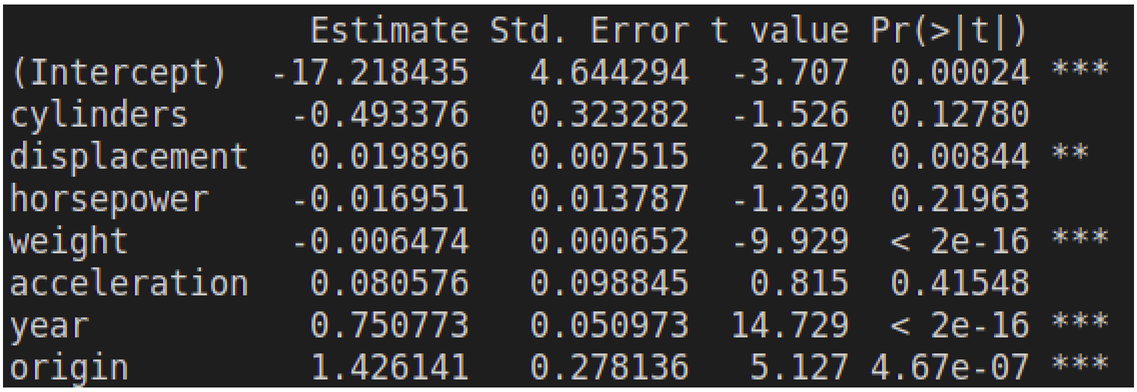
INTRODUCTION TO MACHINE LEARNING (NPFL054)
A template for Homework #2

Name: Lafay Gareth

School year: 2019/20

- **Provide answers for the exercises.**
- **For each exercise, your answer cannot exceed one sheet of paper.**

1.1 Multiple linear regression

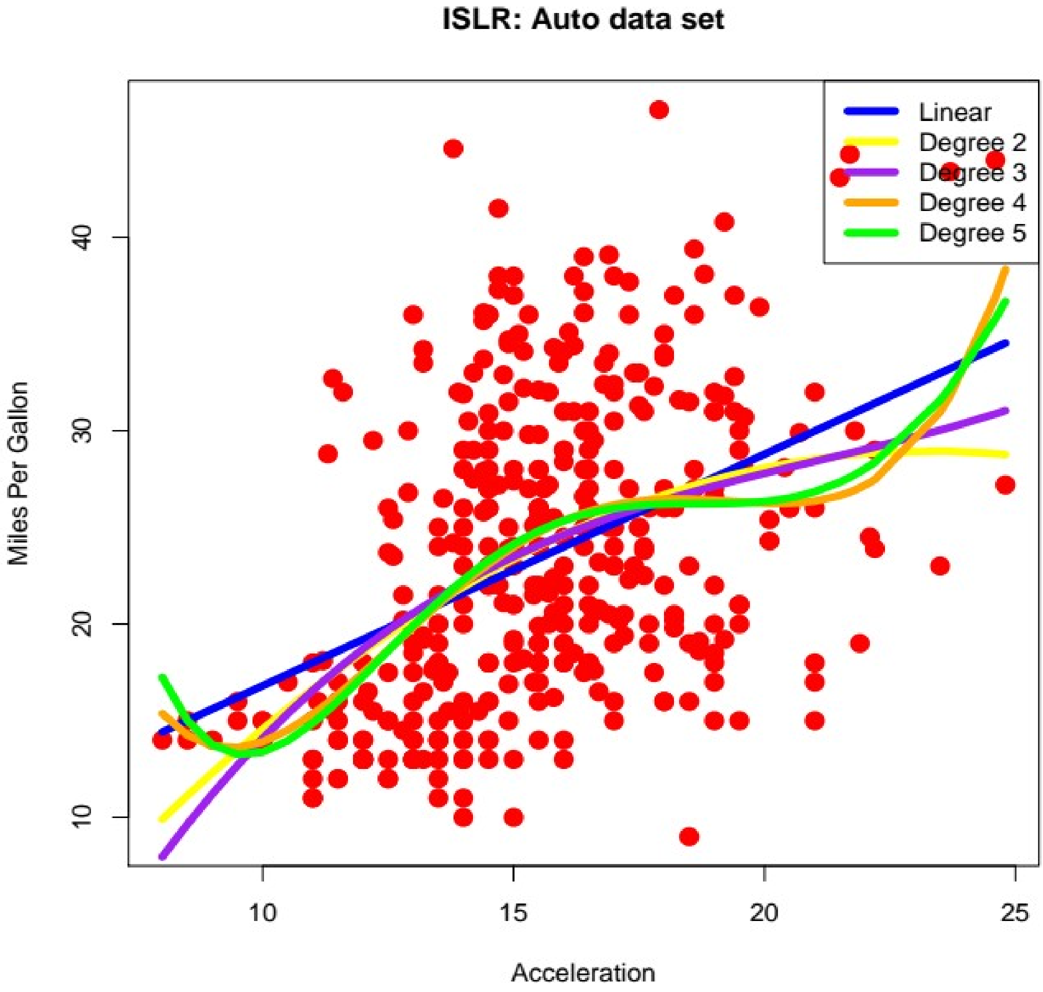


```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780    
displacement   0.019896   0.007515   2.647  0.00844 ** 
horsepower    -0.016951   0.013787  -1.230  0.21963    
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548    
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
```

Interpretation:

The origin, year, acceleration and cylinders are the parameters that have the most impact on the miles per gallon consumption. They are the outlier values for the estimate. The standard deviation is high for cylinders and origin meaning there is a lack of precision for these parameters. The high T value of acceleration, cylinder and horsepower show a strong correlation between them and the miles per gallon consumption.

1.2 Polynomial regression



The associated R^2 values are:

degree 1 : 0.18

degree 2 : 0.19

degree 3 : 0.20

degree 4 : 0.21

degree 5 : 0.21

2.1 Binary attribute `mpg01` and its entropy

---

entropy = 1

The data must be full of low probability values.

2.3 Trivial classifier accuracy

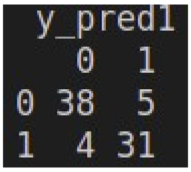
Accuracy = 0.62

This is not very surprising, with all the data it would have been 0.5 but we are only using a randomized portion of it.

2.4 Logistic regression – training and test error rate, confusion matrix, Sensitivity, Specificity, interpretation

a) training error rate: 0.09

b) confusion matrix:



```
  y_pred1
     0  1
0 38  5
1  4 31
```

test error rate: 0.12

sensitivity: 0.89

specificity: 0.90

c) interpretation:

The model is pretty good.

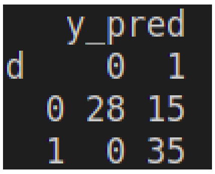
2.5 Logistic regression – threshold 0.1 and 0.9, confusion matrix, Precision, Recall, F1-measure, interpretation

Threshold 0.1:

Precision: 1

Recall: 1

F1-measure: 1



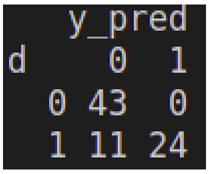
```
   y_pred
d    0  1
  0 28 15
  1  0 35
```

Threshold 0.9:

Precision: 0.69

Recall: 0.67

F1-measure: 0.69

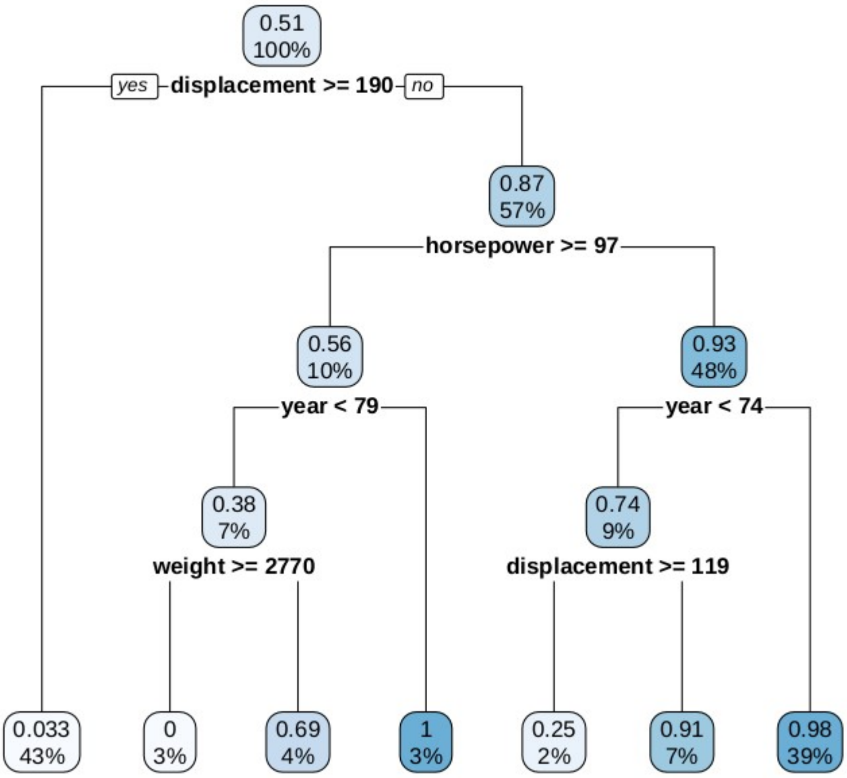


```
   y_pred
d    0  1
  0 43  0
  1 11 24
```

interpretation:

With a threshold of 0.1 we get a perfect model but that might be a problem with my data or me not understanding it, with a threshold of 0.9 our results are close to the trivial classifier, which means it is really bad.

2.6 Decision tree algorithm – training and test error rate, cp parameter



training error rate: 0.95

test error rate: 0.92

cp parameter: The best cp value is 0.01, choosing a lower one does lower the xerror but not by much and a simpler tree is usually better. And the decrease does not seem to be much higher than the standard deviation.