Homework 2

Alex Soupir

January 24, 2020

Packages: ISLR, ggplot2, MASS

Collaborators:

Answer all questions specified on the problem and include a discussion on how your results answered/addressed the question.

Submit your **.rmd** file with the knitted **PDF** (or knitted Word Document saved as a PDF). If you are having trouble with .rmd, let us know and we will help you, but both the .rmd and the PDF are required.

This file can be used as a skeleton document for your code/write up. Please follow the instructions found under Content for Formatting and Guidelines. No code should be in your PDF write-up unless stated otherwise.

For any question asking for plots/graphs, please do as the question asks as well as do the same but using the respective commands in the GGPLOT2 library. (So if the question asks for one plot, your results should have two plots. One produced using the given R-function and one produced from the GGPLOT2 equivalent). This doesn't apply to questions that don't specifically ask for a plot, however I still would encourage you to produce both.

You do not need to include the above statements.

Please do the following problems from the text book ISLR.

1. Question 3.7.5 pg 121

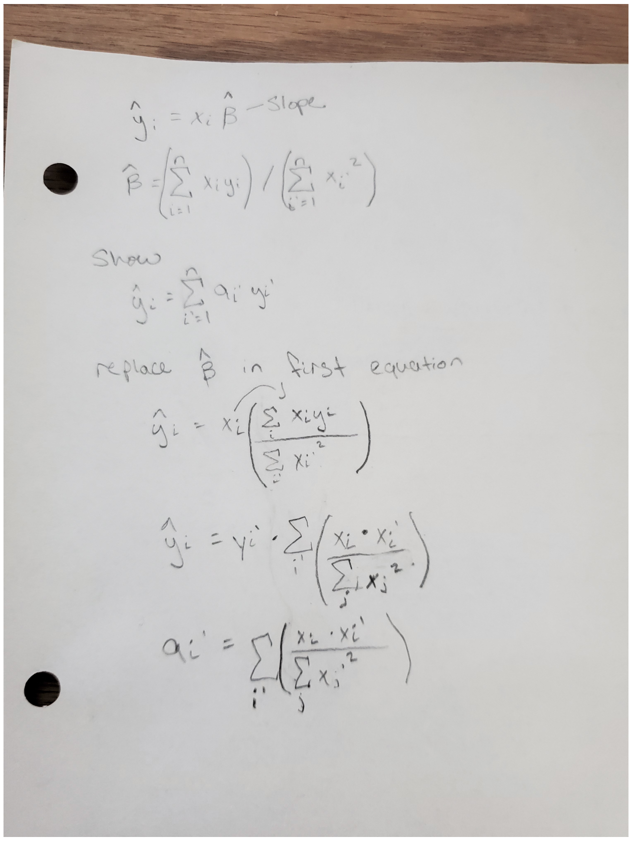
Show that we can write the equation - Page 2

2. Question 3.7.10 pg 123

This question 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!

```
##
## 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
##
```



$\hat{y}_i = x_i \hat{\beta}$ —Slope

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

Show

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

replace $\hat{\beta}$ in first equation

$$\hat{y}_i = x_i \left(\frac{\overset{j}{\sum_i} x_i y_i}{\sum_{i'} x_{i'}^2} \right)$$

$$\hat{y}_i = y_{i'} \cdot \sum_{i'} \left(\frac{x_i \cdot x_{i'}}{\sum_j x_j^2} \right)$$

$$a_{i'} = \sum_{i'} \left(\frac{x_i \cdot x_{i'}}{\sum_j x_{j'}^2} \right)$$

Figure 1: Question 5

```
## 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
```

For price, there seems to be a significant relationship. As the price increases, the sales of that car seat decreases (for every increase in price by 1, sales decreases by 54 seats for that location). For Urban - Yes, the p-value is not significant and therefore whether the location is urban or not doesn't significant contribute to this model. Finally, US - Yes, the p-value is significant indicating there is a relationship, and if the location is in the US sales will increase since the coefficient is positive (since sales are in the thousands and the coefficient for US - Yes is 1.2006, it would be expected that the sales will increase by 1200 seats if being sold in the US.

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

Car seat sales = 13.04 + Price * -0.054 + UrbanYes * -0.022 + USYes * 1.2006.

d) For which of the predictors can you reject the null hypothesis that the coefficient is equal to 0?

The predictors that we can reject the null hypothesis that the coefficient is equal to 0, is for the *Price* and *USYes* coefficients

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

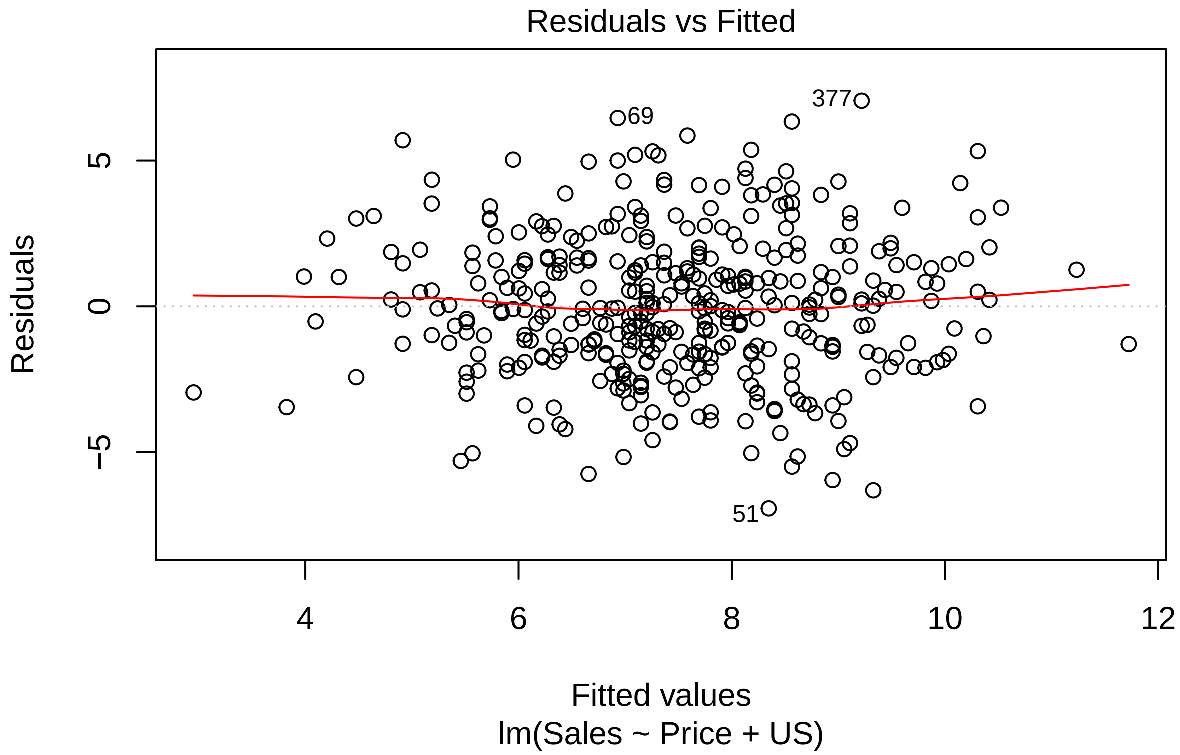
```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## 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
```

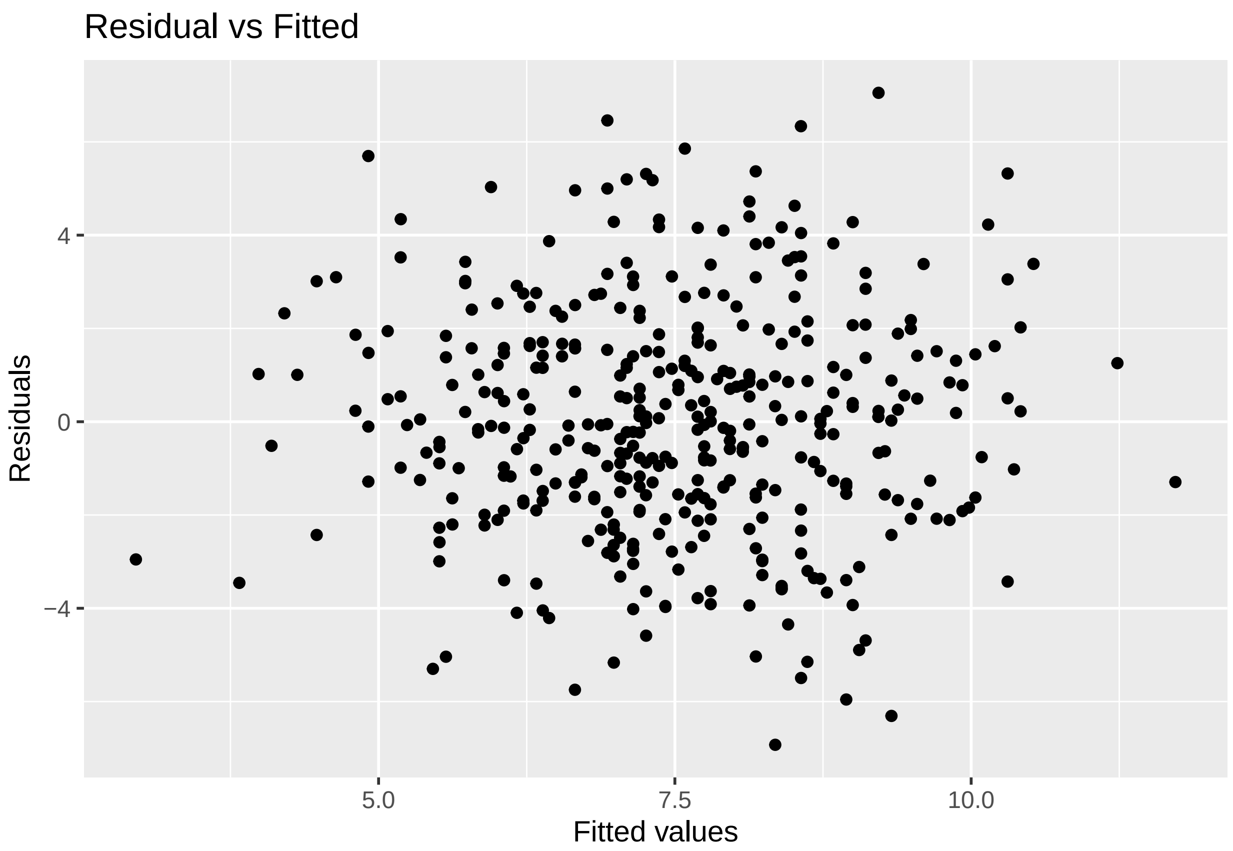
f) How well do the models in part A and part E fit the data?

Both models have R^2 of 0.2393 so they both fit the data in a similar way. However, the F-statistic is greater for the part E model so it is slightly better. The F-statistic shows more insight to the model than does the p-value of the model because the p-values in R can only display 2.2e-16, and they both show that.

g) Using the model from part E, obtain 95% confidence intervals for each of the coefficient(s).

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


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



With the model created in part E, there are a few value that are outliers.

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 this data set. In other words, per capita crime 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.

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

```
## Crime predicted by residential land zoned for lots over 25k sq.ft
```

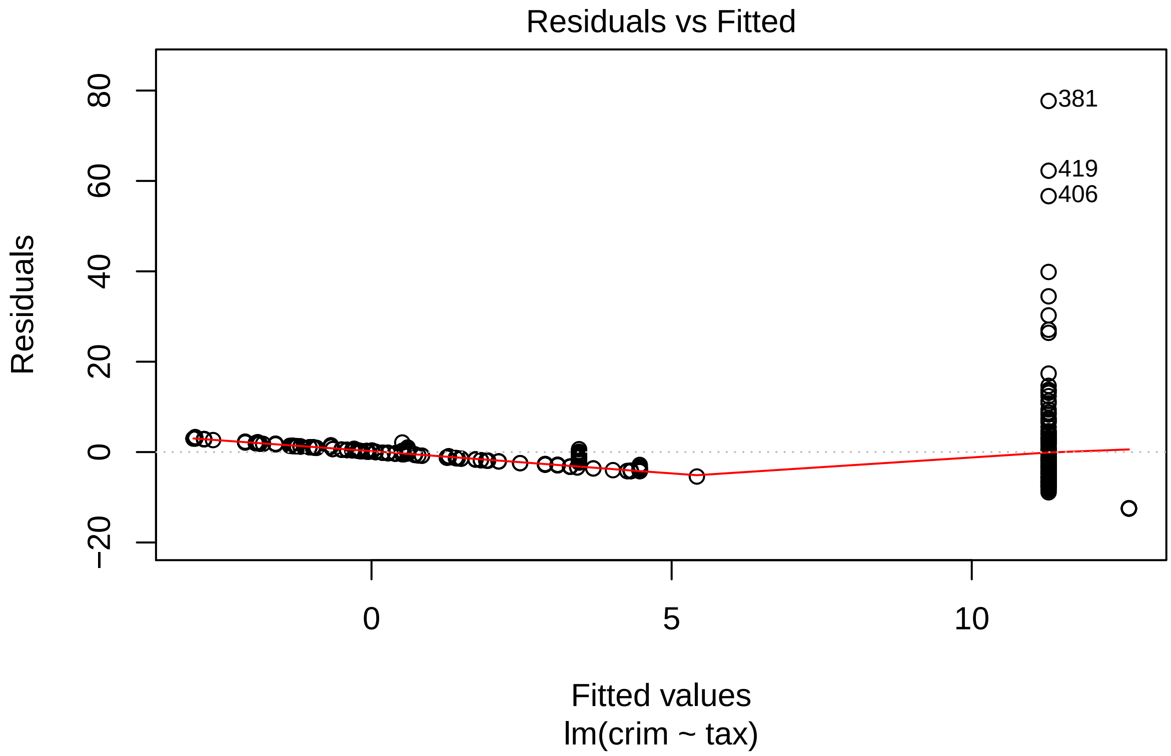
```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  4.45369376  0.4172178 10.674746 4.037668e-24
## zn          -0.07393498  0.0160946 -4.593776 5.506472e-06
```

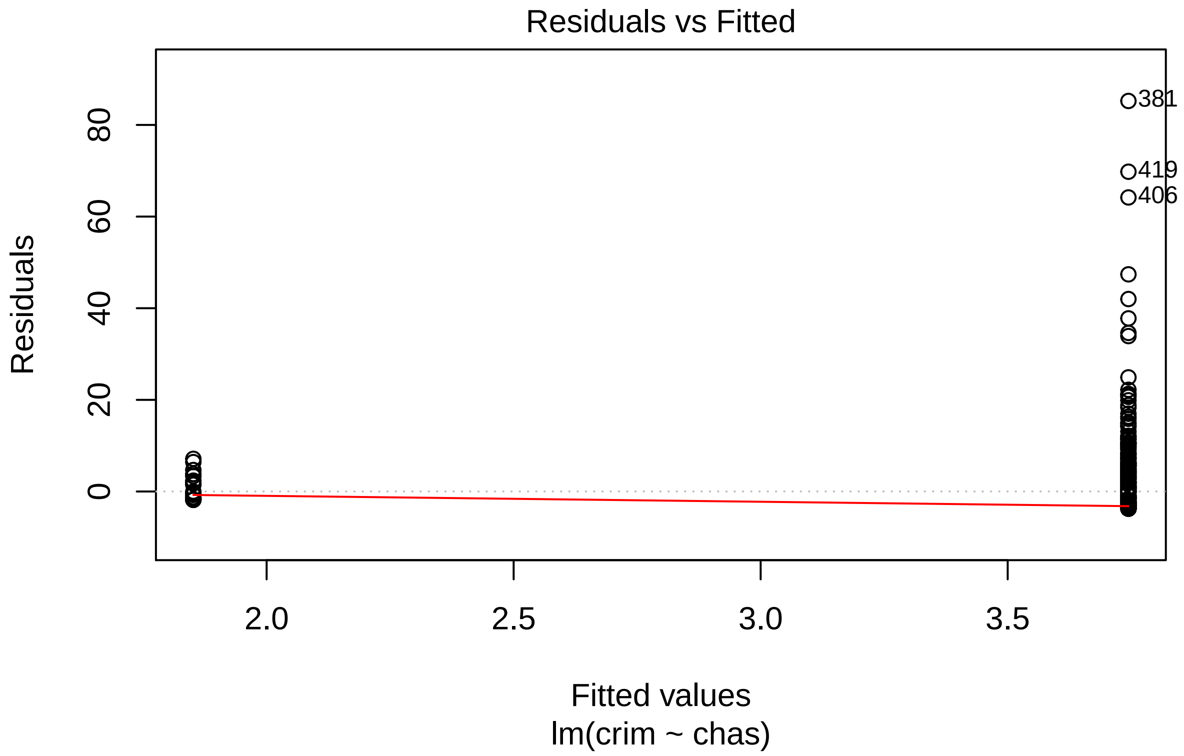
```
## Crime predicted by proportion of non-retail business per acre
```

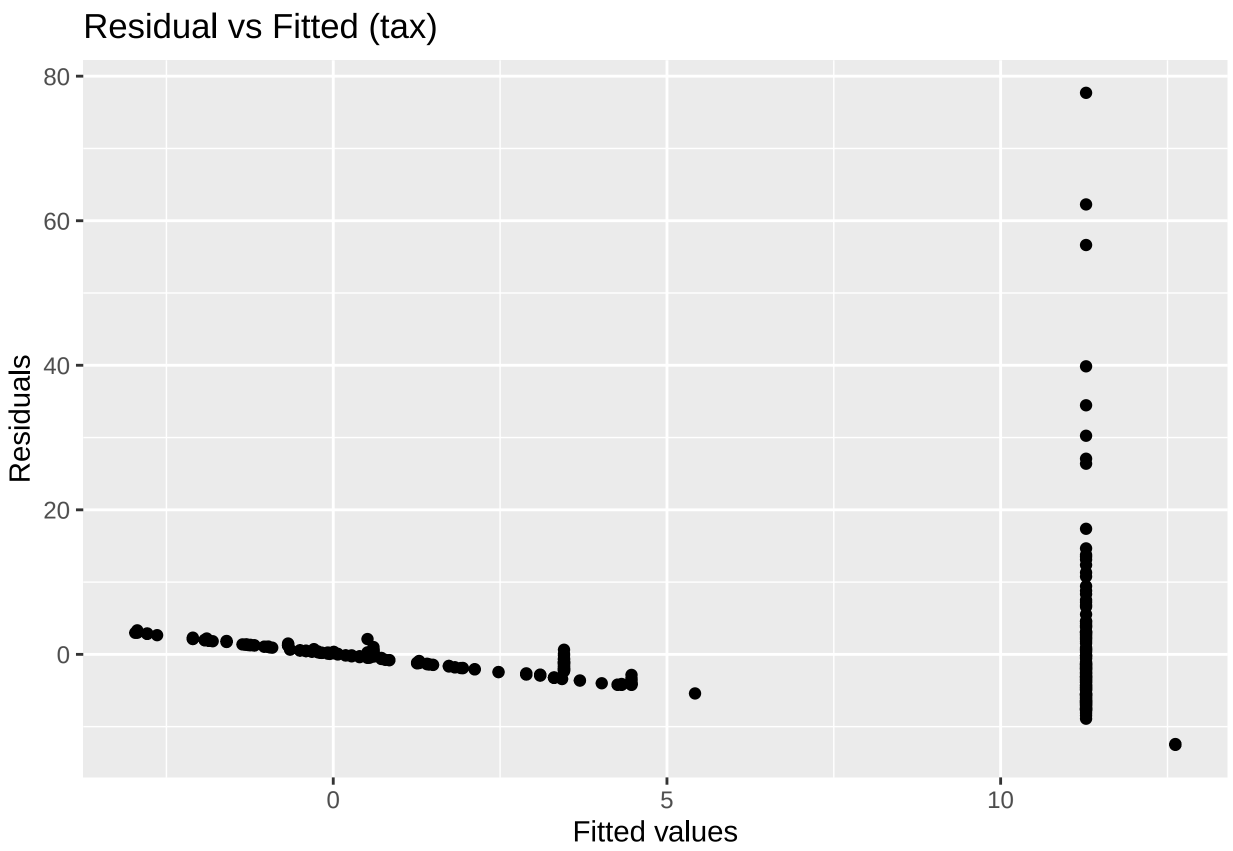
```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -2.0637426 0.66722830 -3.093008 2.091266e-03
## indus        0.5097763 0.05102433  9.990848 1.450349e-21
```

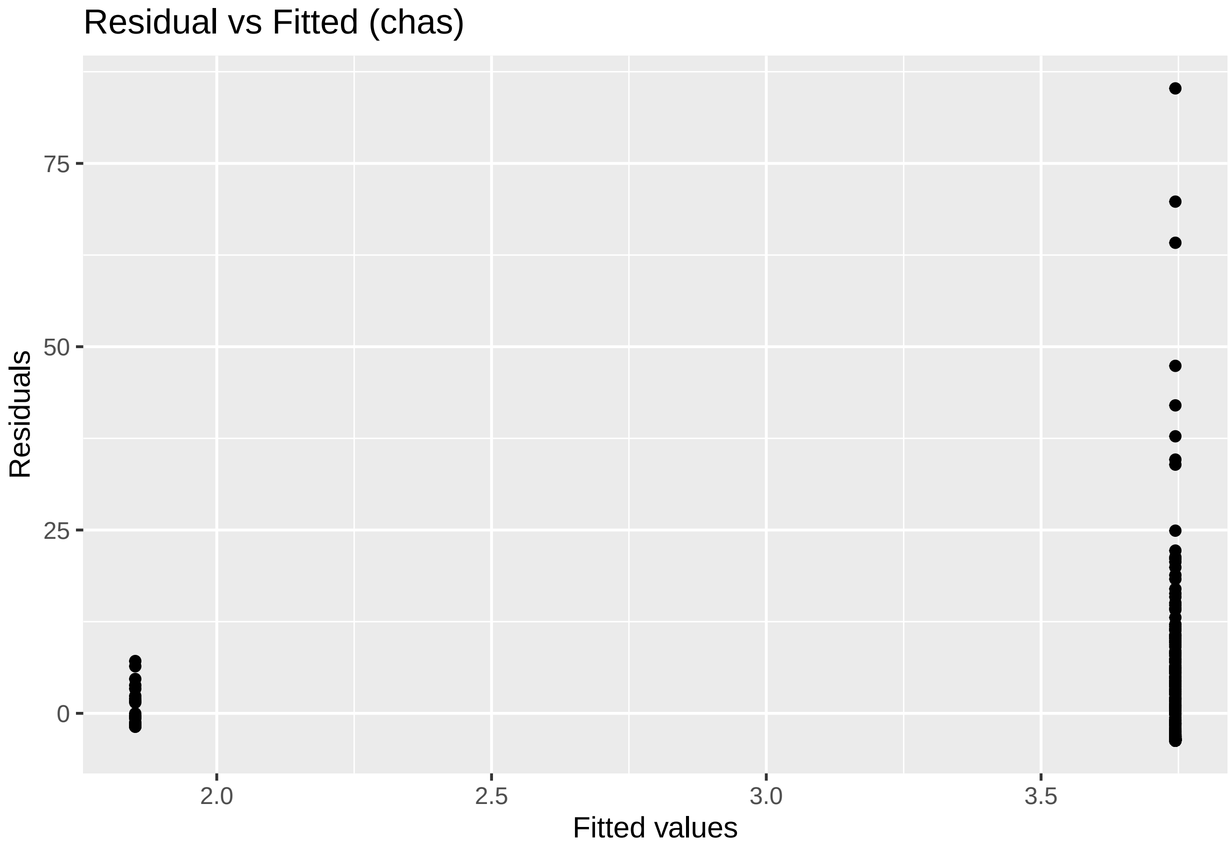
```
## Crime predicted by Charles river dummy variable
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  3.744447  0.3961111  9.453021 1.239505e-19
## chas        -1.892777  1.5061155 -1.256727 2.094345e-01

## Crime predicted by nitrogen oxides concentration

##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -13.71988   1.699479 -8.072992 5.076814e-15
## nox          31.24853   2.999190 10.418989 3.751739e-23

## Crime predicted by average number of rooms per dwelling

##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 20.481804  3.3644742  6.087669 2.272000e-09
## rm          -2.684051  0.5320411 -5.044819 6.346703e-07

## Crime predicted by proportion of owner-occupied unites built before 1940

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -3.7779063 0.94398472 -4.002084 7.221718e-05
## age          0.1077862 0.01273644  8.462825 2.854869e-16

## Crime predicted by weighted mean of distance to Boston employment centers

##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.499262  0.7303972 13.005611 1.502748e-33
## dis         -1.550902  0.1683300 -9.213458 8.519949e-19

## Crime predicted by index of accessibility to radial highways

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -2.2871594 0.44347583 -5.157349 3.605846e-07
## rad          0.6179109 0.03433182 17.998199 2.693844e-56

## Crime predicted by full-value property-tax rate

##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -8.52836909 0.815809392 -10.45387 2.773600e-23
## tax          0.02974225 0.001847415  16.09939 2.357127e-47

## Crime predicted by pupil-teacher ratio

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -17.646933  3.1472718 -5.607057 3.395255e-08
## ptratio       1.151983  0.1693736  6.801430 2.942922e-11

## Crime predicted by proportion of blacks by town

##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 16.55352922 1.425902755 11.609157 8.922239e-28
## black       -0.03627964 0.003873154 -9.366951 2.487274e-19

## Crime predicted by lower status of the population

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -3.3305381 0.69375829 -4.800718 2.087022e-06
## lstat        0.5488048 0.04776097 11.490654 2.654277e-27

## Crime predicted by median value of owner-occupied homes

##                Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 11.7965358 0.93418916 12.62757 5.934119e-32
## medv        -0.3631599 0.03839017 -9.45971 1.173987e-19
```


Residuals vs Fitted
Residuals
80
60
40
20
0
−20
0
5
10
Fitted values
lm(crim ~ tax)
381
419
406


Residuals vs Fitted
Residuals
80
60
40
20
0
381
419
406
2.0
2.5
3.0
3.5
Fitted values
lm(crim ~ chas)


Residual vs Fitted (tax)
Residuals
80
60
40
20
0
0
5
10
Fitted values



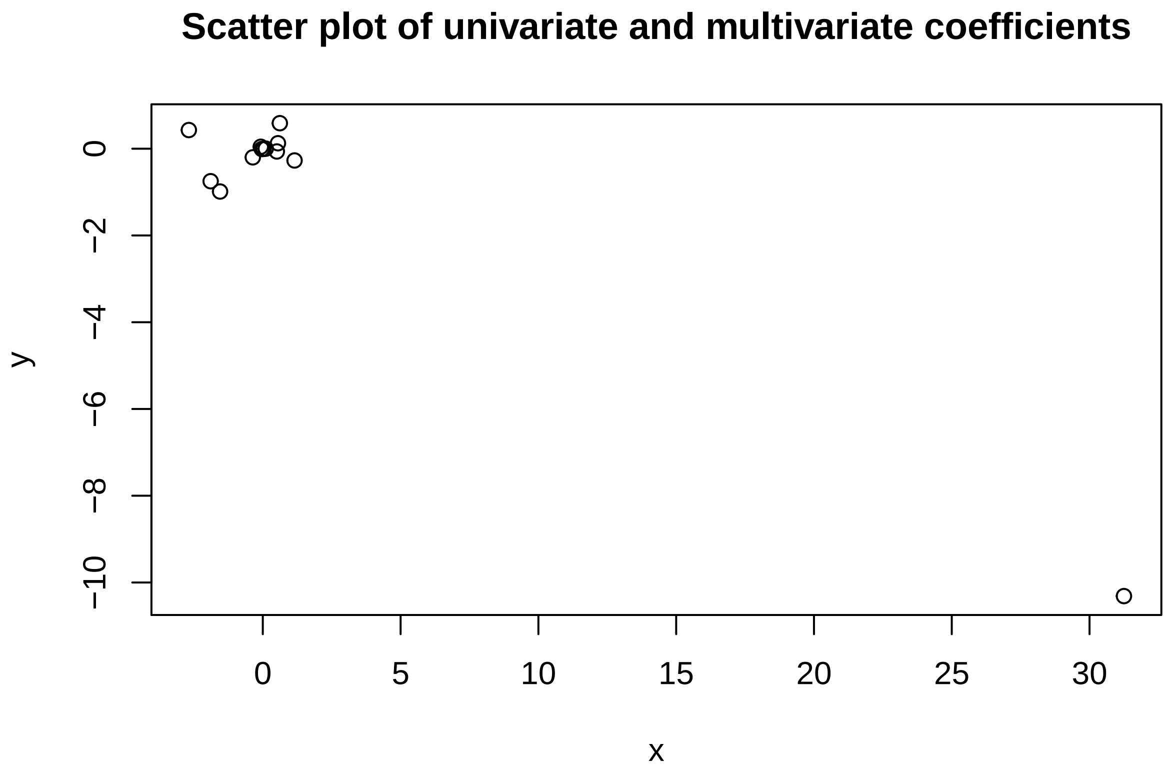
From looking at the summary of all the models, the Charles River dummy variable is the only one that doesn't show a strong relationship to crime.

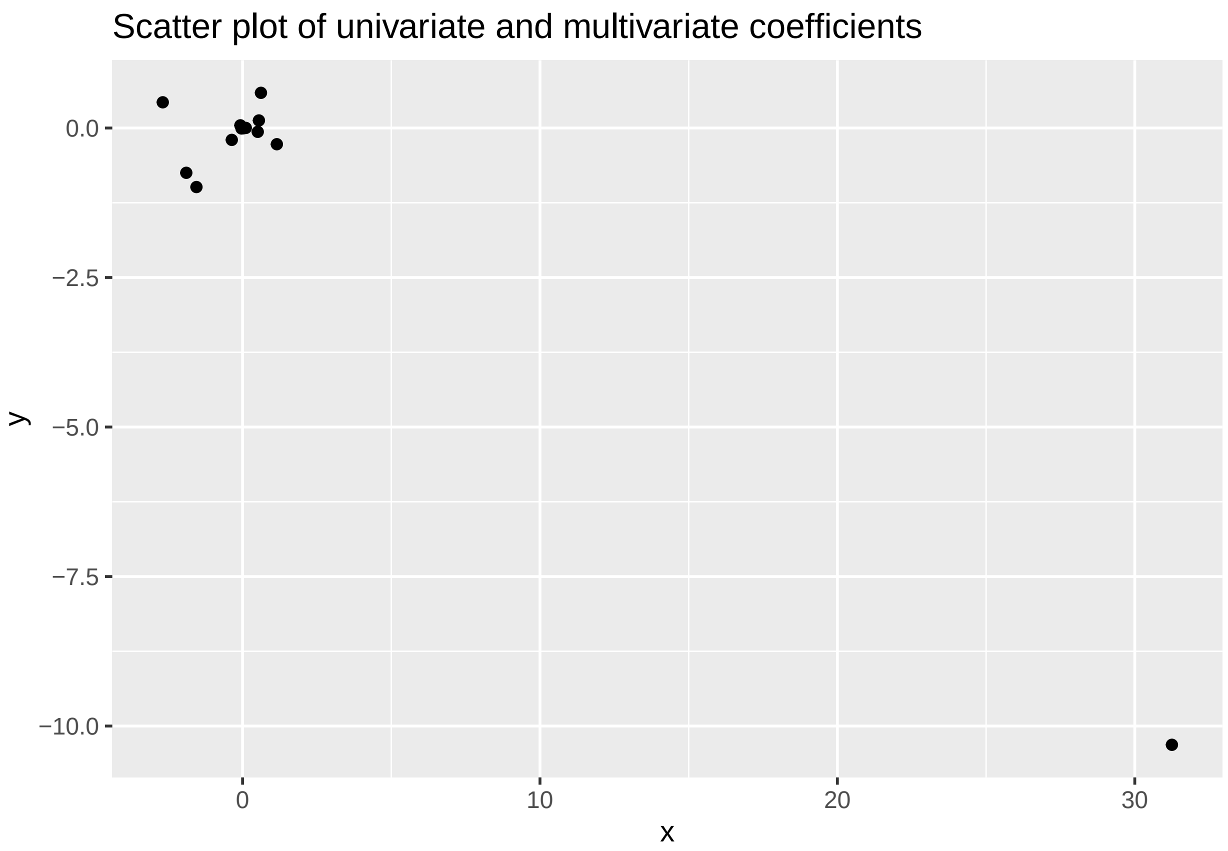
b) Fit a multiple regression model to predict the response using all the predictors. Describe your results. For which predictors can we reject the null hypothesis that the coefficient is equal to 0?

```
##
## 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
```

The results for the linear model using all of the variables to predict the per capita crime rate shows the predictors have different impacts than when they were used individually. When looking at each of the variables as predictors for the crime rate per capita, only the *chas* variable wasn't significantly related to crime, but now combining all of the variables, *indus*, *rm*, *age*, *tax*, and *ptratio* are also not contributing significantly to the model (coefficients are not significant and therefore we cannot say that their coefficients are different from zero). We can say that *zn*, *nox*, *dis*, *rad*, *black*, *lstat*, and *medv* are significant and reject the hypthesis that their coefficients in THIS model are different from zero. This is much like what Cami talked about in her P-value video where although they are not significant in this particular model, does not mean that given a different combination of variables as predictors they wouldn't be significant.

c) How do your results from part A compare to your results form part B? Create a plot displaying the univariate regression coefficients from part A on the x-axis, and the multiple regression coefficients from part B on the y-axis. That is, each predictor 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.





There are some differences between the univariate and multiple regression coefficients, most notably for *nox* which had a positive coefficient in the model where it was the only predictor, but in the model where its combined with other variables to predict the crime rate it become a negative coefficient. The variable *rm* also deviates more from the x=y line with it having been negative as a univariate and positive (0.4) in the multiple regression model. Most of the other variables fall close to a x=y line.

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 y = B0 + B1 * X + B2 * Xˆ2 + B3 * Xˆ3 + e.

```
## Crime predicted by residential land zoned for lots over 25k sq.ft
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     3.613524   0.372190  9.708814 1.547150e-20
## poly(zn, 3)1 -38.749835   8.372207 -4.628389 4.697806e-06
## poly(zn, 3)2  23.939832   8.372207  2.859441 4.420507e-03
## poly(zn, 3)3 -10.071868   8.372207 -1.203012 2.295386e-01
```

```
## Crime predicted by proportion of non-retail business per acre
```

```
##                    Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        3.613524   0.329998 10.950138 3.606468e-25
## poly(indus, 3)1  78.590819   7.423121 10.587301 8.854243e-24
## poly(indus, 3)2 -24.394796   7.423121 -3.286326 1.086057e-03
## poly(indus, 3)3 -54.129763   7.423121 -7.292049 1.196405e-12
```

```
## Crime predicted by nitrogen oxides concentration
```

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)      3.613524   0.321573 11.237025 2.742908e-26
## poly(nox, 3)1  81.372015   7.233605 11.249165 2.457491e-26
## poly(nox, 3)2 -28.828594   7.233605 -3.985370 7.736755e-05
## poly(nox, 3)3 -60.361894   7.233605 -8.344649 6.961110e-16
```

```
## Crime predicted by average number of rooms per dwelling
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)    3.613524  0.3702993  9.7583873 1.026665e-20
## poly(rm, 3)1 -42.379442  8.3296758 -5.0877661 5.128048e-07
## poly(rm, 3)2  26.576770  8.3296758  3.1906128 1.508545e-03
## poly(rm, 3)3  -5.510342  8.3296758 -0.6615314 5.085751e-01
```

```
## Crime predicted by proportion of owner-occupied unites built before 1940
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    3.613524  0.3485173 10.368276 5.918933e-23
## poly(age, 3)1 68.182009  7.8397027  8.697015 4.878803e-17
## poly(age, 3)2 37.484470  7.8397027  4.781364 2.291156e-06
## poly(age, 3)3 21.353207  7.8397027  2.723727 6.679915e-03
```

```
## Crime predicted by weighted mean of distance to Boston employment centers
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)     3.613524   0.325924  11.087013 1.060226e-25
## poly(dis, 3)1 -73.388590   7.331479 -10.010066 1.253249e-21
## poly(dis, 3)2  56.373036   7.331479   7.689176 7.869767e-14
## poly(dis, 3)3 -42.621877   7.331479  -5.813544 1.088832e-08
```

```
## Crime predicted by index of accessibility to radial highways
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     3.613524   0.297069 12.163920 5.149845e-30
## poly(rad, 3)1 120.907446   6.682402 18.093412 1.053211e-56
## poly(rad, 3)2  17.492299   6.682402  2.617666 9.120558e-03
## poly(rad, 3)3   4.698457   6.682402  0.703109 4.823138e-01
```

```
## Crime predicted by full-value property-tax rate
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     3.613524  0.3046845 11.859888 8.955923e-29
## poly(tax, 3)1 112.645827  6.8537074 16.435751 6.976314e-49
## poly(tax, 3)2  32.087251  6.8537074  4.681736 3.665348e-06
## poly(tax, 3)3  -7.996811  6.8537074 -1.166786 2.438507e-01
```

```
## Crime predicted by pupil-teacher ratio
```

```
##                     Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)         3.613524  0.3610484 10.008419 1.270767e-21
## poly(ptratio, 3)1  56.045229  8.1215830  6.900777 1.565484e-11
## poly(ptratio, 3)2  24.774824  8.1215830  3.050492 2.405468e-03
## poly(ptratio, 3)3 -22.279737  8.1215830 -2.743275 6.300514e-03
```

```
## Crime predicted by proportion of blacks by town
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)       3.613524   0.353627 10.2184605 2.139710e-22
## poly(black, 3)1 -74.431199   7.954643 -9.3569505 2.730082e-19
## poly(black, 3)2   5.926419   7.954643  0.7450264 4.566044e-01
## poly(black, 3)3  -4.834565   7.954643 -0.6077665 5.436172e-01
```

```
## Crime predicted by lower status of the population
##                    Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)       3.613524  0.3391698 10.654025 4.939398e-24
## poly(lstat, 3)1  88.069666  7.6294361 11.543404 1.678072e-27
## poly(lstat, 3)2  15.888164  7.6294361  2.082482 3.780418e-02
## poly(lstat, 3)3 -11.574022  7.6294361 -1.517022 1.298906e-01
```

```
## Crime predicted by median value of owner-occupied homes
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)      3.613524  0.2920344  12.373622 7.024110e-31
## poly(medv, 3)1 -75.057605  6.5691520 -11.425768 4.930818e-27
## poly(medv, 3)2  88.086211  6.5691520  13.409069 2.928577e-35
## poly(medv, 3)3 -48.033435  6.5691520  -7.311969 1.046510e-12
```

The variable *chas* doesn't allow for the *poly(chas,3)* function. There is evidence for non-linear relationships between certain variables and the crime rate per capita in this data. The variables with poly 3 coeffiences that are significant and thus different from zero are *indus*, *nox*, *age*, *dis*, *ptratio*, and *medv*. Beyond that, there are variables that show poly 2 coeffients that are significant and therefore different from 0 which are *zn*, *rm*, *rad*, *tax*, and *lstat*. The only variable that didn't have a coefficent at poly 2 or 3 that didn't reject the null hypothesis was *black*, although *chas* wasn't able to be run with poly so there is no evidence for this variable that it has a non-linear relationship with crime.

References:

- STHDA.com
- R-bloggers.com