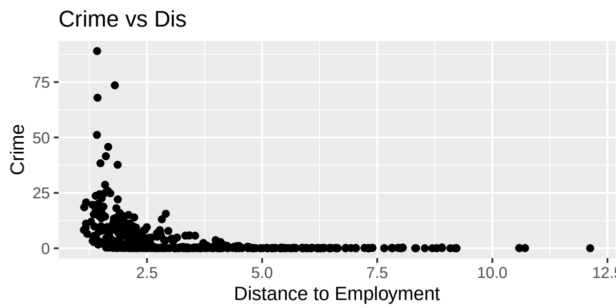
(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistcally significant association between the predictor and the response? Create some plots to back up your assertions.

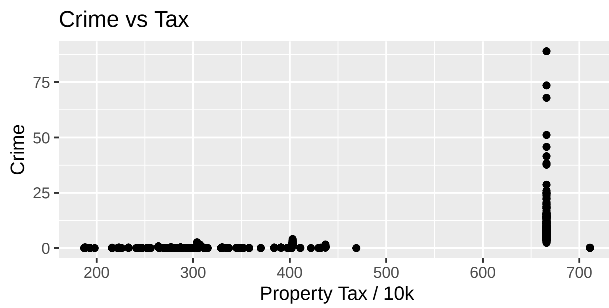
Table 5: P-Value by Variable

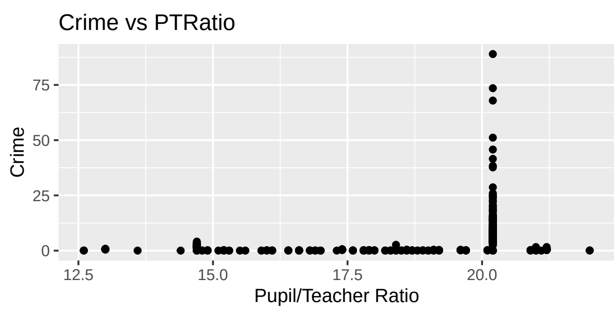
Variable	P-Value
Zone	5.50647210767964e-06
Industrial	1.45034893302756e-21
CharlesRiver	0.209434501535197
Nitrogen Concentration	3.7517392603569e-23
Rooms	6.34670298468749e-07
Age(Before 1940)	2.85486935024409e-16
Distance	8.5199487669261e-19
Highway Access	2.69384439818633e-56
Tax Rate	2.35712683525685e-47
Pupil-Teacher Ratio	2.94292244735967e-11
BlacksProp	2.48727397377375e-19
LowerStatus	2.65427723147327e-27
MedianValue	1.17398708219449e-19

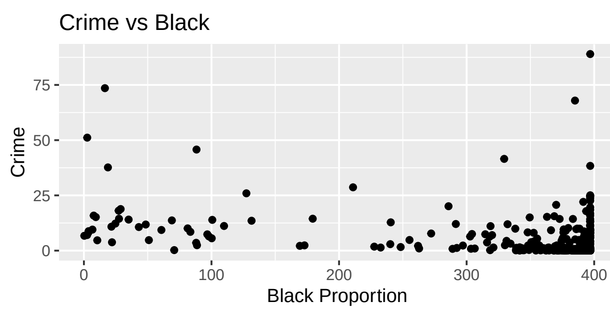

Crime vs Industrial
Crime
Business Acres/Town
Crime vs Age
Crime
Prop of Dwellings pre '40
Crime vs Distance
Crime
Distance to Employment
Crime vs Tax
Crime
Property Tax / 10k
Crime vs PTRatio
Crime
Pupil/Teacher Ratio
Crime vs Black
Crime
Black Proportion
Crime vs Lstat
Crime
% Lower Status
Crime vs Medv
Crime
Median Home Value

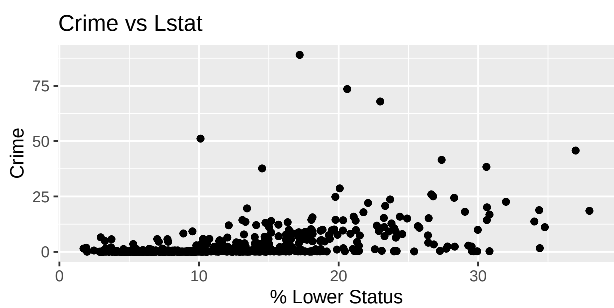


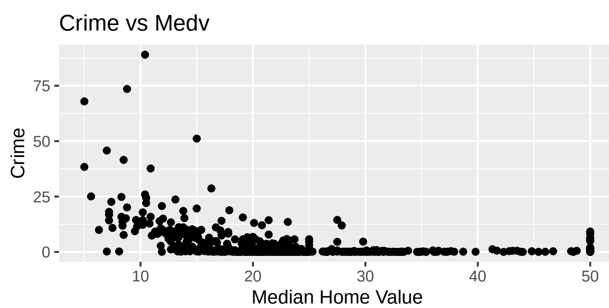












I think there is a lot to take away from the plots above. I only included the plots that are fairly easy for interpretation without getting too deep. Many of the other interactions I believe are open to interpretation to a much greater extent with too many assumptions.

One that I am more curious about is the Crime vs Pupil/Teacher Ratio. The first thing that comes to mine is that classes are often fairly similar in size with a somewhat standardized classroom setting and limitations on capacity. There is very heavy chunk around the ratio of 20:1. This doesn't come at much of a surprise since I am assuming that is a very common ratio. I think this would be an interesting area to dig deeper into.

The other topics that stand out probably have some correlation with each other such as median home value, lower status, and distance to employment. This may also have quite a bit of geographic factoring where the population of race, home value, and status come into play.

Multiple Regression

(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?

$$H_0 : \beta_j = 0$$

From the table below using a threshold of .05 for alpha, we can reject the null hypothesis for five variables: zn, dis, rad, black, and medv.

At first glance, I can agree with thir result. Was expecting the ptratio as well and possibly the lstat variable.

Table 6: Multiple Regression P-Values

	P-Value
(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

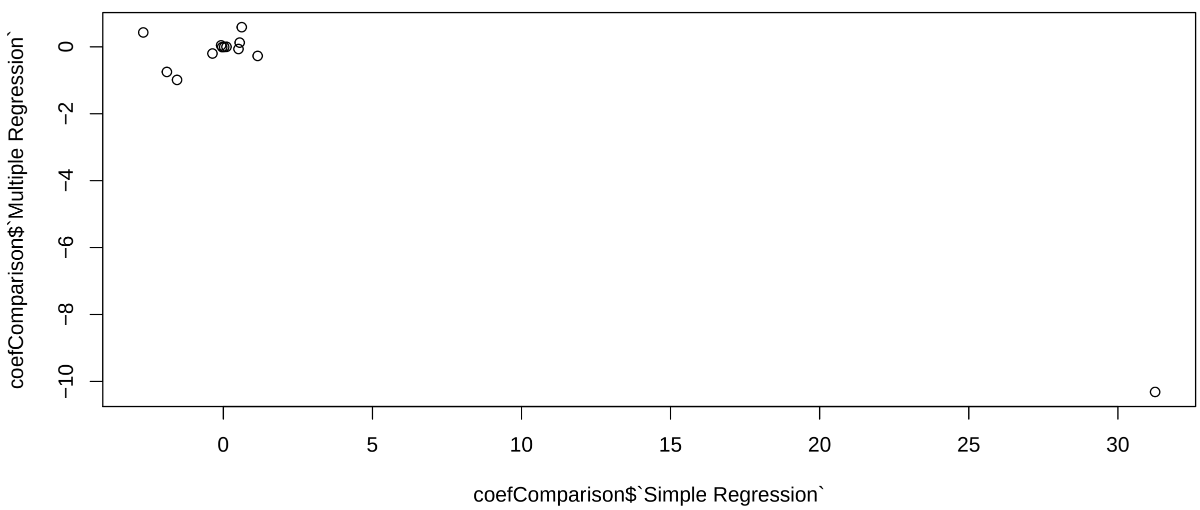
Coefficient Comparison

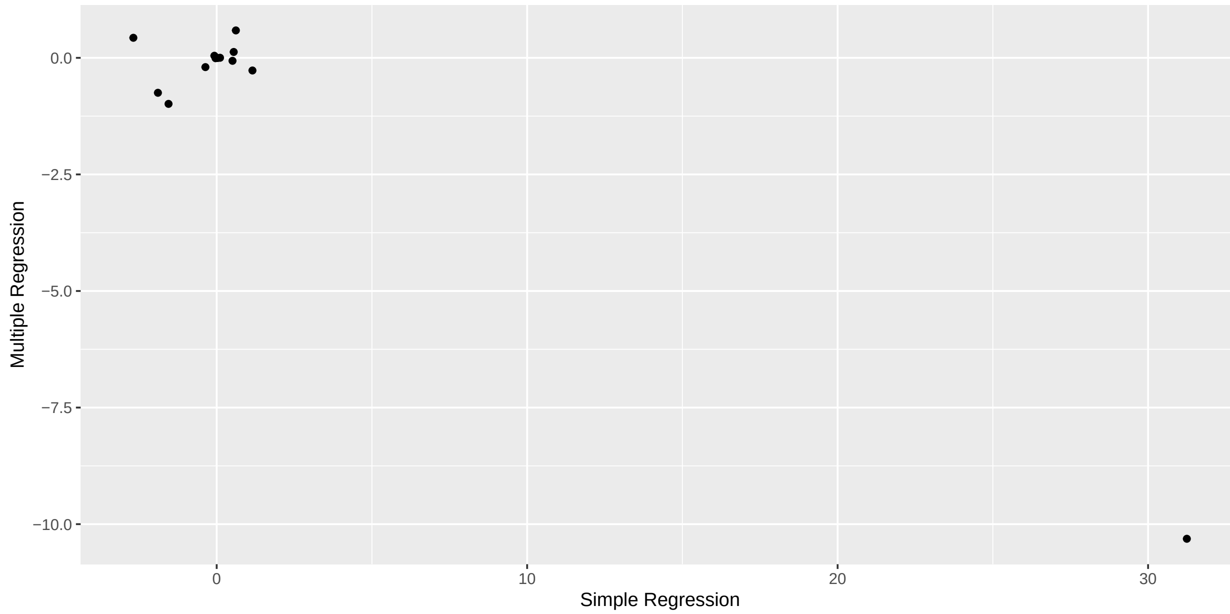
(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point on the plot. Its coefficient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

This part is a little ugly in my opinion and the only thing that should be taken away is that when adding in multiple dependent variables, it can play a big role in the coefficients of each predictor. Clearly the nox variable's coefficient drastically changes when it is not the only predictor used in the model. Some variables even go from having a positive coefficient to a negative coefficient.

Table 7: Coefficient Comparison

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





Non-Linear Associations

(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

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

We can see from the p-values that there is a change in whether a predictor is seen as significant when going from simple linear to cubic and quadratic. Again, some variables show no or very little change (chas, nox, dis, medv) while some that seen as significant before are no longer and vice versa.

Noticeable Variables : age, lstat, zn

Note: This is looking back seven decimal points from output when saying no change.

Table 8: Significance of Quadratic and Cubic Predictors

P-Value	Variable
0.0026123	Zn
0.0937505	Zn^2
0.2295386	Zn^3
0.0000530	Indus
0.0000000	Indus^2
0.0000000	Indus^3
0.2094345	Chas
0.0000000	Nox
0.0000000	Nox^2
0.0000000	Nox^3
0.2117564	Rm
0.3641094	Rm^2
0.5085751	Rm^3
0.1426608	Age
0.0473773	Age^2
0.0066799	Age^3
0.0000000	Dis
0.0000000	Dis^2
0.2143267	Dis^3
0.6234175	Rad
0.6130099	Rad^2
0.4823138	Rad^3
0.1097075	Tax
0.1374682	Tax^2
0.2438507	Tax^3
0.0030287	Ptratio
0.0041196	Ptratio^2
0.0063005	Ptratio^3
0.1385871	Black
0.4741751	Black^2
0.5436172	Black^3
0.3345300	Lstat
0.0645874	Lstat^2
0.1298906	Lstat^3
0.0000000	Medv
0.0000000	Medv^2
0.0000000	Medv^3