Homework 2

Rylie Fleckenstein

1/28/2021

Exercises (ISLR)

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

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

   where

   $$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_{i'}^2}$$

   and we have been asked to show that

   $$\hat{y_i} = \sum_{i'=1}^{n} \alpha_{i'} y_{i'}$$

   also, what is $\alpha_{i'}$?

   So, we know that $\hat{y} = \hat{\beta} + \epsilon_i$ for $i = 1, 2, ..., n$ where ϵ_i is the ith error with mean 0 and variance σ^2.

   By following the OLS (ordinary least squares) principle, we know that in order to approximate the best parameters we must minimize the squared error rate of the function. We represent the squared error of this function with the following formula:

   $$\Delta = \sum_{i=1}^{n} (x_i - y_i\beta)^2$$

   $$\hat{y_i} = x_i\hat{\beta} = x_i * \frac{\sum_{i=1}^{n} x_{i'} y_{i'}}{\sum_{i=1}^{n} x_{i'}^2}$$

   therefore,

   $$\hat{y_i} = \frac{\sum_{i'=1}^{n} x_i x_i'}{\sum_{i=1}^{n} x_{i'}^2} * y_{i'}$$

   meaning

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

   where

   $$\alpha_{i'} = \frac{\sum_{i'=1}^{n} x_i x_i'}{\sum_{i=1}^{n} x_{i'}^2}; i' = 1, 2, ..., n$$

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.

```
## Linear Model Successfully Fit
```

(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
##
## 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
```

Based upon the summary of the linear model we can provide an interpretation of the coefficients of the model. We can see that for one unit increase in the Price of the car seat, the unit Sales in child car seats decreases by -0.054459 (in thousands). So, for $1 increase in the car seat price, you can expect to see, on average, a decrease of about 54.44 units sold when all other variables are constant. We can also see that when the store selling the car seat is located in an Urban area, you can expect to see a decrease of about -0.21916 (in thousands) of unit sales in relative to a rural store location. This parameter, however, does not have strong evidence supporting an association between it and the response variable considering the p-value is almost 1. Lastly, we can see that when the store selling the car seat is located in the US, you can expect to see an increase in the number of car seat units sold by about 1.2 (in thousands) relative to stores found outside of the US.

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

$$\hat{Sales} = 13.043469 - 0.054459Price - 0.021916Urban[Yes] + 1.200573US[Yes]$$

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

The predictors "Price" and "USYes" both have p-values with highest level of significance, meaning we can reject the null hypothesis for them. The predictor "UrbanYes" does not have a statistically significant p-value therefore there is not strong enough evidence for us to reject the null hypothesis for that predictor in this model. It is important to keep in mind, however, that there are no interaction terms in this model, meaning the entire story may not be represented and a different model may present different evidence leading to a different conclusion.

(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 first model in (a) has an adjusted R-squared value of 0.2335 while the model in past (e) has an adjusted R-squared value of 0.2354. I would say neither of these models fits the data in an overwhelming well fashion considering those R-squared values are showing that the model only accounts for a low portion of the variance, R-squared values closer to 1 would show the model fitting the data well. With that, the second model in part (e) does fit the data better then the first model in part (a).

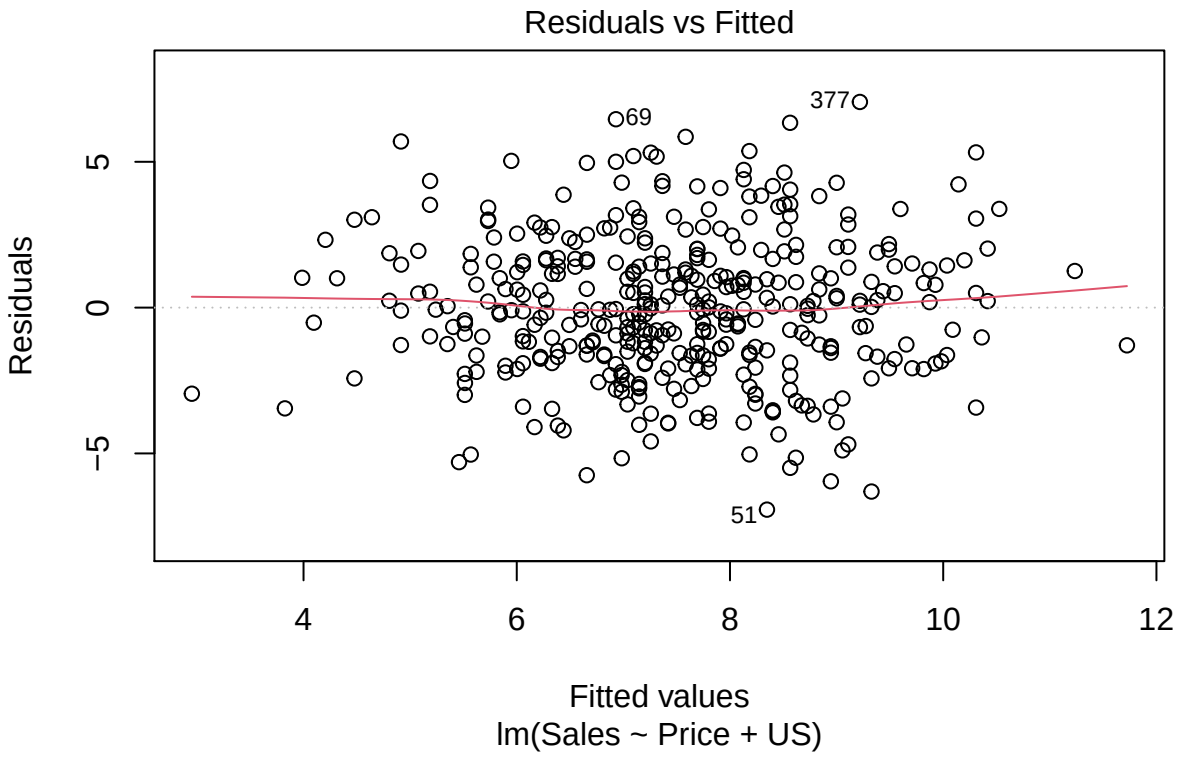
(g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).

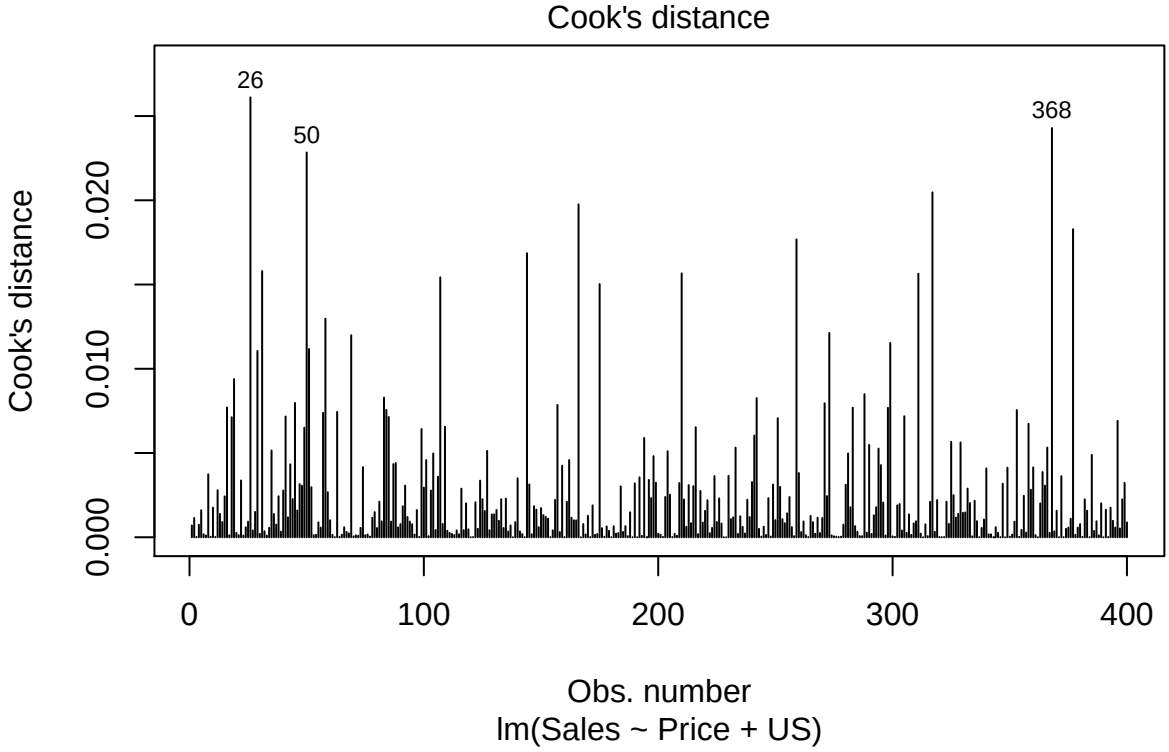
Table 1: Confidence Intervals 95%

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

The 95% confidence intervals for the parameters in the second model are printed above. As we can see, the confidence interval for the variable Price ranges from (-0.0647, -0.044) and the confidence interval for the parameters "US[Yes]" ranges from (0.69,1.707).

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


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



If we take a look at the graph of the studentized residuals against the fitted values we see that all observations fall within the (-3,3) range suggesting there are no outliers in this data set. However, if we take a look at the plot of the residuals vs the fitted values and the standardized residuals against the theoretical quantiles, there is evidence to suggest the three points (51, 69, 377) are outliers in this data set. Next, if we take a look at the standardized residuals plotted against the leverage, we see there is evidence to suggest the three points (25, 50, 368) are high leverage points. Taking a look at the Cook's Distance graph confirms this evidence as well.

3. Question 3.7.15 pg 126

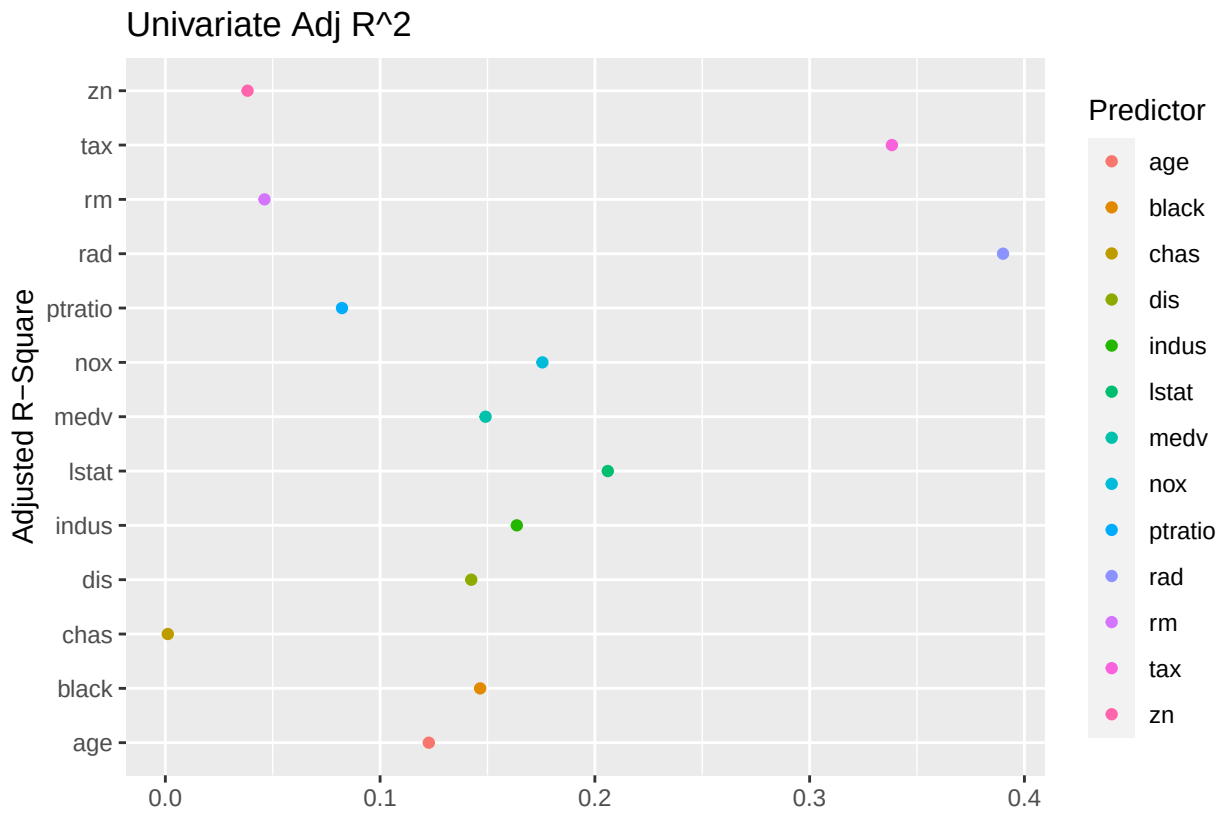
This problem involves the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

(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.

Predictor	AdjR2	PValue	Slope
zn	0.0382835	0.0000055	-0.0739350
indus	0.1636539	0.0000000	0.5097763
chas	0.0011459	0.2094345	-1.8927766
nox	0.1755847	0.0000000	31.2485312
rm	0.0461804	0.0000006	-2.6840512

Predictor	AdjR2	PValue	Slope
age	0.1226842	0.0000000	0.1077862
dis	0.1424513	0.0000000	-1.5509017
rad	0.3900489	0.0000000	0.6179109
tax	0.3383040	0.0000000	0.0297423
ptratio	0.0822511	0.0000000	1.1519828
black	0.1465843	0.0000000	-0.0362796
lstat	0.2060187	0.0000000	0.5488048
medv	0.1490955	0.0000000	-0.3631599

I fit a simple linear regression model for each predictor in order to predict the response "Crime" and found that 12 out of the 13 models suggested their variable is significant in predicting the response and therefore are able to reject the null hypothesis that Ho=0. The only predictor to not be significant was the predictor "chas". As far as how well each model fits the data, I would not consider any of these models to be the best for making predictions. The highest adjusted R-square value is only 0.39 which is pretty low suggesting we can do better. Below I have plotted the adjusted R-square values for all of the univariate models.



(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: Bj=0?

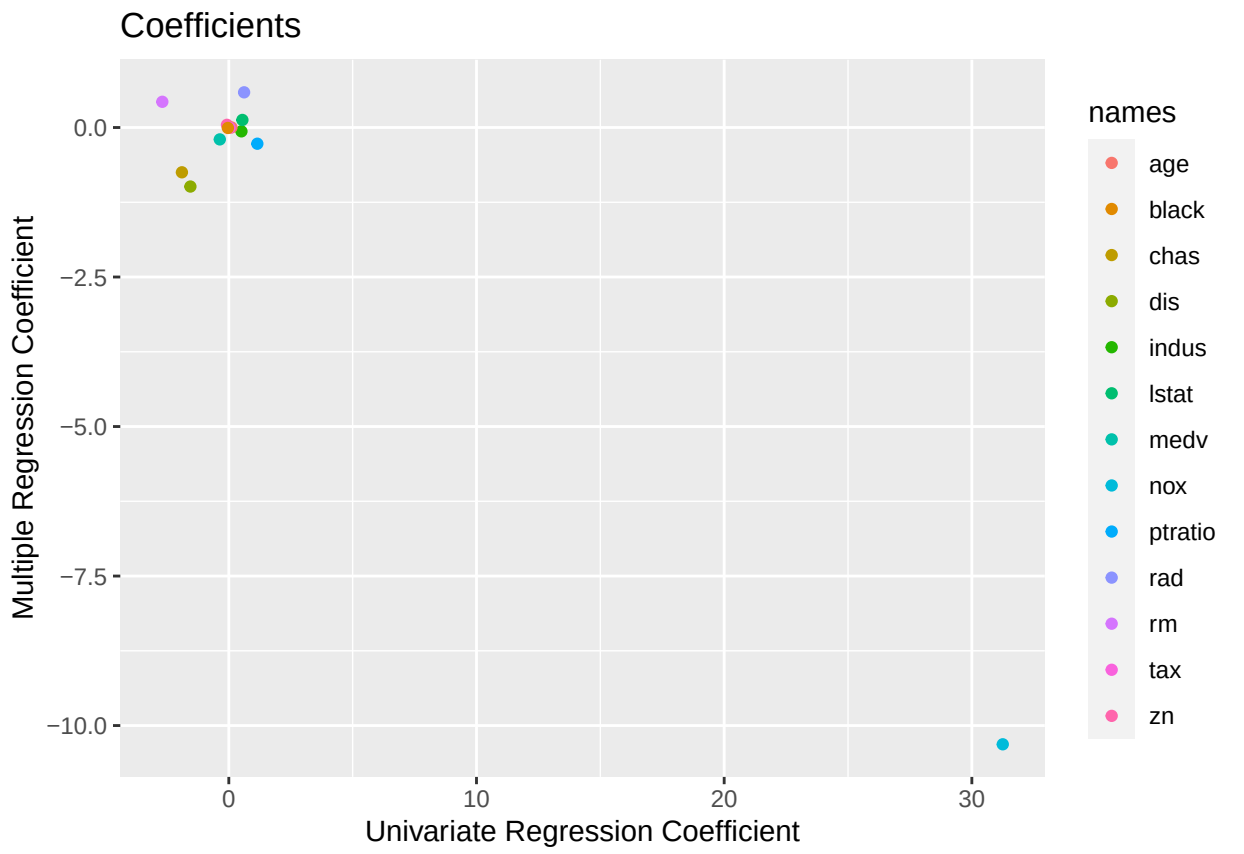
```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
```

```
## 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
```

In the above summary we can see how well the model fits the data when all of the variables are used together to predict the response variable. We can see that the model has greatly improved its fit since the adjusted R-square is now 0.4396 which is better then any previous model but could still be better. Also, we now see that there are only 8 predictors which have a p-value significant enough to reject the null hypothesis ("zn", "nox", "dis", "rad", "black", "lstat", and "medv") and most of them have p-values which suggest they have a lower level of significance in predicting the response variable then they did in their own univariate models.

(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 from (a) compare to (b) some what similarly. We can see all but one predictor is clumped pretty closely together. The multiple linear regression model, however, does perform much better than the the univariate models in predicting 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 = B0 + B1X + B2X2 + B3X3 + error.

After fitting models for each predictor that have each polynomial transformation up to the 3rd power, it appears there are a few predictors that have a non-linear association between themselves and the response. By looking at the p-values of each predictor transformation we can see that the predictors (inus, nox, rm, age, dis, tax, ptratio, lstat) all have non-linear predictors with significant p-values.

Predictor	AdjR2	PredPV	Pred2PV	Pred3PV
zn	0.0497002	0.2437057	0.5318217	0.6516291
indus	0.3024016	0.0003683	0.0000064	0.0000001
chas	0.0011459	0.2094345	NA	NA
nox	0.3189756	0.0000108	0.0000074	0.0000056
rm	0.0839644	0.0007210	0.0008426	0.0010664
age	0.1784578	0.2875856	0.1962969	0.1346755
dis	0.2987607	0.0000007	0.0001732	0.0027019
rad	0.3942002	0.7649367	0.7461524	0.7474796
tax	0.3937992	0.2541226	0.1936155	0.1344614
ptratio	0.2122608	0.0056310	0.0034248	0.0019886
black	0.1439470	0.7809206	0.7781976	0.6243812
lstat	0.2242856	0.3760692	0.2035285	0.0865994
medv	0.4281693	0.0055425	0.2063270	0.6550890

Sources:

Alto, V. (2019, August 17). Understanding the OLS method for Simple Linear Regression. Retrieved February 01, 2021, from https://towardsdatascience.com/understanding-the-ols-method-for-simple-linear-regression-e0a4e8f692cc

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York :Springer,

http://people.hsc.edu/faculty-staff/robbk/math121/lectures/Spring%202012/Lecture%2043%20-%20Residual%20Analysis.pdf