Jiahuan He ID: 912490740

Before reading the data into Matlab, observations with missing data are deleted. Then,

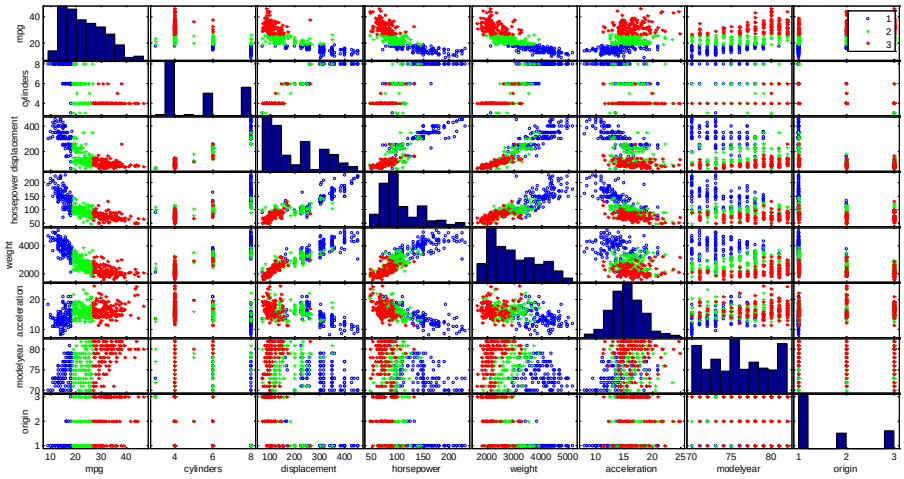
[mpg, cylinders, displacement, horsepower, weight, acceleration, modelyear, origin, carname] = textread('auto-mpg.txt', '%n %d %n %n %n %d %d %q')

Q1.

```
>> quantile1=quantile(mpg, 1/3)

quantile1 =

    18.6667

>> quantile2=quantile(mpg, 2/3)

quantile2 =

    26.9667
```

That is, mpg under 18.6667 will be classified as low, medium is between 18.6667 and 26.9667, while high is those mpg over 26.9667.

Q2.



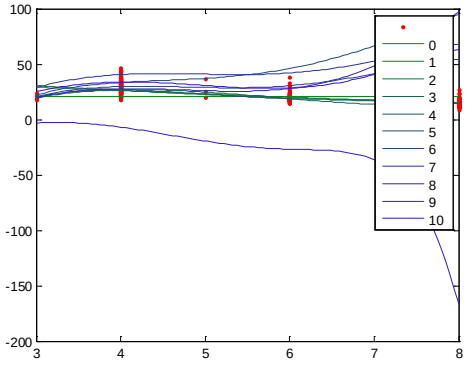
In the legend of the figure, 1 is for “low”, 2 is for “medium” and 3 is for “high”. There are some visible relationships between “mpg”, “displacement”, “horsepower” and “weight”. The plot “mpg-modelyear” shows that as years passed, cars become more and more efficient.

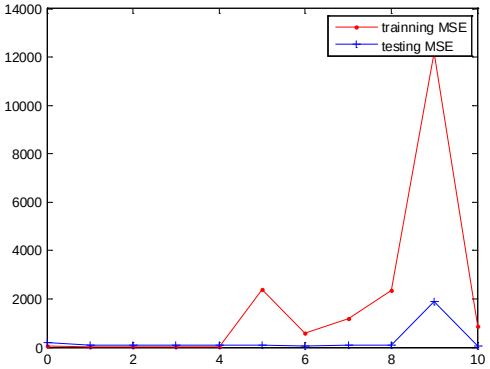
Q3.

See file “hw1q3.m”.

Q4.

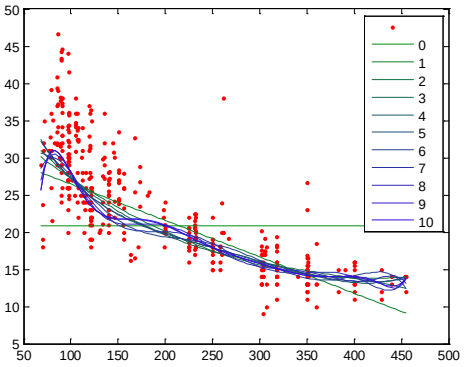
Cylinders:

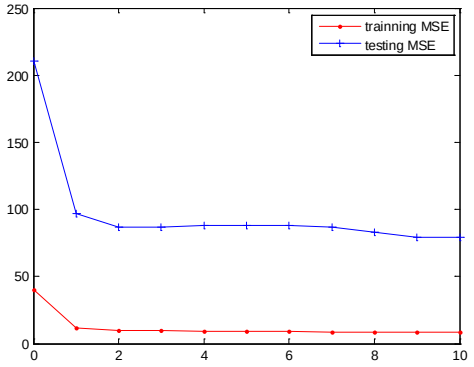




Cylinders seems to have little effect on mpg.

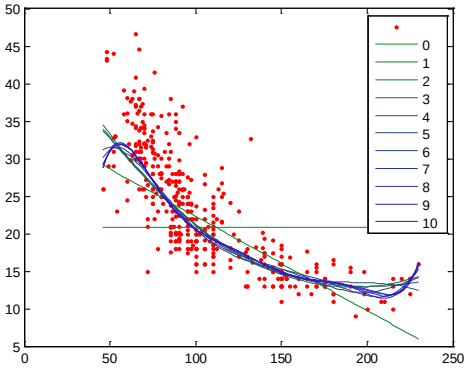
Displacement:

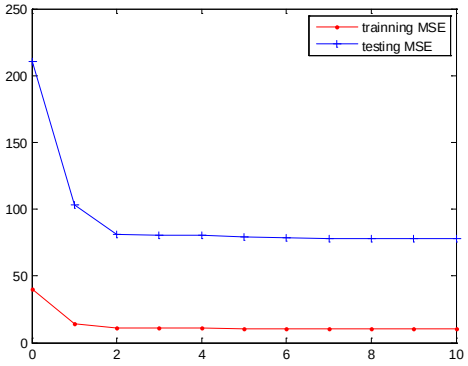




Displacement seems to have negative correlation with mpg. The second order and ninth order performed well.

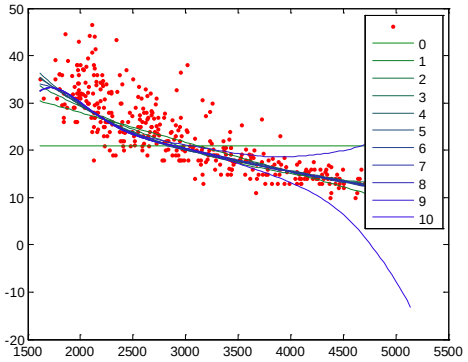
Horsepower:

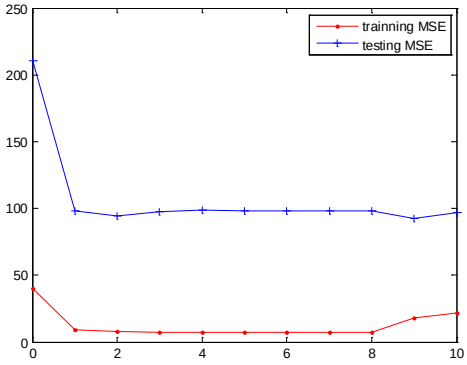




Similar to displacement, horsepower has a negative correlation with mpg, which is likely to be non-linear. That is perhaps why second order and those above have better performance.

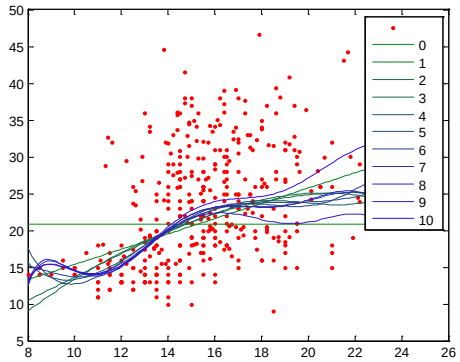
Weight:

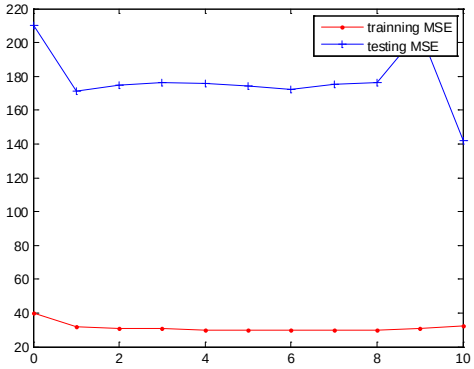




The situation with weight is similar to displacement and horsepower.

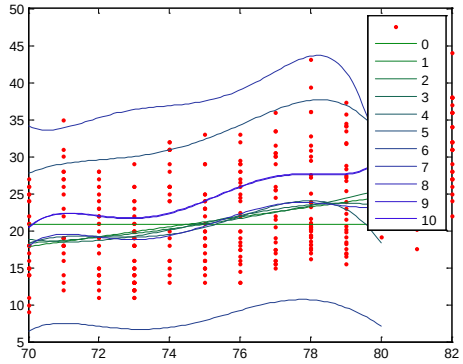
Acceleration:

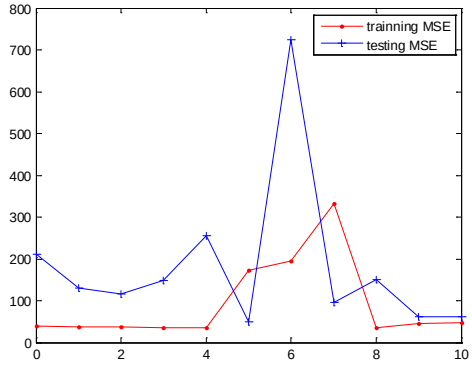




The acceleration show, however not very strong, a positive correlation with mpg. The tenth order has the least MSE, which is probably over-fitting.

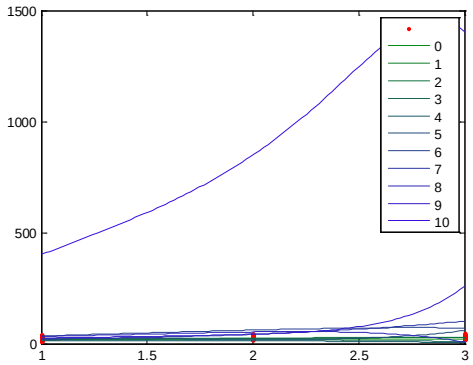
Modelyear:

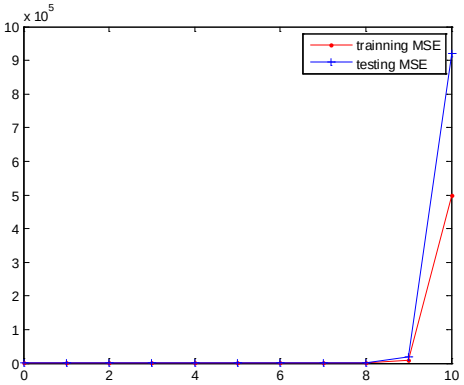




As show in the picture, there is positive correlation between modelyear and mpg. The errors fluctuate a lot. But in my opinion, first or second order will be more reliable than others.

Origin:

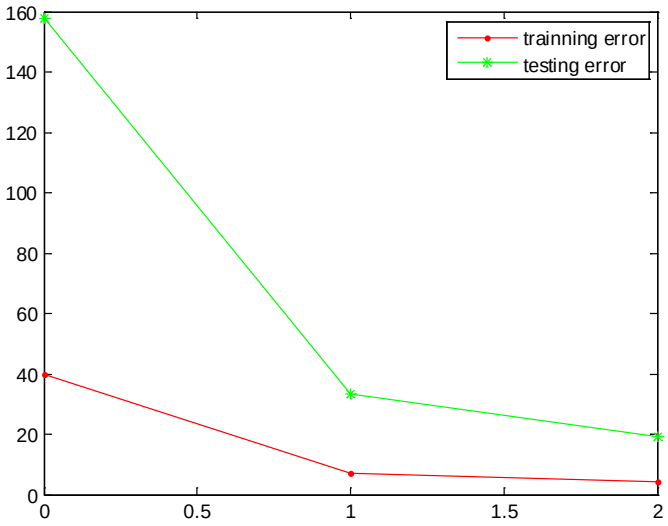




Origin might not be informative regarding mpg.

Q5.

A new solver is modified from “hw1q3.m”, see “hw1q5f.m”.



Q6.

For the logistic regression solver, see file “GDlogit.m”.

We calculate the ratio of misclassification of the test set(see “hw1q6.m”), for “low or not low”, the ratio is 0.0326; for “medium or not medium”, the ratio is 0.8152; for “high or not high”, the ratio is 0.5000.

Q7.

See file “hw1q7.m”.

The predict value of mpg is 21.8524. From this predictor, it should belong to “medium” category. However, by using logistic regression. The predicted value for “low or not low” is close to 1, while other two value is close to zero, which suggests that it should be classified as “low mpg”.

Q8.

The predicted mpg is equal to 0. Since its innate mechanism is totally difference from those vehicle

in the data set, it is impropriate to use any model based on the data to predict its “mpg”.