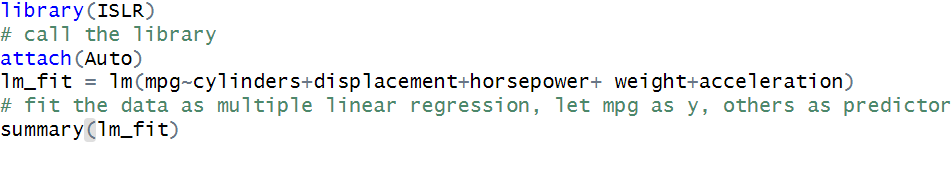
Assignment 4: Linear Regression

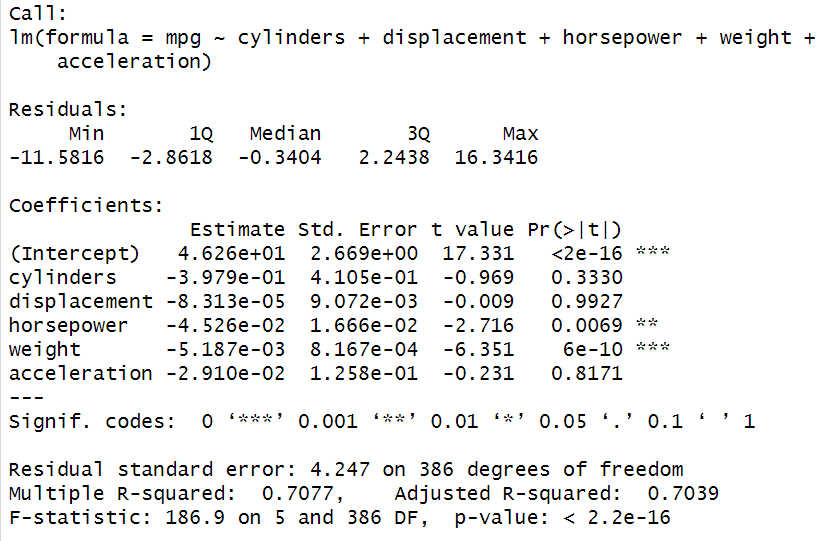
Use the Auto data set from the ISLR package to answer the following questions:

(a) Perform a simple linear regression with mpg as the response and cylinders, displacement, horsepower, weight, and acceleration as the predictors.

i. Is there a relationship between each predictor and the response?

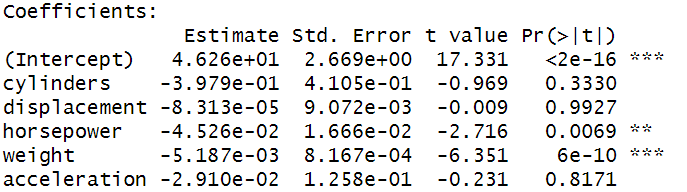
Assuming the critical value is 0.05

Coding: 

Output: 

From the above output we can see that F-statistic is 186.9, so that means at least one indicator has relationship with the response. And there are some indicator’s p values is smaller than 0.05. So, there is a relationship between each predictor and the response.

ii. How strong is the relationship between each predictor and the response?



From above output, we can see cylinder’s p value is 0.333 which is larger than 0.05, so based on multiple linear regression, there is no relationship between cylinders and MPG.

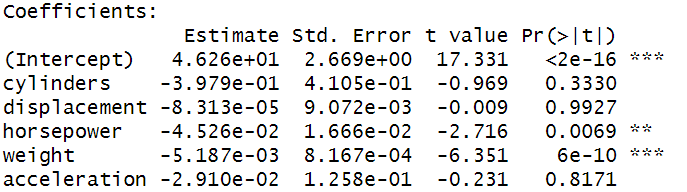
For displacement, its’ p value is 0.9927 which is larger than 0.05, so based on multiple linear regression, there is no relationship between displacement and MPG.

For horsepower, its’ p value is 0.0069 which is smaller than 0.05, so based on multiple linear regression, there is a strong relationship between horsepower and MPG.

For weight, its’ p value is 6\*10^(-10) which is significant smaller than 0.05, so based on multiple linear regression, there is a significant relationship between weight and MPG.

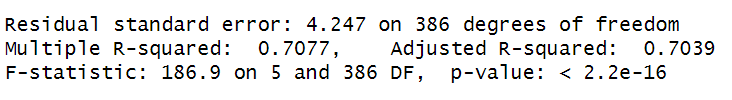
For acceleration, its’ p value is 0.8171 which is larger than 0.05, so based on multiple linear regression, there is no relationship between acceleration and MPG.

iii. Is the relationship between each predictor and the response positive or negative?



Based on the multiple linear regression output, every indicator’s coefficient is smaller than 0. So, it is negative relationship between each predictor and the response.

iv. Characterize the fitness of the developed model.



The residual standard error (RSE) is 4.247 on 386 degrees of freedom, which suggests that the average difference between the observed values of the response variable and the predicted values from the model is around 4.247 units.

The multiple R-squared value is 0.7077, which indicates that the model explains 70.77% of the variation in the response variable. This is a relatively high value, suggesting that the model is a good fit for the data.

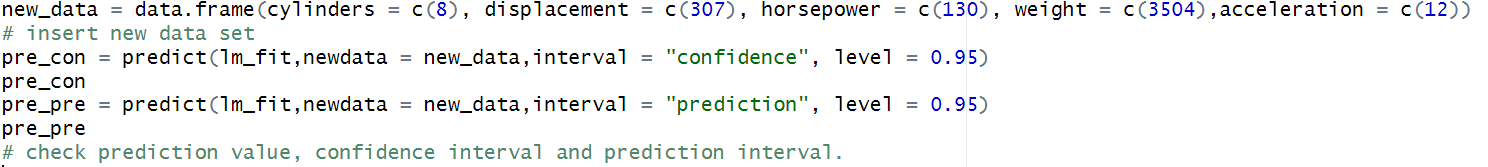
The adjusted R-squared value is 0.7039, which takes into account the number of predictor variables in the model. This value is slightly lower than the multiple R-squared value, which suggests that adding more predictor variables to the model may not substantially improve its fit.

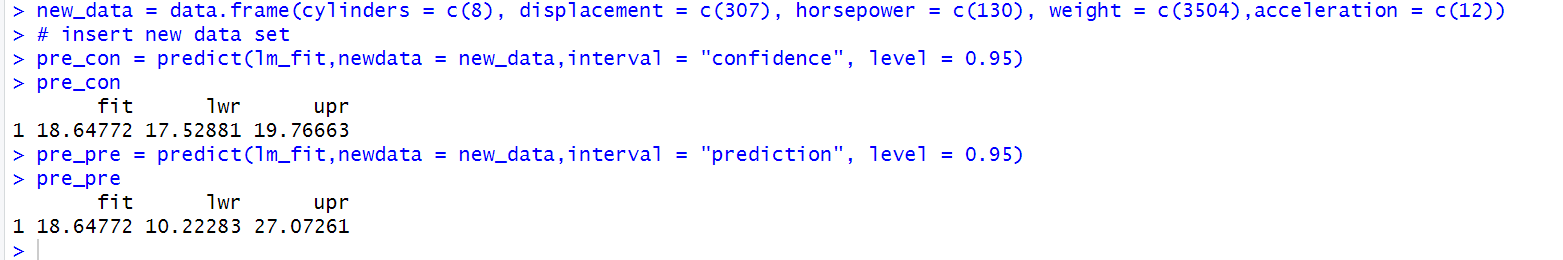
The F-statistic is 186.9 on 5 and 386 degrees of freedom, with a p-value of less than 2.2e-16. This indicates that the overall fit of the model is statistically significant.

With a high R-squared value and a significant F-statistic, we can see that the model appears to be a good fit to the data.

v. What is the predicted mpg associated cylinders = 8, displacement = 307, horsepower = 130, weight = 3504, and acceleration = 12? What are the associated 95% confidence and prediction intervals?

Code:

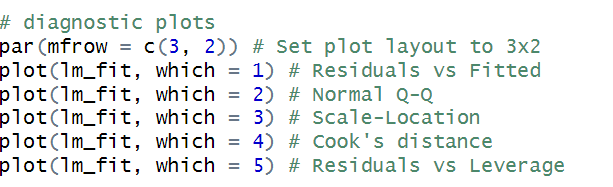


Output: 

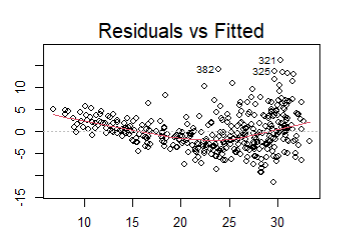
So, the prediction value of cylinders = 8, displacement = 307, horsepower = 130, weight = 3504, and acceleration = 12 is **18.64772,** and confidence interval is (17.52881,19.7663) while prediction interval is (10.22283,27.07261).

(b) Produce the diagnostic plots of the least squares regression fit. Comment on each plot.

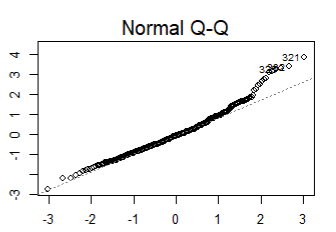
Code:



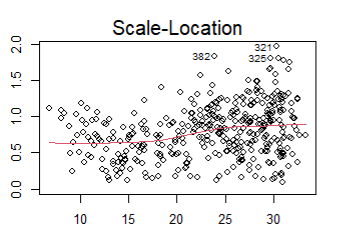
Output:



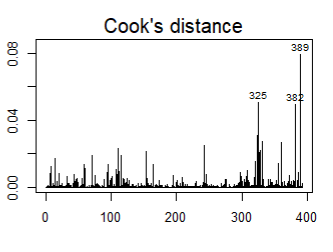
From above we can see that there is slight non-linearity.



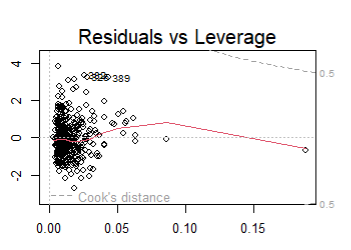
From above we can see that most of residuals follows normal distribution, until the end set of the data which are not on the line.



There are not very obvious trend shows up for this picture, so it may indicate constant variance.



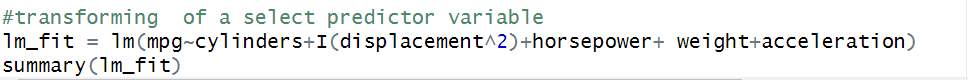
From above picture we can see that there are 3 points which are influential points. Because it is significantly higher than other points.



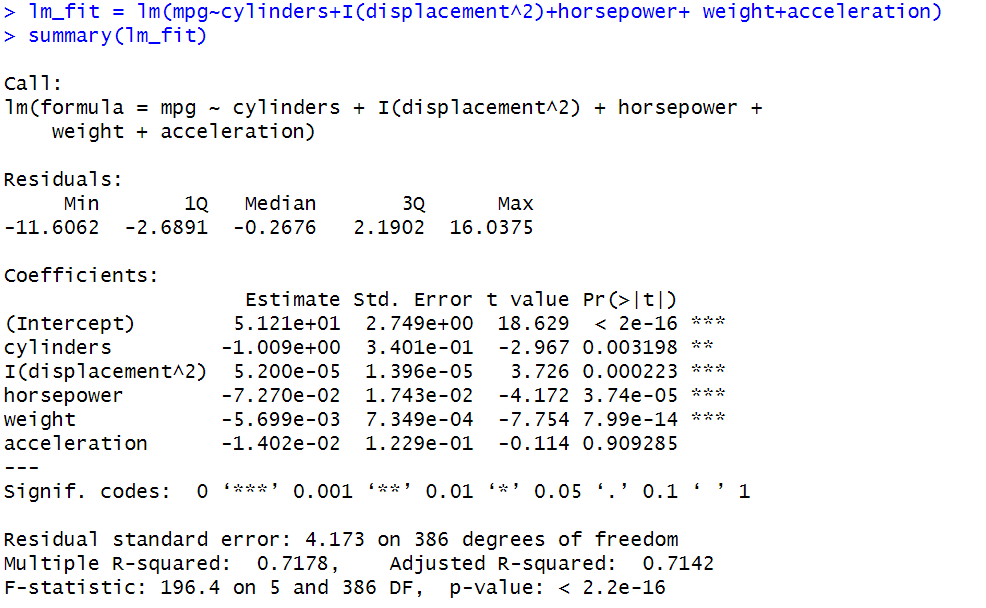
From above we can see that there is one point which is out off center. So, it means there is unusual value for Xi, we need limit the values of x.

(c) Try one transformation of a select predictor variable (such as log(𝑋), √𝑋, 𝑋^2). Comment on your findings. Did it improve model fit?

Code:



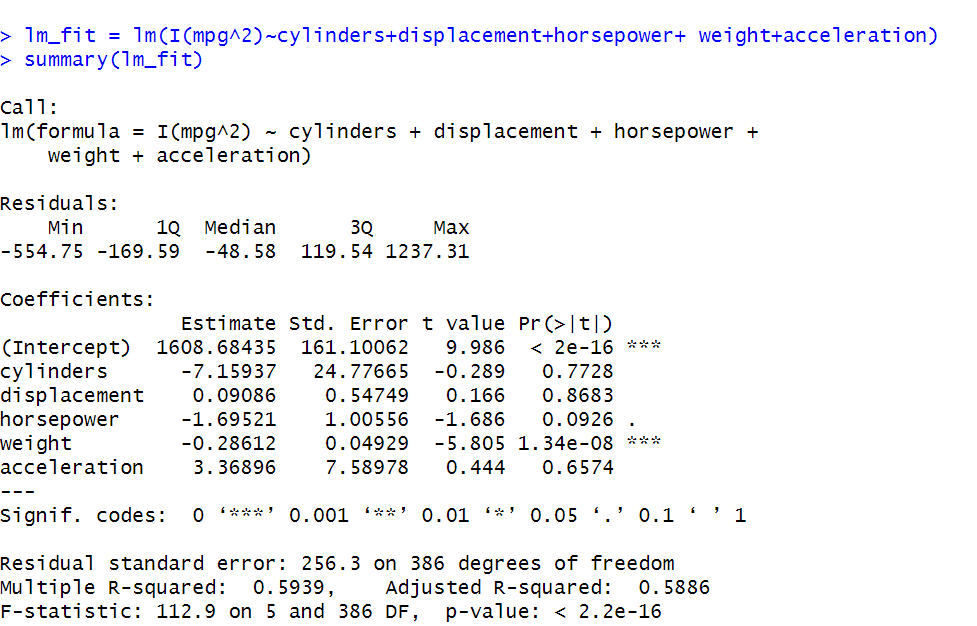
Output:



So, after transforming displacement to ^2, the R^2 and adjusted R^2 value is better than before. It means that there might be a quadratic relationship between MGP and displacement and it improves model fit.

(c) Try one transformation of the response variable. Comment on your findings. Did it improve model fit?

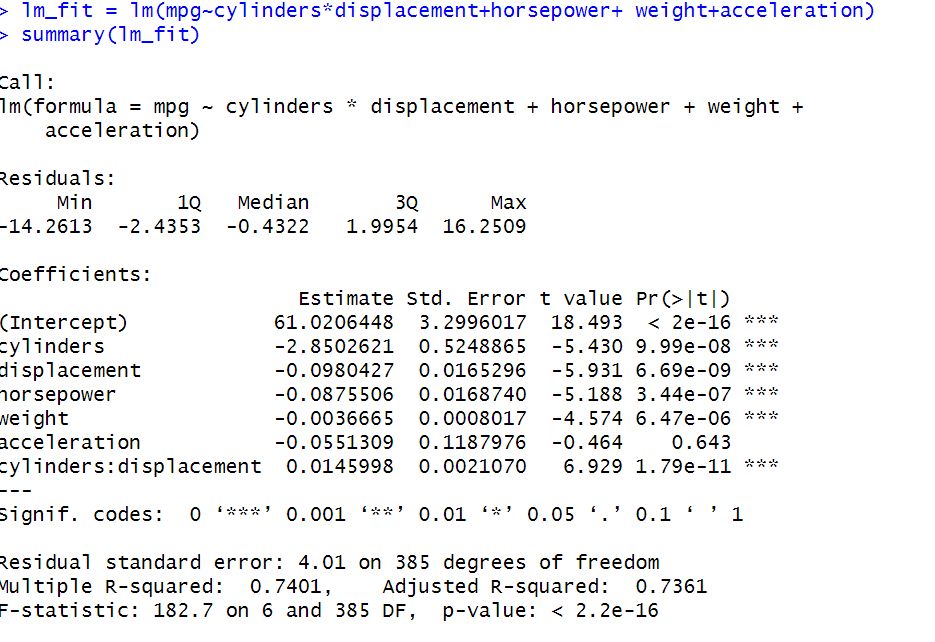
Code and output:



After transforming dependent variable to square, we can see that R^2 value and adjusted R^2 value is much lower than before. It does not improve model fit.

(d) Try one interaction effect between any two predictors. Comment on your findings. Did it improve model fit?

Code and output:



So, after interaction independent variables between cylinders and displacement, the R^2 value and adjusted R^2 value improved a lot than before. So, that means that including the interaction term in the regression model has increased the amount of variance in the dependent variable that is explained by the independent variables. And it improves the model fit.