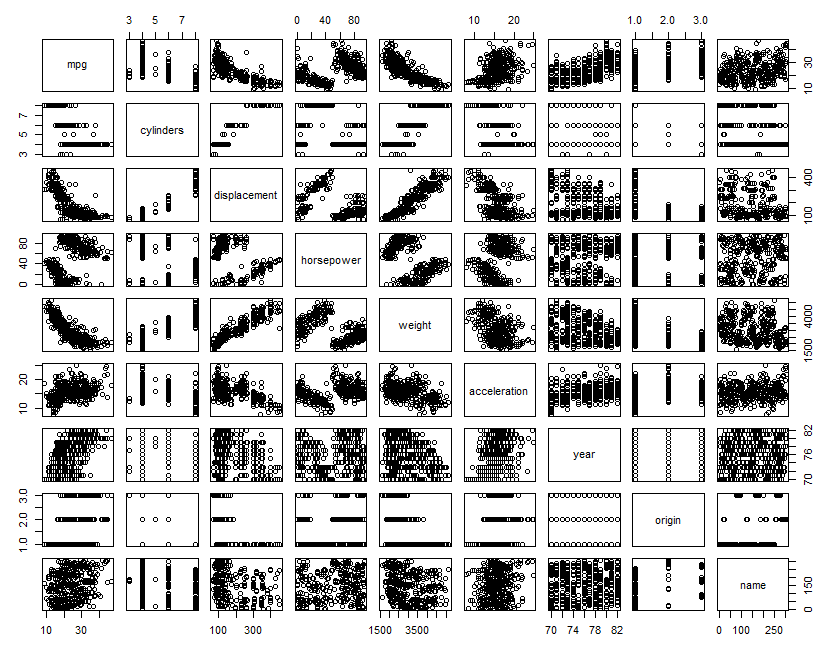
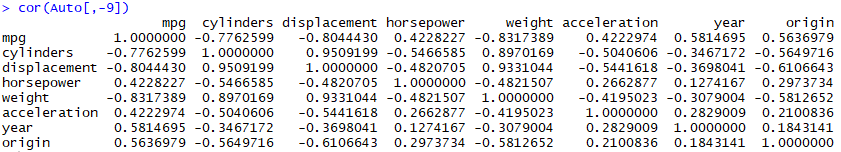
9) This question involves the use of multiple linear regression on the Auto data set.

1. Produce a scatterplot matrix which includes all of the variables in the data set.

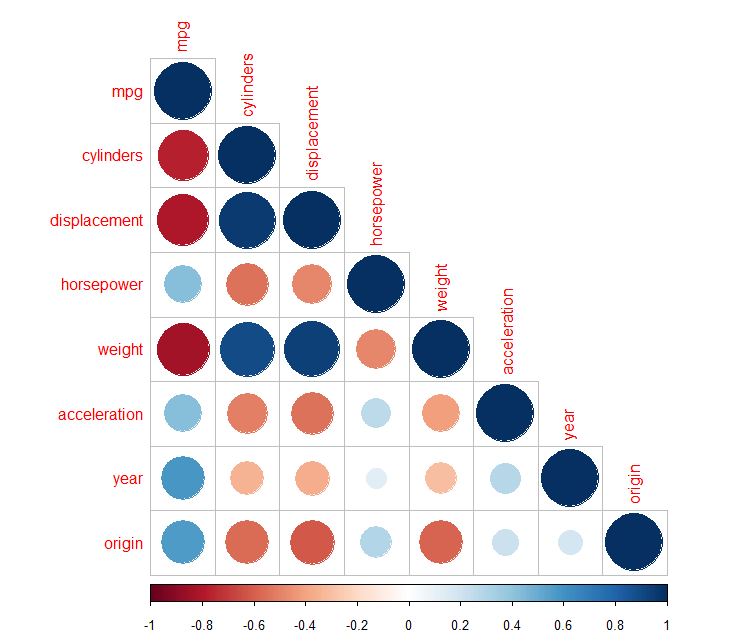


1. Compute the matrix of correlations between the variables using the function cor (). You will need to exclude the name variable, which is qualitative.

R output:

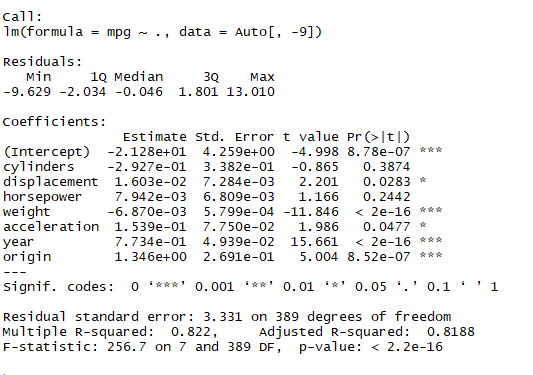


Plot Correlation:



1. Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output.

R output:



1. Is there a relationship between the predictors and the response?

Yes, as R2= 82.2% i.e. Based on independent variables the linear regression model can predict 82.2% variability in dependent variable.

1. Which predictors appear to have a statistically significant relationship to the response?

If alpha=5% then following predictors appear to be statistically significant:

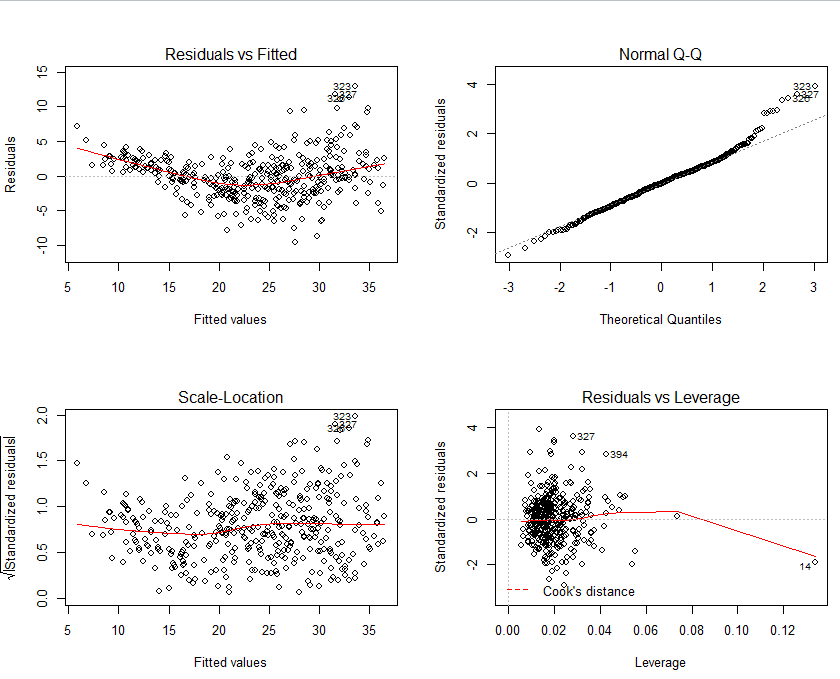
|  |  |
| --- | --- |
| Predictors | P-value |
| displacement | 0.0283 |
| weight | < 2e-16 |
| acceleration | 0.0477 |
| year | < 2E-16 |
| origin | 8.52E-07 |

1. What does the coefficient for the year variable suggest?

For every 1 unit increase in year there is 0.7734 unit increase in mpg.

1. Use the plot () function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

R output:



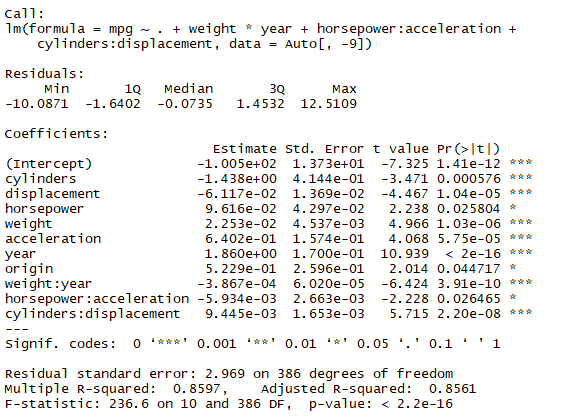
Observations:

1. Residual vs Fitted: The plot shows that the residuals have non-linear patterns. There is a non-linear relationship between predictor variables and the response variable as the residuals are not equally spread around the horizontal line.
2. Normal Q-Q: Plot shows that the residuals are normally distributed, although there are some exceptions as observation 323, 326 and 327 which seem off.
3. Scale-Location: This plot shows that the residuals are not spread equally along the ranges of predictors i.e. the plot shows heteroscedasticity.
4. Residuals vs Leverage: It looks like there are no influential cases, although there are outliers (observations 394, 327 and 14) but all of them are inside the cook’s distance line. That means they even though they have extreme values, they would not affect the regression line and the results wouldn’t be much different if we either include or exclude them from analysis.
5. Use the \* and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

Ans: Following interactions are statistically significant:

* Weight and Year
* Horsepower and acceleration
* Cylinder and displacement

R output:



1. Try a few different transformations of the variables, such as log(*X*), *√X*, *X*2. Comment on your findings.

