Team 5
Branden Ho
Anna Nguyen
2/27/2022

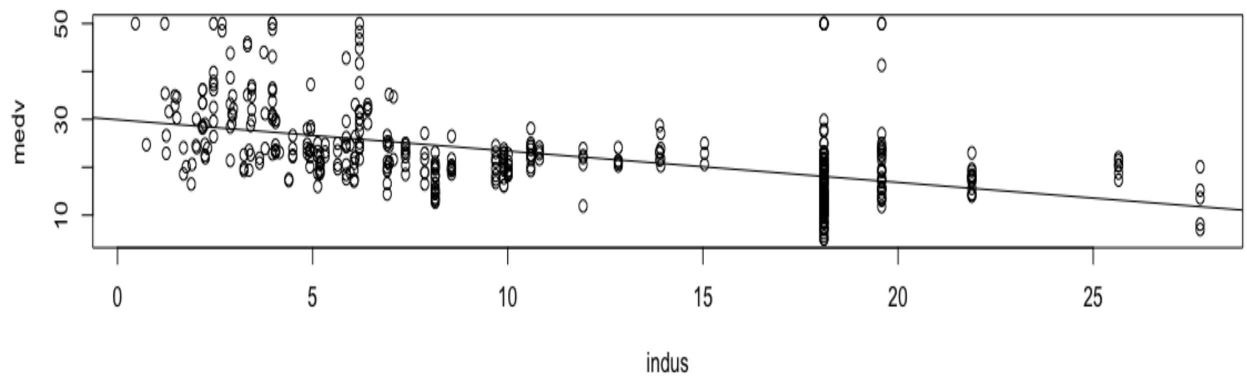
Case 2 Report

#(i) Start with exploratory data analysis. Are there outliers? (use boxplot)
Explanatory analysis

boxplot(boston_train$crim)Many outliers
boxplot(boston_train$zn) SOme outliers
boxplot(boston_train$indus) No outliers/
boxplot(boston_train$chas) One outlier
boxplot(boston_train$nox)No outliers
boxplot(boston_train$rm)Many outliers
boxplot(boston_train$age)No outliers/
boxplot(boston_train$dis) few outliers
boxplot(boston_train$rad)No outliers
boxplot(boston_train$tax)No outliers
boxplot(boston_train$ptratio)No outliers
boxplot(boston_train$black) many outliers
boxplot(boston_train$lstat)Some outliers

No outliers	**Some Outliers**	**Many outliers**
indus	Zn	Crim
Nox	Chas	Rm
age	dis	black
rad	lstat	
tax		
ptratio		

#(ii) Conduct linear regression on the training data.



=> There is a negative correlation between medv and indus

```
Residuals:
    Min      1Q  Median      3Q     Max 
-15.365  -2.737  -0.518   1.598  25.576 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.964e+01  5.357e+00   7.399 6.98e-13 ***
crim        -1.221e-01  3.696e-02  -3.304  0.00103 ** 
zn           4.518e-02  1.446e-02   3.124  0.00190 ** 
indus       -1.396e-02  6.618e-02  -0.211  0.83304    
chas         3.036e+00  9.319e-01   3.257  0.00121 ** 
nox         -1.814e+01  4.059e+00  -4.469 1.00e-05 ***
rm           3.511e+00  4.481e-01   7.835 3.53e-14 ***
age         -2.625e-04  1.372e-02  -0.019  0.98474    
dis         -1.576e+00  2.129e-01  -7.402 6.87e-13 ***
rad          3.347e-01  7.043e-02   4.752 2.73e-06 ***
tax         -1.280e-02  3.966e-03  -3.227  0.00134 ** 
ptratio     -9.639e-01  1.410e-01  -6.834 2.76e-11 ***
black        9.187e-03  2.920e-03   3.146  0.00177 ** 
lstat       -5.304e-01  5.371e-02  -9.875  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.788 on 441 degrees of freedom
Multiple R-squared:  0.7289,    Adjusted R-squared:  0.7209 
F-statistic: 91.19 on 13 and 441 DF,  p-value: < 2.2e-16
```

=> this output shows us the p-value, t-value, standard error, and coefficient estimate for each independent variable, along with the adjusted R-squared, residual standard error, and F-statistic for the model with all independent variables
=> if we use F-test and t-test to conduct hypothesis tests, we will reject the null hypothesis if p-value < alpha

#(iii) Conduct variable selection. Find the best linear model. Show residual diagnosis

```
Step:  AIC=1435.04
medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
    black + lstat

          Df Sum of Sq   RSS    AIC
<none>                  10112 1435.0
- black    1    229.17 10341 1443.2
- zn       1    230.13 10342 1443.3
- chas     1    243.09 10355 1443.8
- crim     1    249.58 10362 1444.1
- tax      1    313.34 10425 1446.9
- nox      1    553.27 10665 1457.3
- rad      1    578.10 10690 1458.3
- ptratio  1   1112.19 11224 1480.5
- dis      1   1435.27 11547 1493.4
- rm       1   1484.78 11597 1495.4
- lstat    1   2503.61 12616 1533.7
```

=> This is the best model according to backward elimination

```
Step:  AIC=1435.04
medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn +
    rad + tax + crim

        Df Sum of Sq   RSS  AIC
<none>                10112 1435
+ indus  1   1.02071 10111 1437
+ age    1   0.00907 10112 1437
```
=> This is the best model according to forward selection

```
> summary(model_1)

Call:
lm(formula = medv ~ ., data = boston_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.365  -2.737  -0.518   1.598  25.576 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.964e+01  5.357e+00   7.399 6.98e-13 ***
crim        -1.221e-01  3.696e-02  -3.304  0.00103 ** 
zn           4.518e-02  1.446e-02   3.124  0.00190 ** 
indus       -1.396e-02  6.618e-02  -0.211  0.83304    
chas         3.036e+00  9.319e-01   3.257  0.00121 ** 
nox         -1.814e+01  4.059e+00  -4.469 1.00e-05 ***
rm           3.511e+00  4.481e-01   7.835 3.53e-14 ***
age         -2.625e-04  1.372e-02  -0.019  0.98474    
dis         -1.576e+00  2.129e-01  -7.402 6.87e-13 ***
rad          3.347e-01  7.043e-02   4.752 2.73e-06 ***
tax         -1.280e-02  3.966e-03  -3.227  0.00134 ** 
ptratio     -9.639e-01  1.410e-01  -6.834 2.76e-11 ***
black        9.187e-03  2.920e-03   3.146  0.00177 ** 
lstat       -5.304e-01  5.371e-02  -9.875  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.788 on 441 degrees of freedom
Multiple R-squared:  0.7289,	Adjusted R-squared:  0.7209 
F-statistic: 91.19 on 13 and 441 DF,  p-value: < 2.2e-16
```

=> This is the summary of model 1 that includes all independent variables

```
> summary(model_2)

Call:
lm(formula = medv ~ lstat + rm + ptratio + dis + nox + chas +
    black + rad + crim + zn + tax, data = boston_train)

Residuals:
    Min      1Q  Median      3Q     Max
-15.372  -2.727  -0.525   1.604  25.565

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.702975   5.327856   7.452 4.86e-13 ***
lstat        -0.531334   0.050734 -10.473  < 2e-16 ***
rm            3.521499   0.436629   8.065 6.90e-15 ***
ptratio      -0.968598   0.138762  -6.980 1.08e-11 ***
dis          -1.564946   0.197355  -7.930 1.81e-14 ***
nox         -18.400125   3.737385  -4.923 1.20e-06 ***
chas          3.013621   0.923459   3.263 0.001186 **
black         0.009206   0.002906   3.169 0.001638 **
rad           0.338928   0.067348   5.033 7.05e-07 ***
crim         -0.121878   0.036858  -3.307 0.001021 **
zn            0.045452   0.014315   3.175 0.001602 **
tax          -0.013169   0.003554  -3.705 0.000238 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.778 on 443 degrees of freedom
Multiple R-squared:  0.7288,    Adjusted R-squared:  0.7221
F-statistic: 108.2 on 11 and 443 DF,  p-value: < 2.2e-16
```

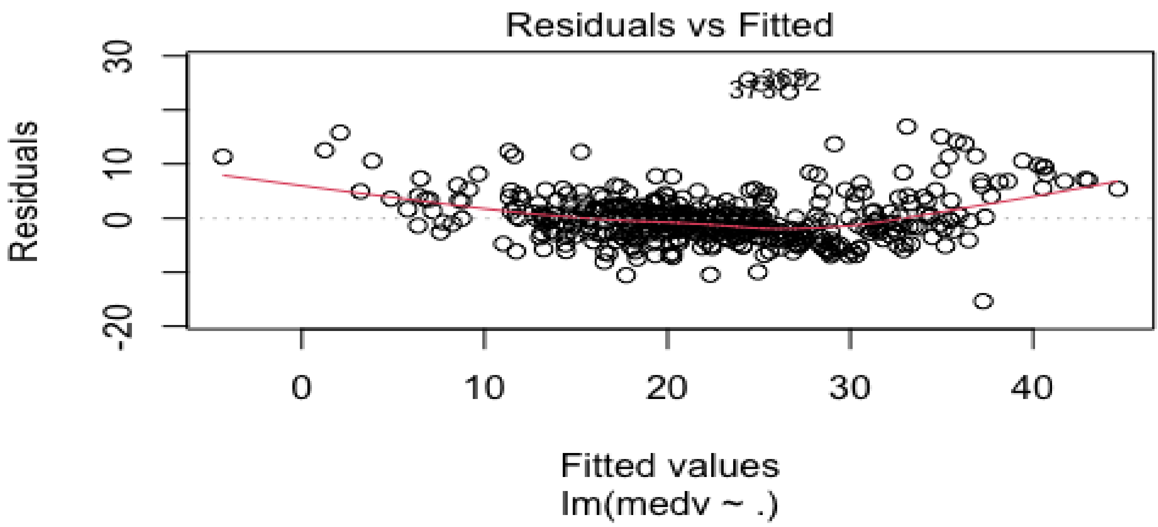
=> This is the summary of model 2 that includes selected variables from backward elimination

```
> AIC(model_1)
[1] 2732.224
> AIC(model_2)
[1] 2728.27
```

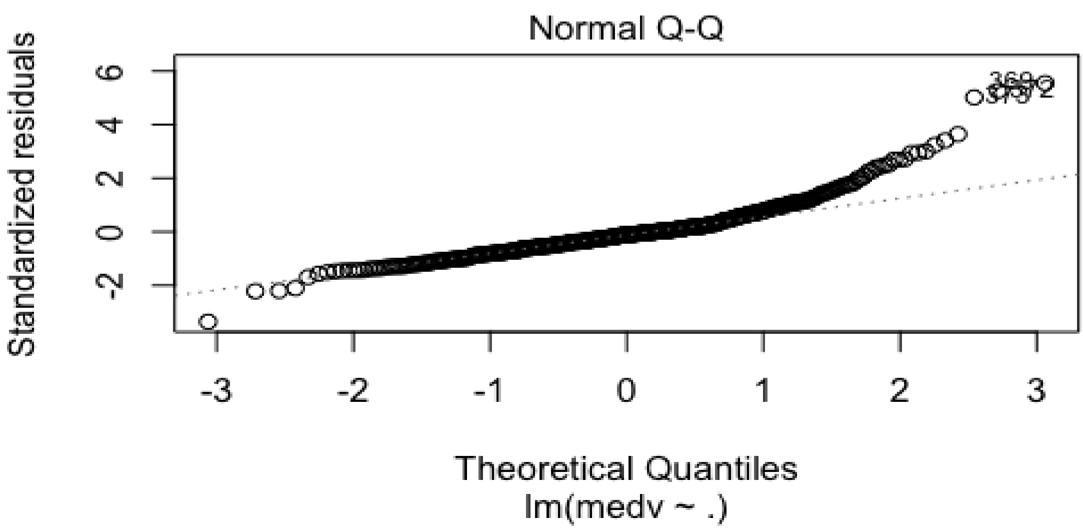
=> This is the AIC of model 1 and 2

=> Based on the output, I conclude that model 2 is better because it has smaller AIC and bigger adjusted R squared

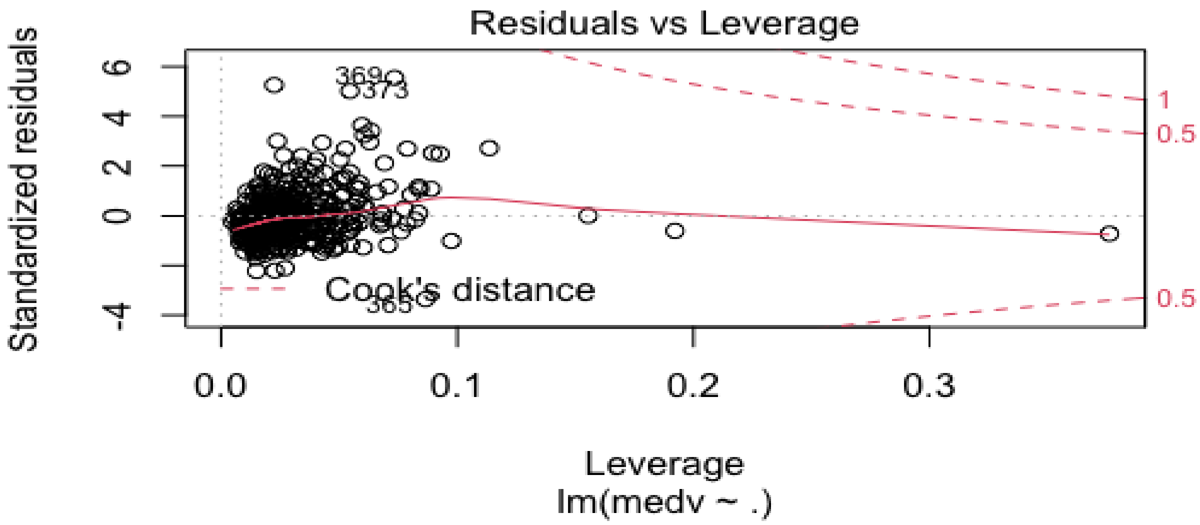
#Plots of model 1



=> This is a good graph because there is no pattern and most residuals scattered around 0 lines which indicates there is a linear relationship

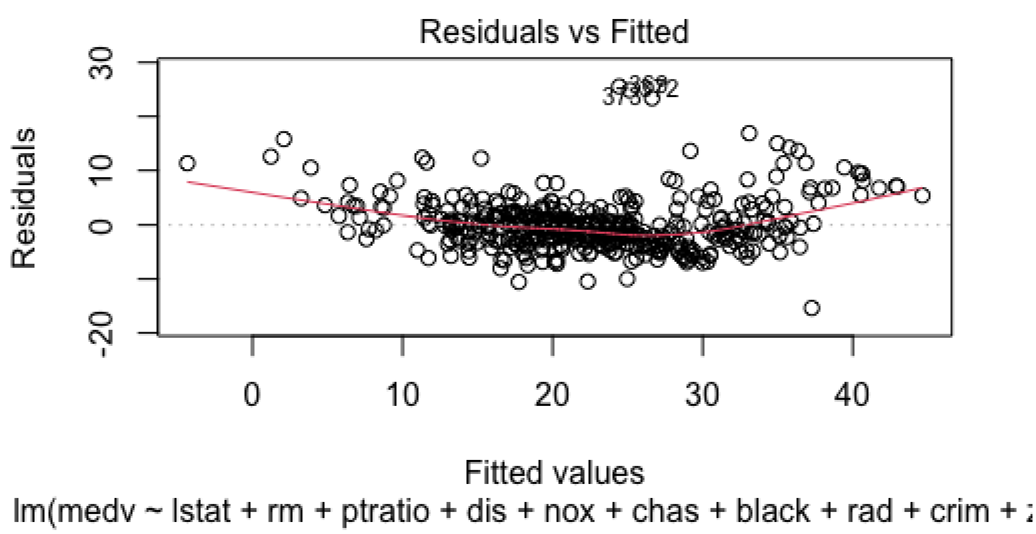


=> Dots fall on the dash line which means that this is a good plot as well

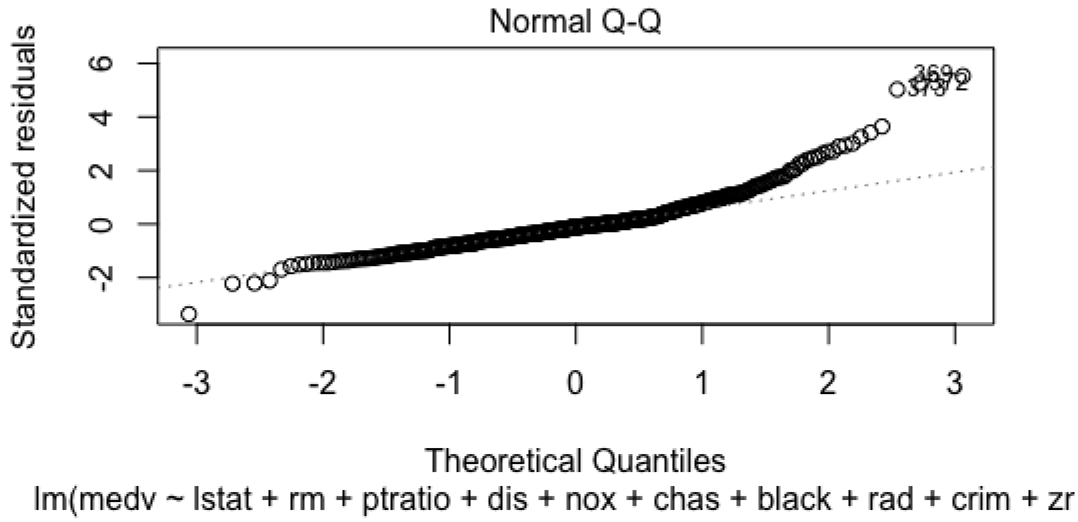


=> This is a good plot too because there is no observation of large Cook's distance, most residuals fall to the left where the distance value is small

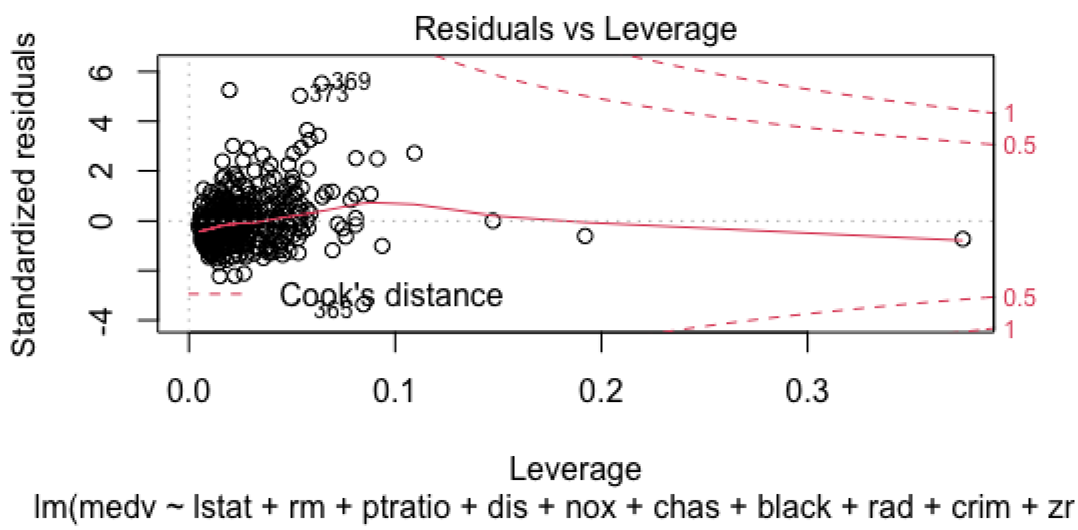
#Plots of model 2



=> This is a good plot because there is no pattern and most residuals scattered around 0 line



=> This is a good plot too because most dots fall on dashed line



=> This is a good plot too because there is no observation with large Cook's distance

#(iv) Test the out-of-sample performance. Using final linear model built from (iii) on the 90% of original data, test with the remaining 10% testing data. Report out-of-sample model MSE etc.

In Sample Evaluation:

	MSE	R squared	Adjusted R squared	AIC	BIC
Model 1	22.92733	0.7288603	0.7208676	2732.224	2794.029
Model 2	22.82614	0.7288327	0.7220995	2728.27	2781.834

As we see in this table, the MSE, R squared, AIC and BIC all come down a bit from the first model to the second model. These values being down shows that there is less error in the second model and that the second model is a better fit.

Out of Sample Using the Testing Data:

	MSE	MAE
Model 1	19.6293	3.526744
Model 2	19.51502	3.519002

Here we have the testing result of use using that leftover 10% to test how good our models were. With both MSE and MAE being down it proves that model 2 is a better fit model and will be able to predict future data points with less error than model 1.

V:**Cross validation on the original data. Use 10-fold cross validation. Does (v) yield a similar answer as (iv)?**

10-Fold Cross Validation:

	10-Fold	Leave-one-out	10-Fold using MAE
Model 2	22.90228	23.51161	3.384988

The 10 fold cross validation is created using our variables from the best fit model, but uses 100% of the data from the data set. The fact that the 10 fold result has a MSE of 22.9 and our original model 2 has an MSE of 22.8 shows that our model was actually very good. The 10-fold should produce a slightly higher MSE due to the fact that it is using the entire data set.

VI:

In sample Evaluation:

	MSE	R squared	Adjusted R squared	AIC	BIC
Model 1	22.68837	0.7402259	0.7325682	2727.457	2789.262
Model 2	22.60116	0.7400509	0.7335962	2723.764	2777.327

Out of Sample Using the Testing Data:

	MSE	MAE
Model 1	21.53126	3.453904
Model 2	21.35655	3.442893

Cross Validation:

	10-Fold	Leave-one-out	10-Fold using MAE
Model 2	23.60327	23.51161	3.3898997

As we can see from the table above, the second run shows pretty similar numbers and similar decreases from model 1 to model 2. One of the things to note is that our numbers are lower on average compared to the first run through and we can attribute that to the random selection of data points. Our conclusion from this data is that these models that were created through the stepwise regression process are more accurate and have a better fit. The meaning behind this is that the variable uses in our best fit model(model_2) are the most significant variables to predicting the other variables