Assignment 1

Mrunal Ghorpade

UIN: 677441117

1. Compare the classification performance of linear regression and k–nearest neighbor classification on the zipcode data. In particular, consider only the 2’s and 3’s, and k = 1, 3, 5, 7 and 15. Show both the training and test error for each choice. The zipcode data are available from the book website www-stat.stanford.edu/ElemStatLearn.

Linear Regression:

For linear regression R2 is 90.05% which means model can explain 90.05% variation in the training data. The RMSE error for training data is 0.1575174 and for test data is 0.3894424.

As linear regression gives output as continuous variable, for converting the predicted output in the form of classifier i.e. 2’s or 3’s, threshold value of 2.5 is selected and so if the predicted value is more than 2.5 then it is converted to 3 else 2.

Following is the confusion matrix;

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  | Training data | | | Test data | | |  |
|  |  |  |  |  |  |  |  |
|  | |  | | --- | |  | |  |  | |  | | --- | |  | |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  | Error Rate= (3+5)/(1389) =0.005759 | | | Error Rate = (7+8)/(364)= 0.0412 | | |  |
|  | Accuracy: 1- Error= 99.4241% | | | Accuracy = 95.879% | |  |  |

K nearest neighbor:

Following are the Error rates for different values of K;



By comparing linear Regression model and K-n-n with different values of K it can be seen that for k=1 classification performance has the accurate predicted output.

2) What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

In less flexible approach we assume a model which gives a functional form of original function (f). Now we just estimate a set of parameters rather than estimating f. Disadvantage of this is that there is always a probability that the functional form used to estimate f is very different than true f, in which case the resulting estimate will be different from the data. This issue can be addressed by using more flexible approach where no assumptions regarding the functional form of f is made. Flexible approach considers all predictors from the data due to which they have more probability to accurately fit the original function. In general, fitting a more flexible model requires estimating a greater number of parameters. But as the model is too flexible, it can overfit the train data leading to large test error.

Less flexible approach can be used when we are interested in interference. When we want to understand how an individual predictor is associated with the response we use less flexible approach. However, when we are only interested in prediction and not the interpretability we use more flexible approach.

3) The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Obs. | X1 | X2 | X3 | Y |
| 1 | 0 | 3 | 0 | Red |
| 2 | 2 | 0 | 0 | Red |
| 3 | 0 | 1 | 3 | Red |
| 4 | 0 | 1 | 2 | Green |
| 5 | -1 | 0 | 1 | Green |
| 6 | 1 | 1 | 1 | Red |

Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.

1. Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

Ans: Euclidean distance between each observation in training data and the test point is as follows:



(Note: Also computed in R as shown below and validated the R code results by above calculations)

R output:



1. What is our prediction with *K* = 1? Why?

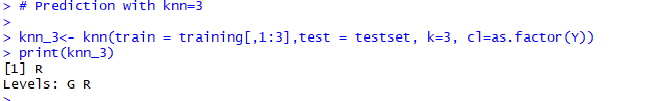
R output:



The prediction is Green with K=1. For K=1, we pick 1 nearest neighbor for the data set from the training data. The closest one is Green i.e. X1 = -1, X2 = 0, X3 = 1 from the training data set at 1.414214 and therefore the prediction was Green.

1. What is our prediction with *K* = 3? Why?

R output:



The prediction is Red with K=3. For K=1 we pick 3 nearest neighbors for the data set from the training data predict on the bases of whichever color occurs most number of time. The 3 closest neighbors are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Y | X1 | X2 | X3 | Distance |
| Green | -1 | 0 | 1 | 1.414214 |
| Red | 1 | 1 | 1 | 1.732051 |
| Red | 2 | 0 | 0 | 2 |

As there are 2 Red’s and 1 Green the prediction was Red.

1. If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the *best* value for *K* to be large or small? Why?

When K is small the decision boundary of K-n-n is more flexible but as K grows, the method becomes less flexible and produces a decision boundary that is close to linear. Now if the Bayes decision boundary is highly non-linear then we would need small K values. Whereas when the boundary becomes more rigid, we need large K values.

4) This exercise involves the Boston housing data set.

1. How many rows are in this data set? How many columns? What do the rows and columns represent?

Ans:

Number of rows: 506

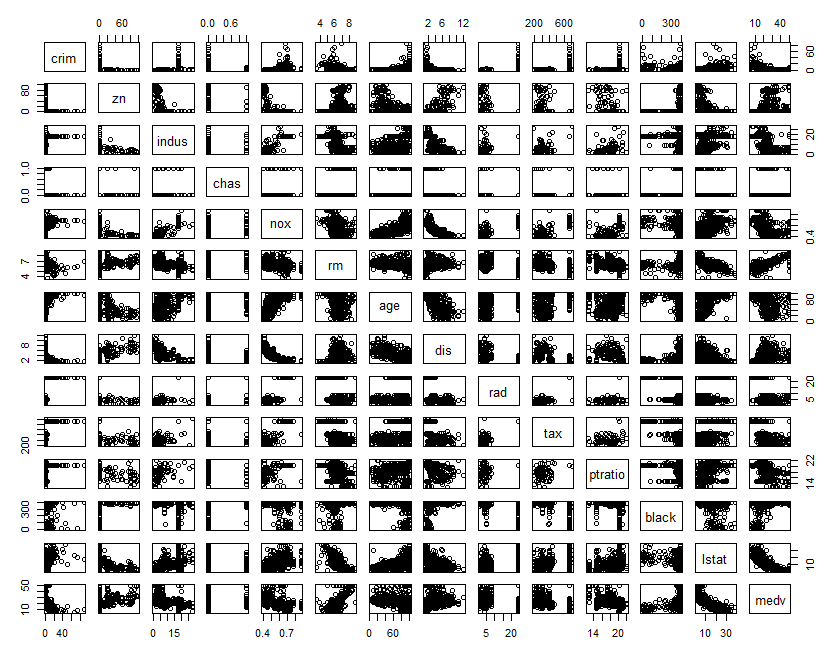
Number of columns: 14

The data represents number of variables for 505 different records of housing values in suburbs of Boston. The variables are as follows:

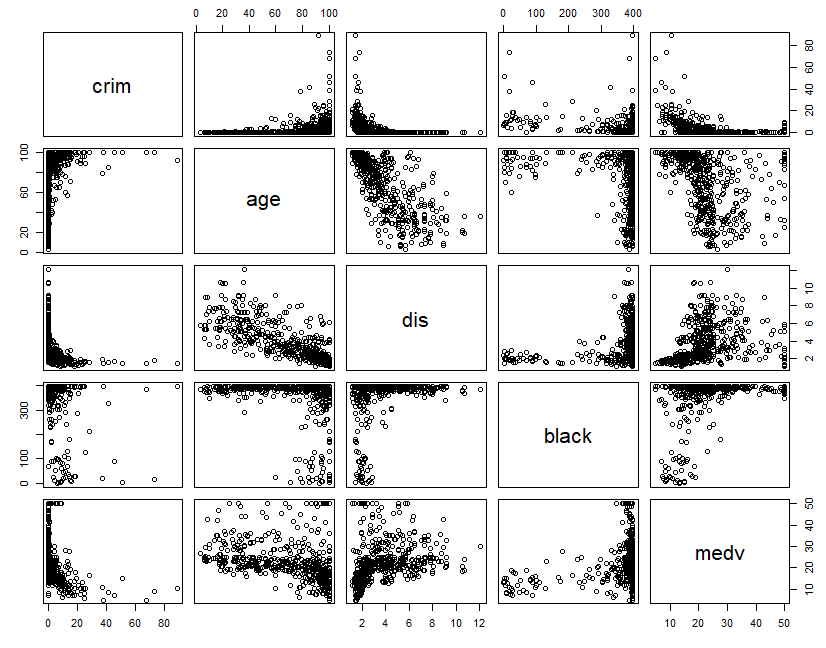
|  |  |
| --- | --- |
| Column Names | Description |
| crim | per capita crime rate by town. |
| zn | proportion of residential land zoned for lots over 25,000 sq.ft. |
| indus | proportion of non-retail business acres per town. |
| chas | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). |
| nox | nitrogen oxides concentration (parts per 10 million). |
| rm | average number of rooms per dwelling. |
| age | proportion of owner-occupied units built prior to 1940. |
| dis | weighted mean of distances to five Boston employment centers. |
| rad | index of accessibility to radial highways. |
| tax | full-value property-tax rate per \$10,000. |
| ptratio | pupil-teacher ratio by town. |
| black | 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town. |
| lstat | lower status of the population (percent). |
| medv | median value of owner-occupied homes in \$1000s. |

1. Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings

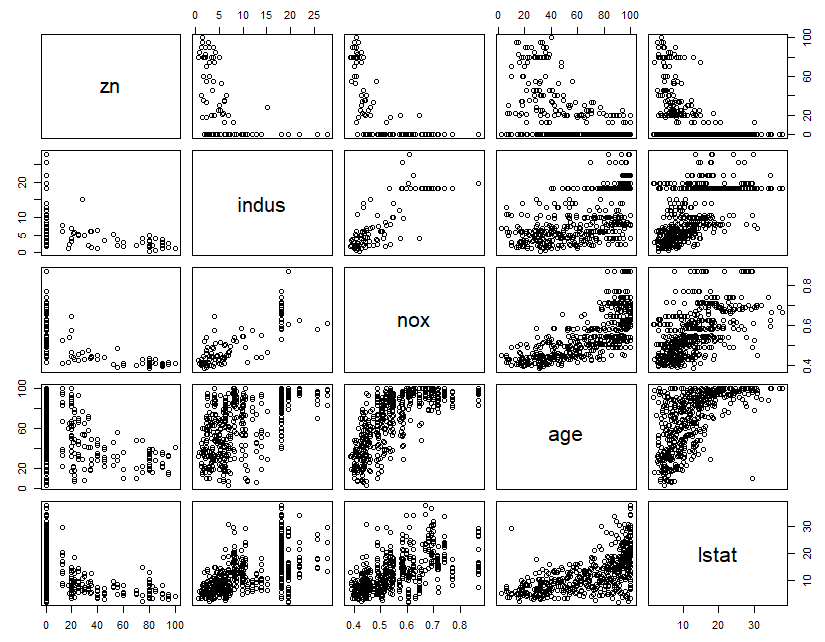
R output: Scatterplot for all the Predictor columns is as follows;



It is very difficult to understand the relation between all the columns when we plot all of them together, so visualizing them pairwise as follows;



The scatterplot shows that the crime is more near older homes. Near the employment centers the crime rate is more. As distance to the Employment centers increases, more newer houses are observed. Crime rate is less where the owner owns expensive houses.

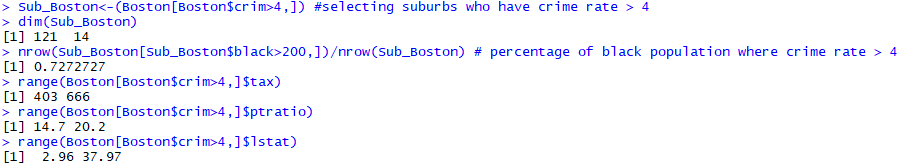


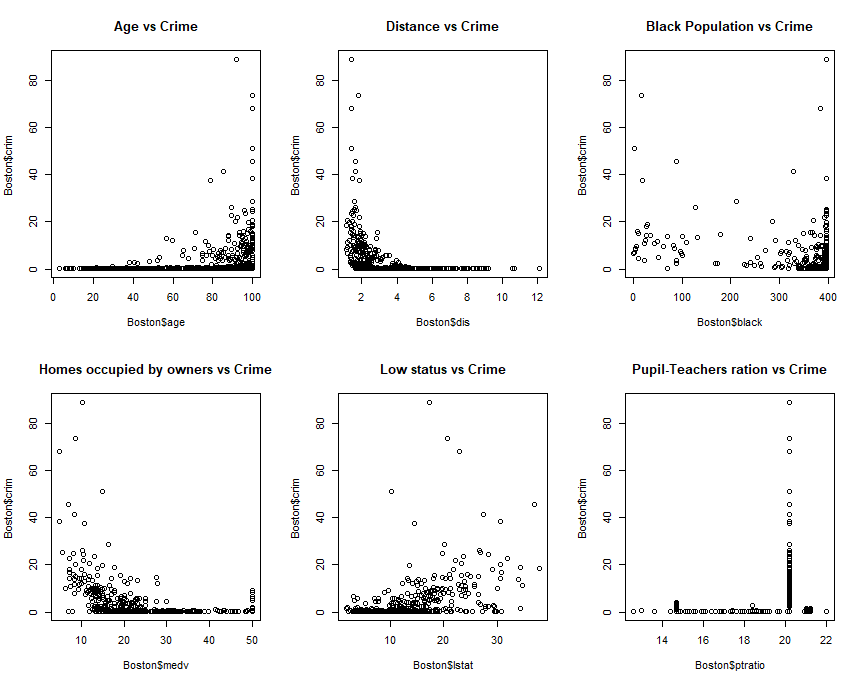
The proportion of residential land decreases as the proportional of non-residential business acres increase. Nitrogen oxide concentration is lower where proportion of residential land more. For the older houses Nitrogen oxide concentration is high. Percentage of lower status population staying in residential area is less.

1. Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

Ans: For the suburbs were the crime rate higher than 4, tax rates are high (ranging from 403 – 666), 72.72% of the black population stays in those suburbs, lower status population is more (range: 2.96-37.97) and pupil-teacher ratio is high (range: 14.7 – 20.2). Also, the plots show that crime rate is more near many older buildings. Near the employment centers the crime is more. Crimes are less where the owner owned expensive houses.

R output:



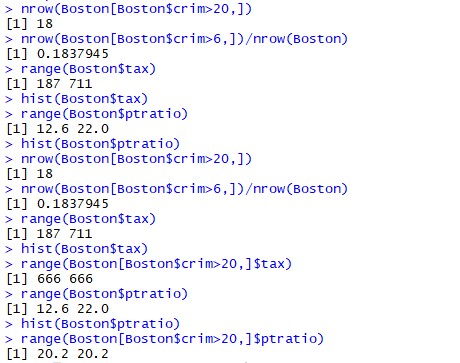


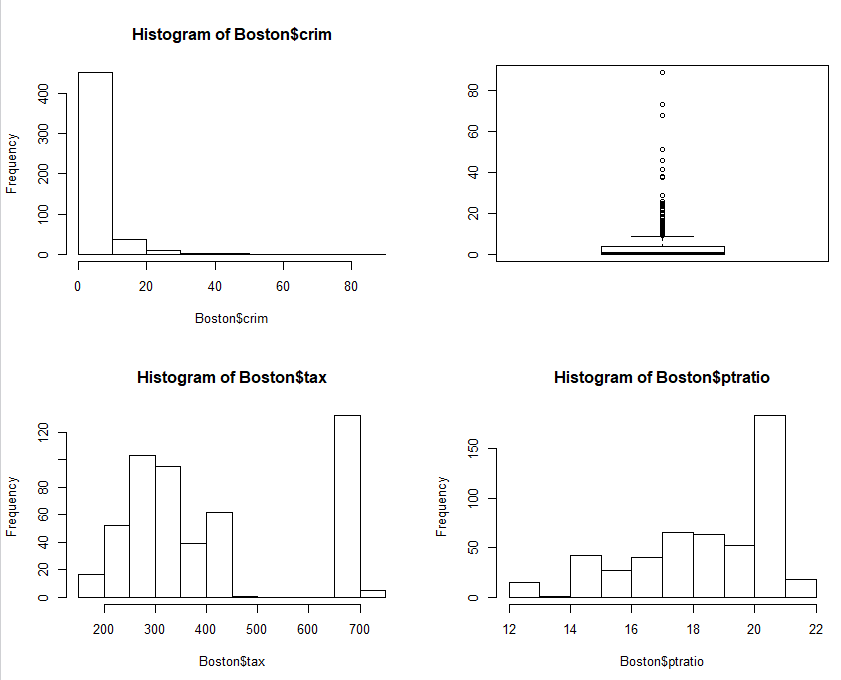
1. Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

Ans:

* The histogram of crime is right tailed which shows most cities have less crime rate, but 18 suburbs have crime rate more than 20 and only 18% of suburbs have crime rate more than 6
* Tax rates are lower in most of the suburbs, but have a peak near 700 (Tax Rates Range: 187-711) for some suburbs and for those suburbs crime rate is greater than 20 (tax rates are 666)
* Histogram for Pupil-teacher ratio is left skewed and most of the suburbs have higher pupil-teacher ratio. For Crime-rate greater than 20 the pupil teacher ratio is 20.2

R output:





1. How many of the suburbs in this data set bound the Charles river?

Ans: 35 suburbs in data set bound to Charles river

R output:



1. What is the median pupil-teacher ratio among the towns in this data set?

Ans: The median pupil-teacher ratio = 19.05

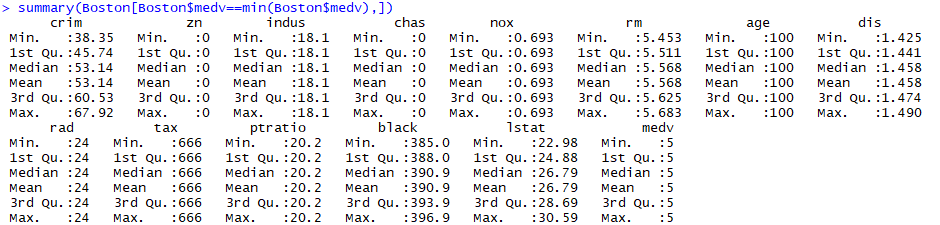
R output:



1. Which suburb of Boston has lowest median value of owner occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

Ans: Suburbs that have following predictors have lowest value of owner occupied homes;

|  |  |  |
| --- | --- | --- |
| Columns | Range | Comments |
| crim | 38.3518-67.9208 | above 3rd quintile |
| zn | 0 | Least |
| indus | 18.1 | at 3rd quintile |
| chas | 0 | Not near the river |
| nox | 0.693 | above 3rd quintile |
| rm | 5.453-5.683 | below 1st quintile |
| age | 100 | Maximum |
| dis | 1.4896-1.4254 | below 1st quintile |
| rad | 24 | Maximum |
| tax | 666 | at 3rd quintile |
| ptratio | 20.2 | at 3rd quintile |
| black | 396.9-384.97 | above 1st quintile |
| lstat | 30.59-22.98 | above 3rd quintile |
| medv | 5 | Least |

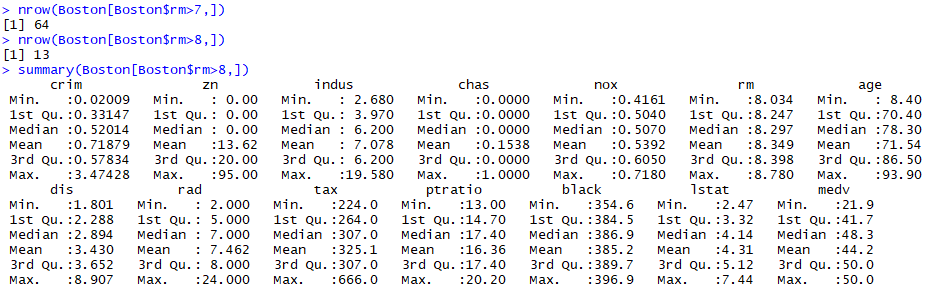
R output:

1. In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

Ans: 64 suburbs have average more than seven rooms per dwelling. 13 suburbs have average more than 8 rooms per dwelling. The Suburbs with average more than 8 rooms per dwelling have following predictors;

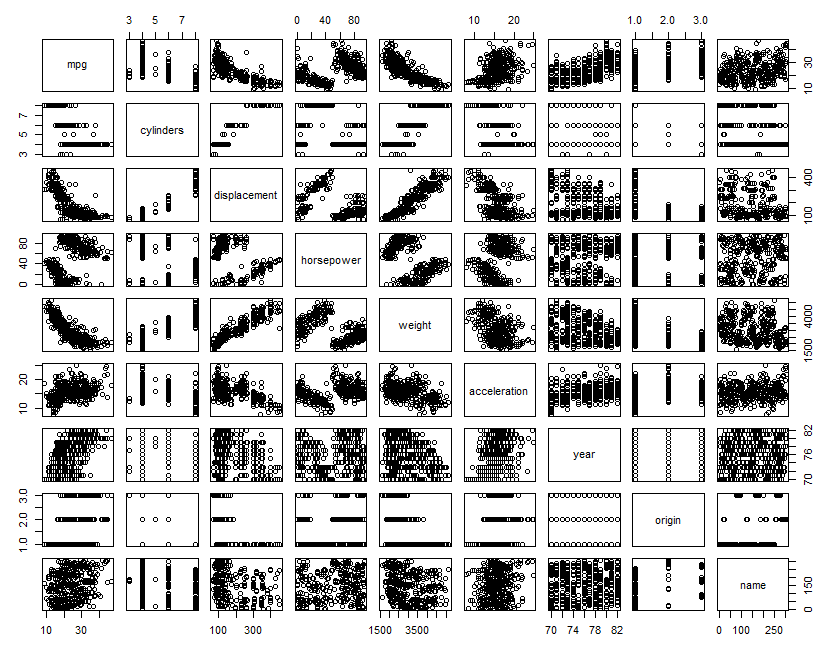
|  |  |  |
| --- | --- | --- |
| Columns | Range | Comments |
| crim | 0.02-3.47 | below 3rd quintile |
| zn | 0-95 | No effect |
| indus | 2-19.58 | below 3rd quintile |
| chas | 0-1 | not bounded by river |
| nox | 0.41-0.71 | above 1st quintile |
| rm | 8.03-8.78 | Maximum |
| age | 8.40-93.9 | Minimum |
| dis | 1.8 - 8.9 | No effect |
| rad | 2.0-24.0 | No effect |
| tax | 224-666 | above 1st quintile |
| ptratio | 13-20.2 | below 3rd quintile |
| black | 354.6-396.9 | above 1st quintile |
| lstat | 2.47-7.44 | below 1st quintile |
| medv | 21.9-50 | above 3rd quintile |

R output:



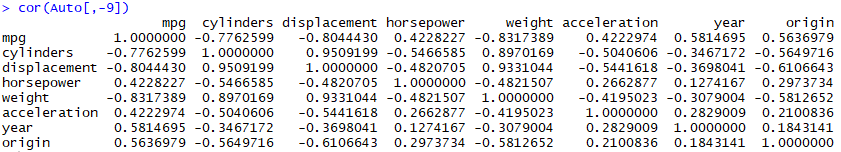
9) This question involves the use of multiple linear regression on the Auto data set.

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

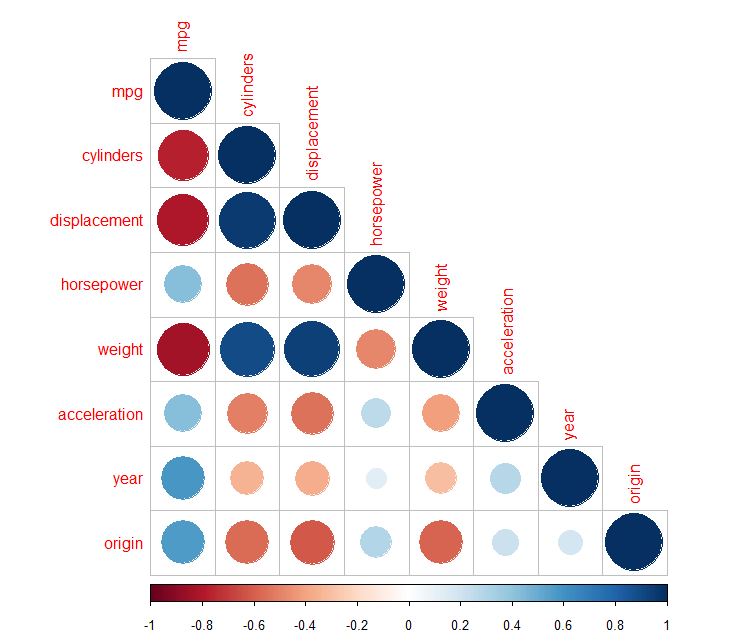


1. Compute the matrix of correlations between the variables using the function cor (). You will need to exclude the name variable, which is qualitative.

R output:

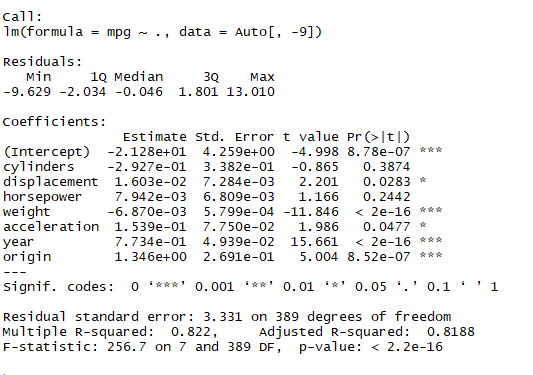


Plot Correlation:



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?

Ans:

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?

Ans:

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 |

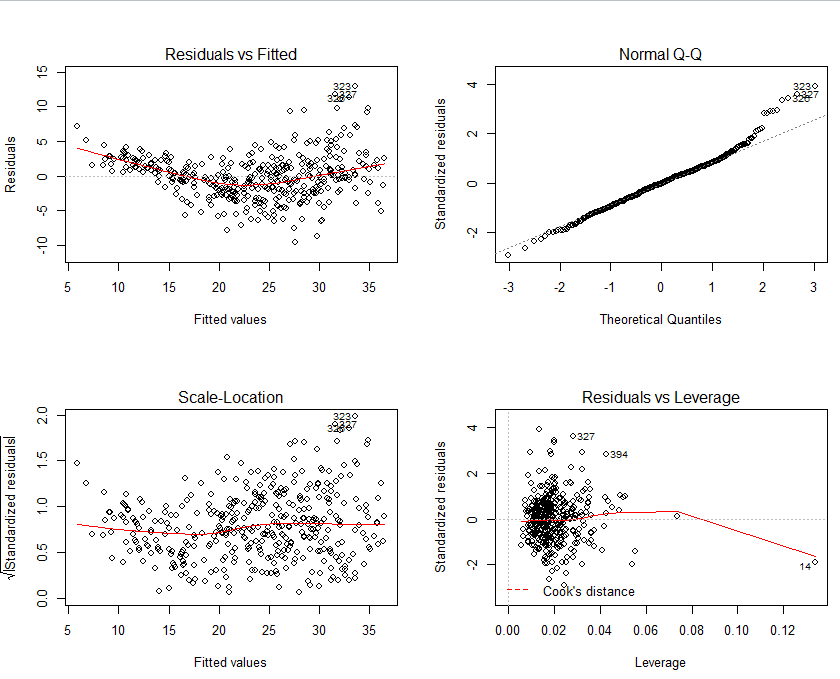
1. What does the coefficient for the year variable suggest?

Ans:

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:

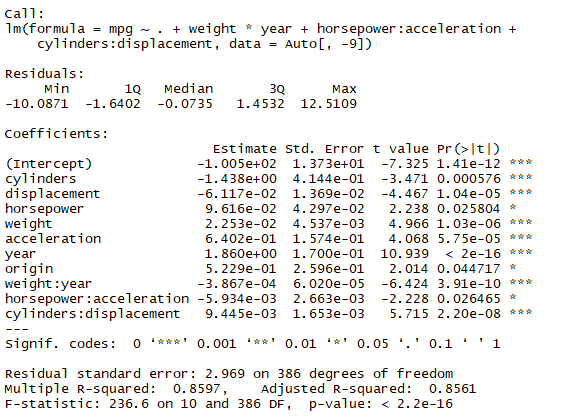
1. Residual vs Fitted: The plot shows that the residuals have non-linear patterns. There is a non-linear relationship between predictor variables and the response variable as the residuals are not equally spread around the horizontal line.
2. Normal Q-Q: Plot shows that the residuals are normally distributed, 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.

Ans:

R2 has increased from 82.2% to 86.21% which mean model is able to explain higher amount of variation in data. Also, Residual vs Fitted plot shows that the non-linearity in the data is decreased and Residuals are equally spread around the horizontal line showing homoscedasticity.

R output:

