Assignment 1

STAT862

Hardik Uppal

20132445

**3. (Multiple linear regression) Install and load the R package ISLR and consider the**

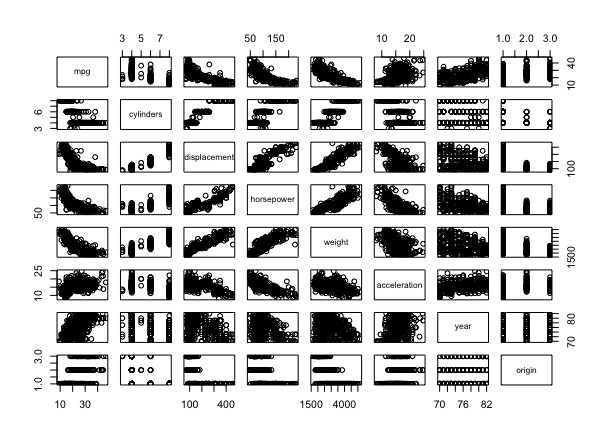
**dataset Auto in the package. Treat mpg as the dependent variable and all the other**

**variables except name as the independent variables (predictors). Note that the 1,2,3**

**of the variable origin correspond to American;European; Japanese.**

**(a) Create a pairwise scatter plot for dependent and independent variables. Show the**

**plot and make comment on the plot.**



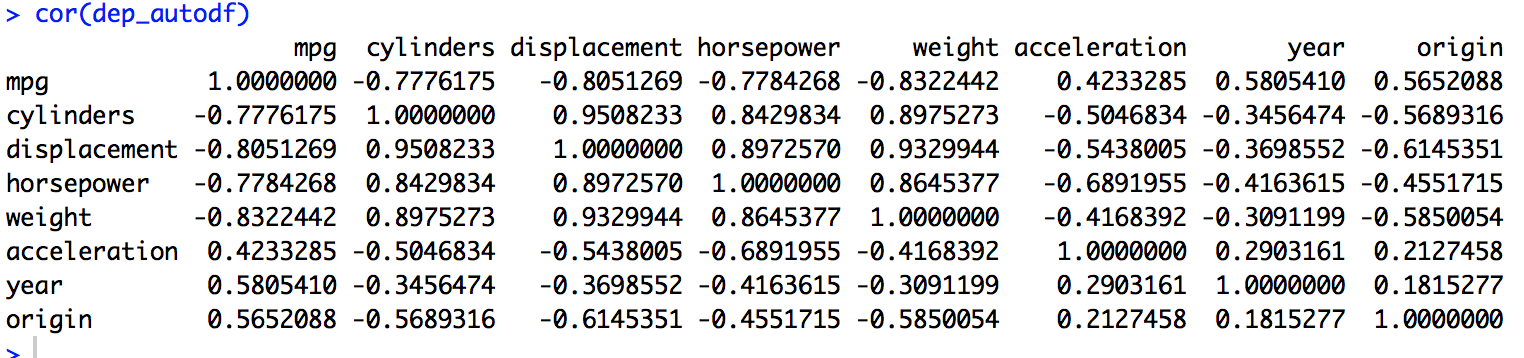
We can see that our dependent variable MPG is linearly correlated with our dependent variables: -

With increase in cylinders, displacement, horsepower, weight the value of MPG is decreasing. It shows a negative correlation.

For variables such as acceleration, year and origin the correlation is present but it is not as evident as it it with other variables. Also, it seems with increase in these variables(acceleration ,year ,origin) the value of MPG increases, positive correlation.

Also, the relation in between the independent variable is also worth noting. Cylinders, displacement, horsepower, weight are in-fact highly correlated to each other. We can clearly see in the pattern that with the increase in cylinders the other variables (displacement, horsepower, weight) are increasing, and with increase in displacement the other three variables increase linearly.

**(b) Computer the correlation matrix between the variables using cor() function.**



**(c) Fit the multiple linear regression model. Show the table of the fitted model:**

**coefficients estimation, their standard deviation, t-statistic, and p-values. Show**

**R2 and the estimation sigma^\_2.**

For the model we take the null hypothesis that all of the coefficients for the independent variables, or the attributes are not correlated with the dependent variable.

We have the following basic assumptions: -

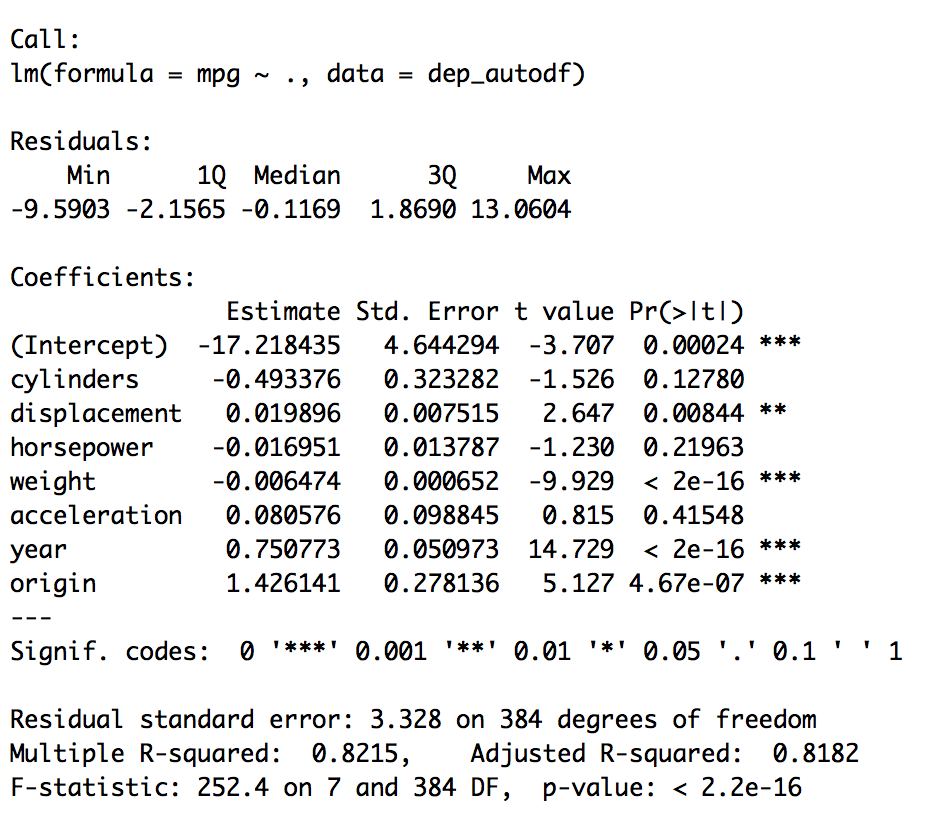
* There is a linear relationship between dependent variable and the independent variable.
* For given values of the independent variables our dependent variable is distributed normally.
* Variation of observed values is constant around the regression line (Homoscedasticity).
* No or little multicollinearity.
* No auto-correlation.

Also we take the value of alpha/significance level =0.05

**Model: -**

fit = lm(data = dep\_autodf, formula = mpg ~ .)

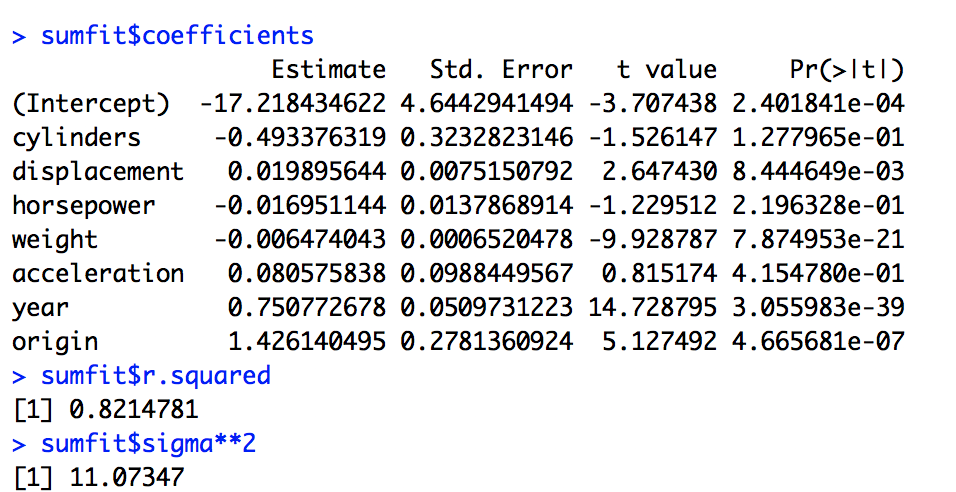
**Model summary: -**



From the model summary we can see that the P-value**(<2.2e-16)** is far less than the significance level or **alpha (0.05)** and hence we can conclude that our model fit the data better than the only intercept model.

Below are the coefficients - estimation, their standard deviation, t-statistic, and p-values

Also, the R2 and the estimation sigma2



**(d) Obtain the prediction of mean response, its associated prediction error and 100(1-alpha) % confidence interval based on the fitted model for the new input cylinders =**

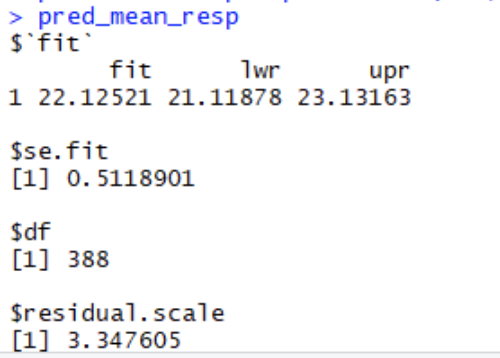
**8; displacement = 300; horsepower = 150; weight = 3600; acceleration = 13; year =**

**77; origin = 3.**

R code attached

The predicted mean response to the new data set is 22.12521 with 95% confidence intervals as 21.11878 and upper limit as 23.13163

The prediction standard error is 0.5118901



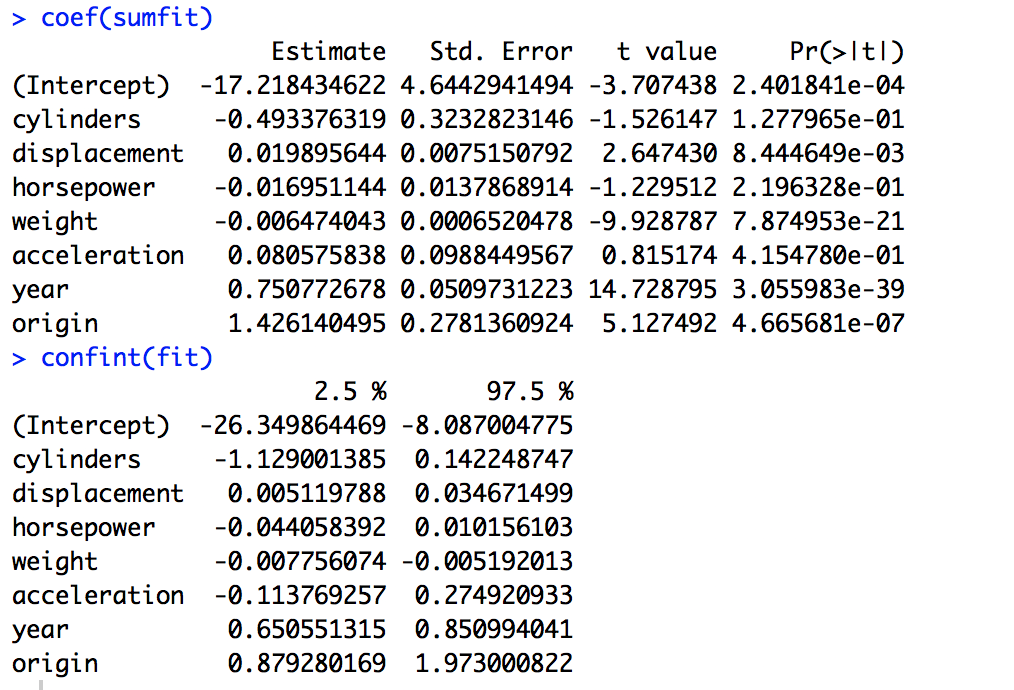
**(e) Is there relationship between the independent variables and the response variable?**

From the model summary we can see the p-value(<2.2e-16) for the model is far less than the alpha (0.05) and hence there is a relationship between out response variable and independent variable and our null hypothesis is false that there exist no relationship between Response variable and Independent variable.

Also, the **F-statistic value (252.428**) is large enough to state that we can reject our null hypothesis.

**(f) Which predictors appear to have a statistically significant relationship to the**

**response?**



We can see from the Coefficient table that some of the variables have a lower p-value than the alpha (0.05) – **year, origin, weight, displacement** all have **p-values less than 0.05** hence they are more significant to the model than the other variables

We can also the the weights of each variable given by then coefficients which would indicate the change in mpg given all other variables are kept constant. Year, origin have comparatively the largest values in coefficients.

Confint() gives us the 95% confidence interval values of the coefficients.

**(g) Produce the residuals plots.**

R code attached