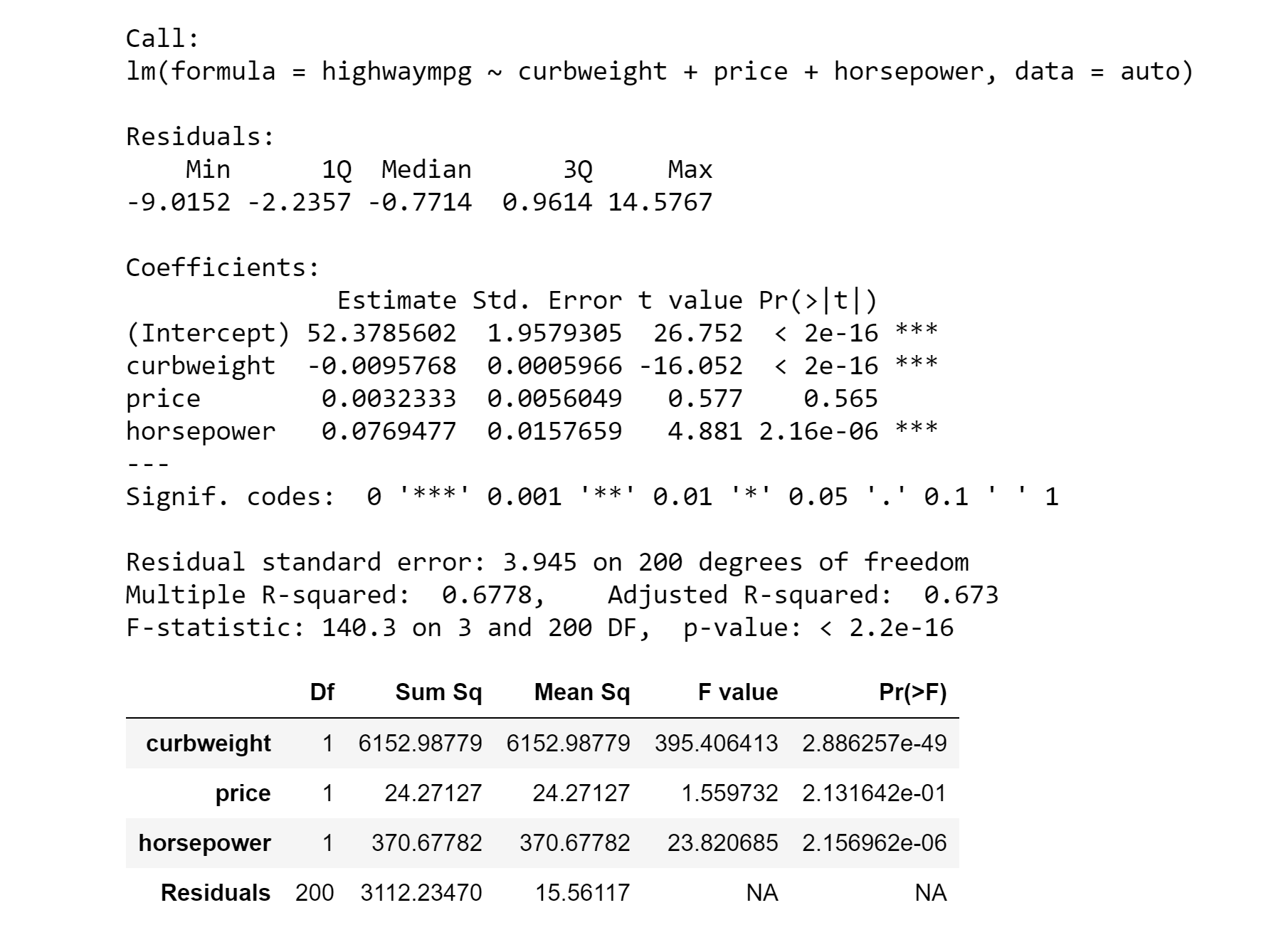
# **MULTIPLE LINEAR REGRESSION ANALYSIS**

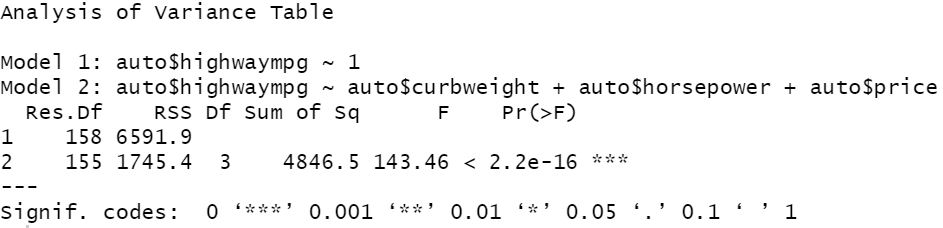
Automobile dataset (<https://archive.ics.uci.edu/ml/datasets/Automobile>).

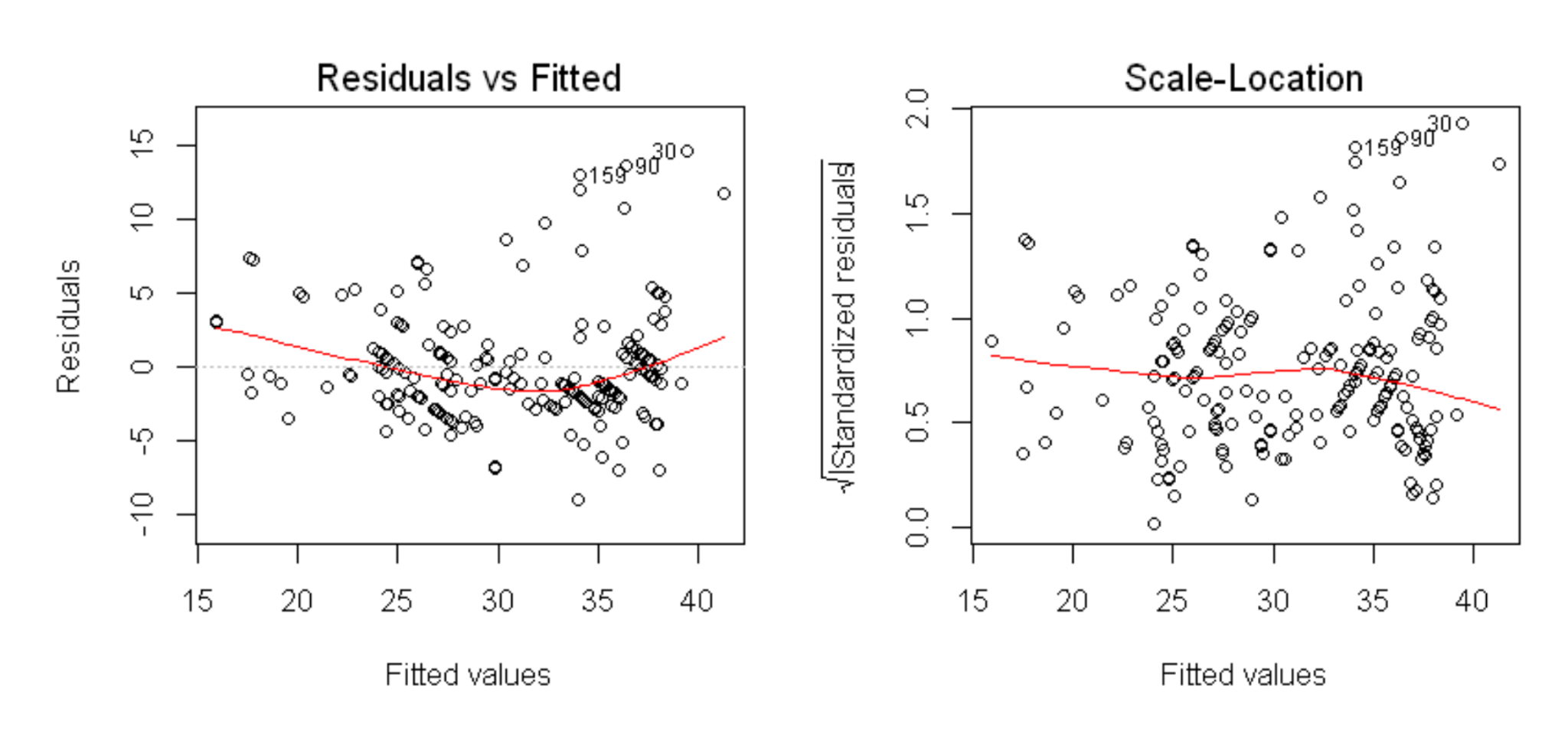
## **Model #1: Regression model including three numerical predictor variables**

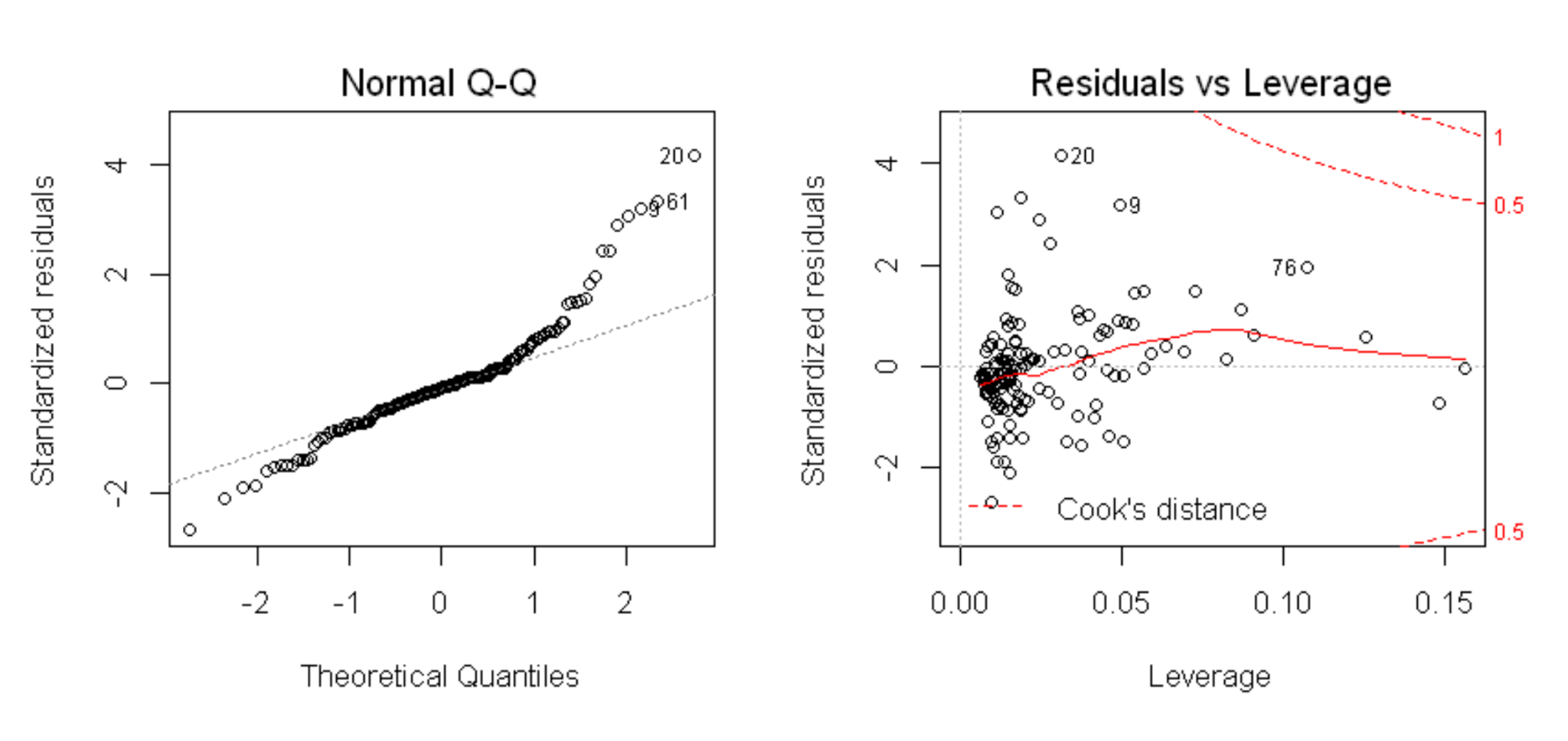
## **ANOVA table Report**

We conducted an analysis of variance with the response (y) variable being highway mileage and predictor (x) variables horsepower, curb-weight, and price.

****

****

****

****

*SSE*= 4846.5 (DF: 3)

*SSR*= 1745.4 (DF: 155)

*SST=SSR+SSE=* 6591.9 (DF: 158)

### **F test of overall linear relationship. Hypotheses is stated, the value of test statistics, p-value, and conclusion.**

Hₒ: =0

Hₐ: ≠ 0

F Statistic = 143.46

p-value = 2.2e -16

The null hypothesis is that the given predictor variables have no effect on the y variable highway mpg. Because the p-value is smaller than our chosen alpha level (0.05) we reject the null hypotheses in favor of the alternative hypothesis. Thus we can conclude that using the multiple regression model is better than using just the mean.

### **Testing whether is helpful, given that and are in the model. Hypotheses is stated, the value of test statistics, p-value, and conclusion.**

Hₒ:= ==0

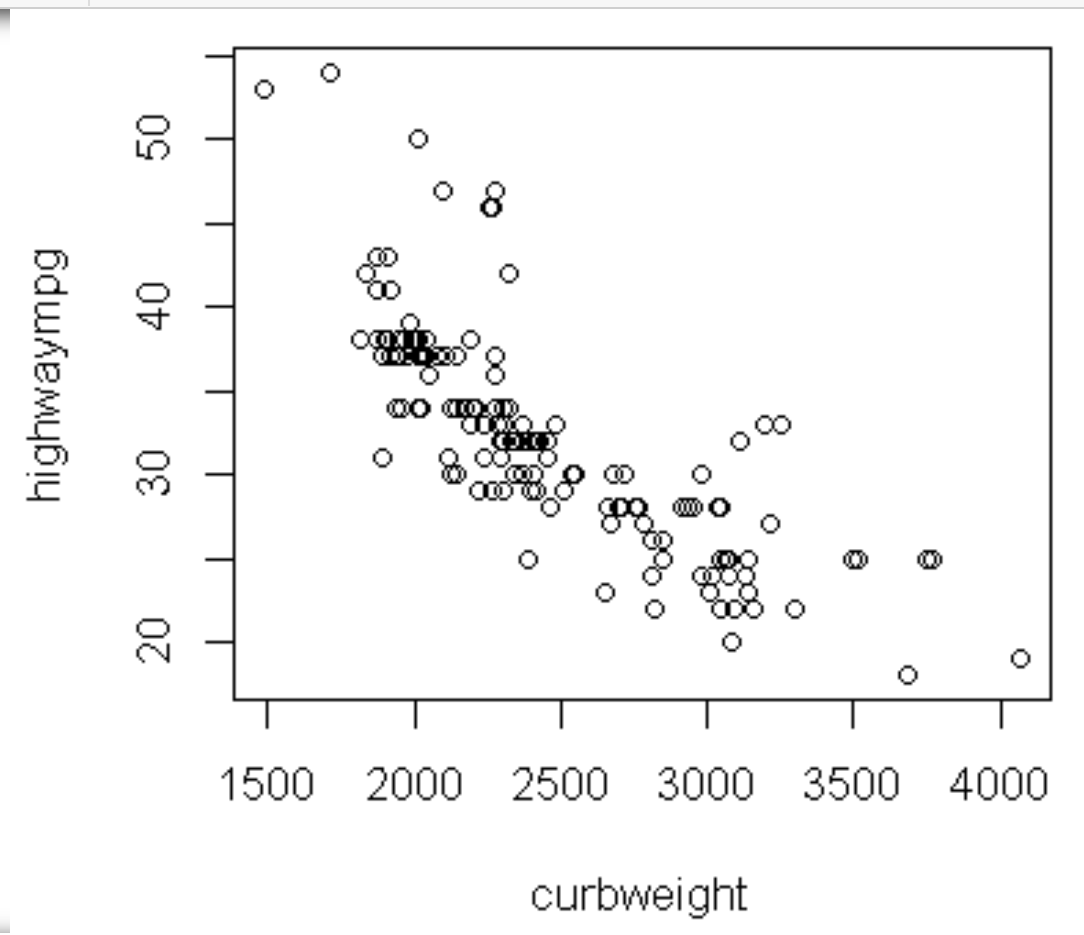
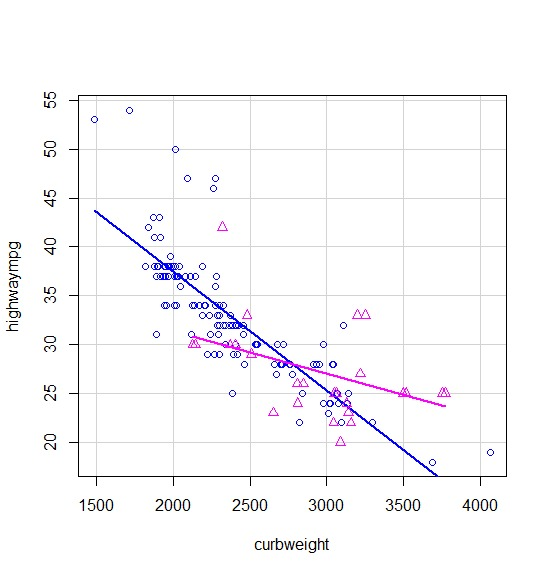
Hₐ: at least one ≠ 0

F Statistic = 1.5597

p-value = 0.2132

For this test we consider the p-value associated with the F statistic. If we fail to reject the null hypothesis this means that the additional predictor variable contributes nothing to the prediction of y after allowing for all other predictors. While we do not reject the null at =0.05 we may at =0.01 since the p value is larger than 0.05 for

## **Model #2: Numerical and categorical predictor variable with an interaction term.**



Turbo's are devices powered by small turbines that increase an internal combustion engine's efficiency and power output by forcing more air into the combustion chamber.

There is a negative linear relationship between highway mpg and curb weight in the first scatterplot and the spread is somewhat equal. We could investigate the association by considering including Aspiration Turbo, which divides car into two groups.

Group 1: Aspiration Turbo = 0 (Does not have a turbo engine) and Group 2: Aspiration Turbo = 1 (Has turbo engine), as seen in the second scatterplot. What we notice is that both groups have different intercepts and different slopes. The pink line is Group 2 and the blue line is Group 1. It seems that the unequal spread can partially be explained by Aspiration Turbo, this is why we will include it in the model.

The addition of the indicator Aspiration Turbo accounts for the different intercepts and Aspiration Turbo \* curb weight accounts for the different slopes - this is why we include the interaction term which gives us the fitted interaction model.

Highway mpg (hat) =

6.192e+01 - 2.170e+01 (Aspiration Turbo) - 1.219e-02 (curb weight) + 7.800e-03 (Aspiration Turbo:curb weight)

P-value:

(Intercept) < 2e-16 \*\*\*

Aspiration Turbo 1.86e-05 \*\*\*

curb weight < 2e-16 \*\*\*

Aspiration Turbo :curb weight 1.17e-05 \*\*\*

Coefficients:

(Intercept) 6.192e+01

Aspiration Turbo -2.170e+01

curb weight -1.219e-02

Aspiration Turbo :curb weight 7.800e-03

Group 1: Aspiration Turbo = 0 (Does not have turbo engine)

Highway mpg (hat) =

6.192e+01 Intercept + (-1.219e-02) slope

Group 2: Aspiration Turbo = 1 (Has turbo engine)

Highway mpg (hat) =

(6.192e+01 Intercept - (-2.170e+01) Aspiration Turbo) + ( -1.219e-02 Curb Weight + 7.800e-03 Aspiration Turbo : Curb Weight)

Therefore:

The indicator Aspiration Turbo is insignificant:

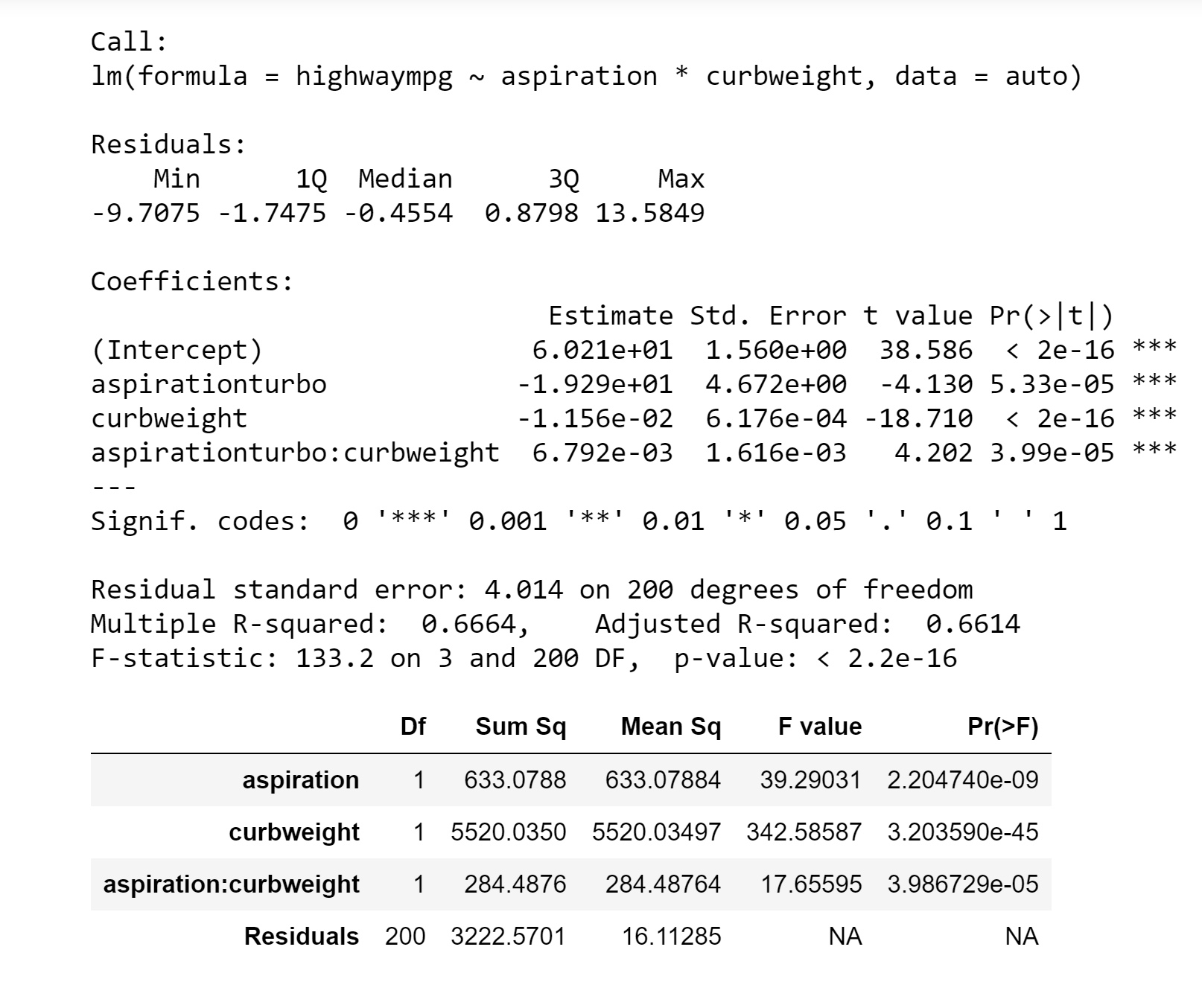
* For Aspiration Turbo = 0, the intercept is : 6.192e+01 = 16.83
* For Aspiration Turbo = 1, the intercept is :6.192e+01 - (-2.170e+01)

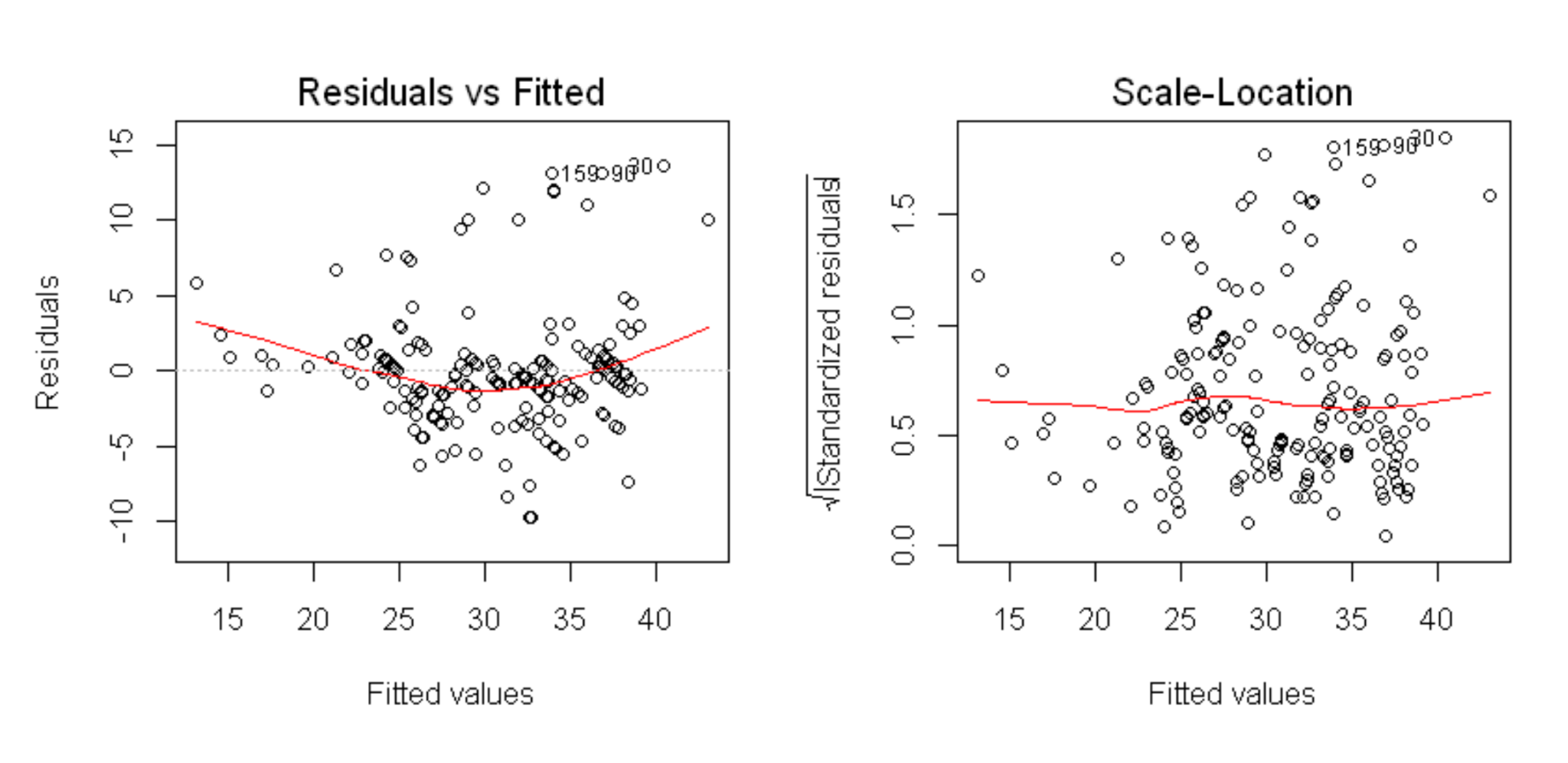
The drop of -2.170e+01 in intercept is insignificant

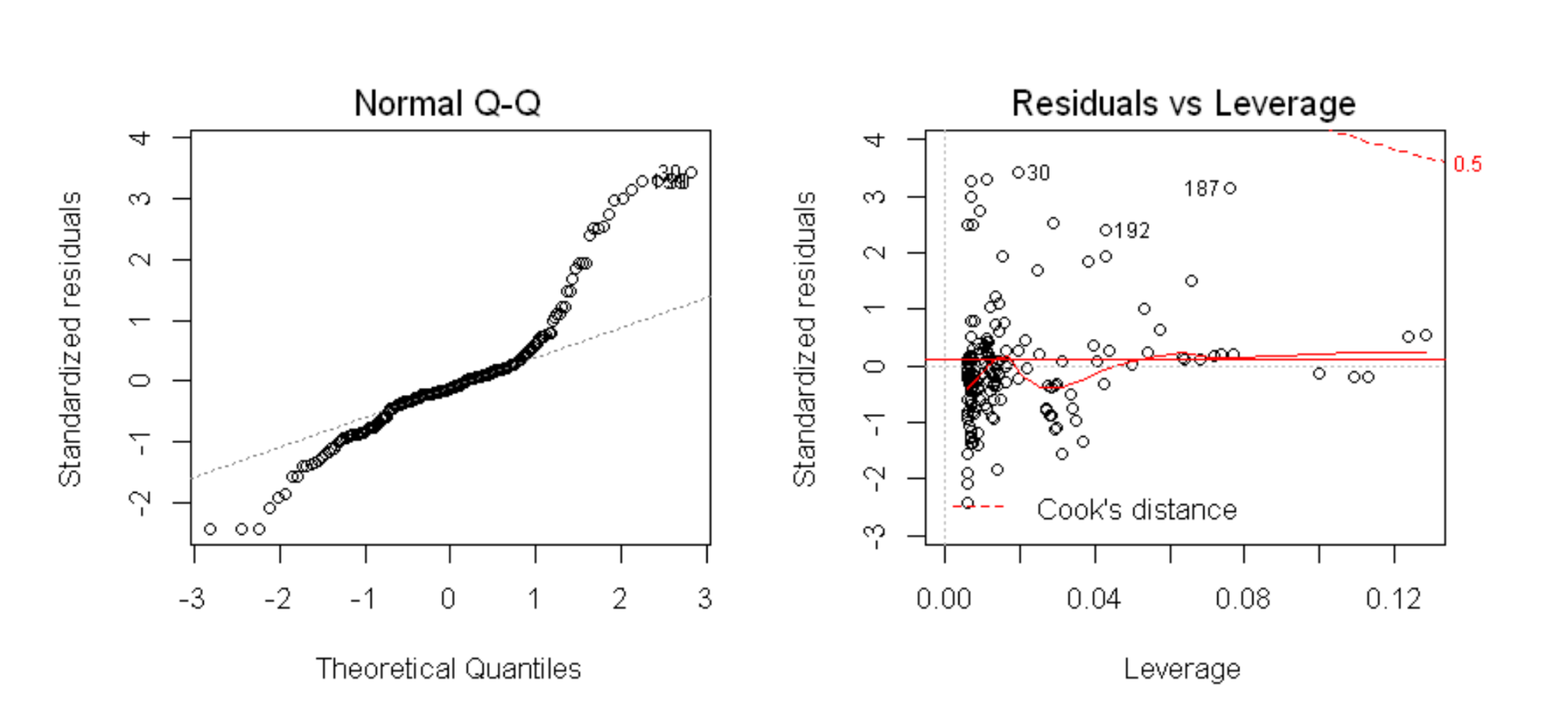
Whereas the interaction term Aspiration Turbo \* Curb Weight is significant

* For Aspiration Turbo = 0, the slope is : -1.219e-02
* For Aspiration Turbo = 1, the slope increases to: 0.2233

The increase of 7.800e-03 in slope is significant







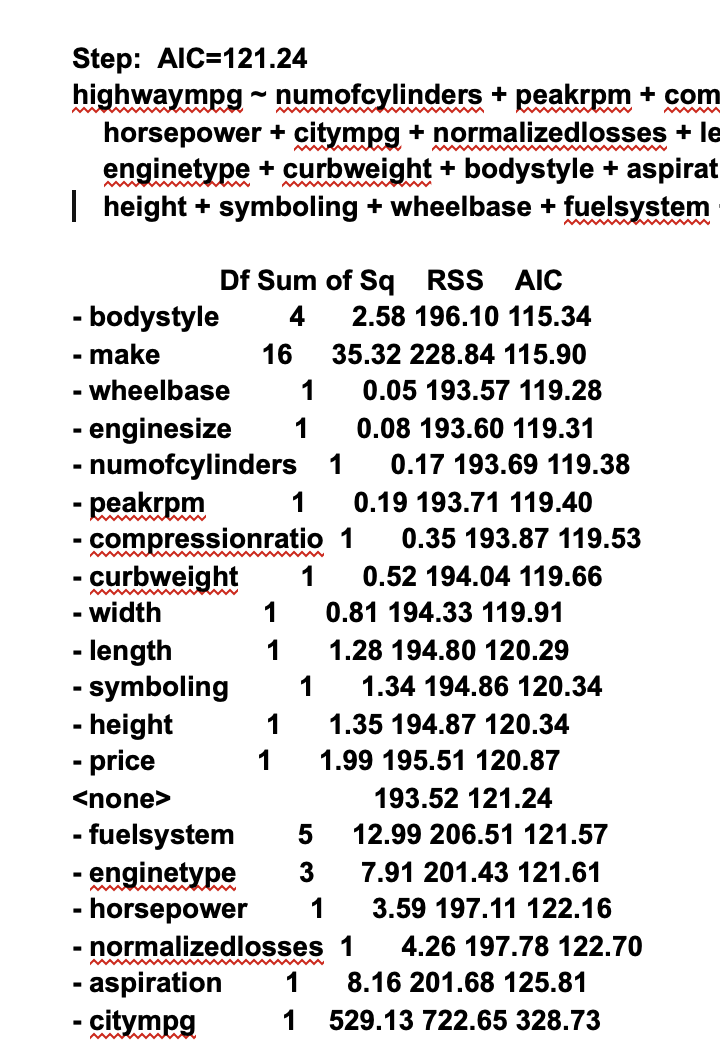
## **Model # 3: 𝐴IC criterion | Backward Elimination | Best model**

Step: AIC=95

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Degrees Freedom** | **Sum of Sq** | **RSS** | **AIC** |
| **<none>** |  |  | 254.84 | 95.00 |
| **length** | 1 | 14.16 | 269.00 | 101.60 |
| **numofcylinders** | 1 | 14.37 | 269.20 | 101.72 |
| **enginetype** | 4 | 35.35 | 290.18 | 107.66 |
| **symboling** | 1 | 24.70 | 279.54 | 107.71 |
| **curbweight** | 1 | 26.94 | 281.78 | 108.98 |
| **citympg** | 1 | 1865.24 | 2120.07 | 429.86 |

This is the best model because in the output, these predictor variables had the lowest AIC in comparison to the first one the backwards regression found (95 VS 121.4). In addition, this model had significantly less predictor variables to the other ones the step function created, which also reduces the AIC and improves the relationship between the variables.

This is the result for the highest AIC the backwards step function, compared to our best model:



**Best Model:**

**Coefficients:**

|  |  |
| --- | --- |
| **Intercept** | **-2.528367** |
| **numofcylinders** | **0.840016** |
| **citympg** | **0.951897** |
| **length** | **0.060041** |
| **enginetype** | **1.712361** |
| **enginetypeohc** | **1.751094** |
| **enginetypeohcv** | **1.161575** |
| **curbweight** | **-0.002176** |
| **symboling** | **0.380379** |