DAR/IDAR Coursework 1

1. Statistical learning methods **(**12% | 12%**)** For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The number of predictors p is extremely large, and the number of observations n is small.

An inflexible method is better because an overly complex model with many parameters compared to the number of observations can overfit and therefore not generalise. The variance of an inflexible model on with a large p and small s will be large.

(b) The sample size n is extremely large, and the number of predictors p is small.

A flexible model is better. The large sample size reduces risk of overfitting and the bias is reduced.

(c) The relationship between the predictors and response is highly non-linear.

A flexible model is better otherwise the non-linearity will not be modeled. An inflexible model will oversimplify and not detect the complicated relationship between the predictors and response.

(d) The standard deviation of the error terms, i.e. $\sigma = \text{sd}(\varepsilon)$, is extremely high.

An inflexible model is better. A flexible model will overfit, and try to model the noise of the high variance of the errors.

2. Bayes' rule **(**12% | 12%**)**

Marking scheme: - MSc: 1% for each probability. - BSc: 1% for each probability.

Given a dataset including 20 observations (S_1, ..., S_20) about the temperature (i.e. hot or cool) for playing golf (i.e. yes or no), you are required to use the Bayes' rule to calculate by hand the probability of playing golf according to the temperature, i.e. P(Play Golf | Temperature).

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_10
Temperature	cool	hot	hot	hot	cool	cool	hot	cool	hot	hot
Play Golf	yes	no	yes	no	yes	yes	no	yes	no	yes

	S_11	S_12	S_13	S_14	S_15	S_16	S_17	S_18	S_19	S_20
Temperature	hot	hot	hot	cool	hot	hot	cool	cool	cool	hot

	S_11	S_12	S_13	S_14	S_15	S_16	S_17	S_18	S_19	S_20
Play Golf	no	no	yes	no	no	no	yes	no	no	no

We can calculate the following probabilities from the dataset:

P(Play Golf = yes) = 8/20 = 0.4

P(Play Golf = no) = 12/20 = 0.6

P(Temperature = hot) = 12/20 = 0.6

P(Temperature = cool) = 8/20 = 0.4

P(Temperature = hot | Play Golf = yes) = 3/8 = 0.375

P(Temperature = cool | Play Golf = yes) = 5/8 = 0.625

P(Temperature = hot | Play Golf = no) = 9/12 = 0.75

P(Temperature = cool | Play Golf = no) = 3/12 = 0.25

Now, using Bayes' rule, we can calculate:

P(Play Golf = yes | Temperature = hot) = (0.375 * 0.4) / 0.6 = 0.25

P(Play Golf = yes | Temperature = cool) = (0.625 * 0.4) / 0.4 = 0.625

P(Play Golf = no | Temperature = hot) = (0.75 * 0.6) / 0.6 = 0.75

P(Play Golf = no | Temperature = cool) = (0.25 * 0.6) / 0.4 = 0.375

3. Descriptive analysis **(**22% | 22%**)** This question involves the `Auto` dataset included in the "ISLR" package.

(a) Which of the predictors are quantitative, and which are qualitative?

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                       name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

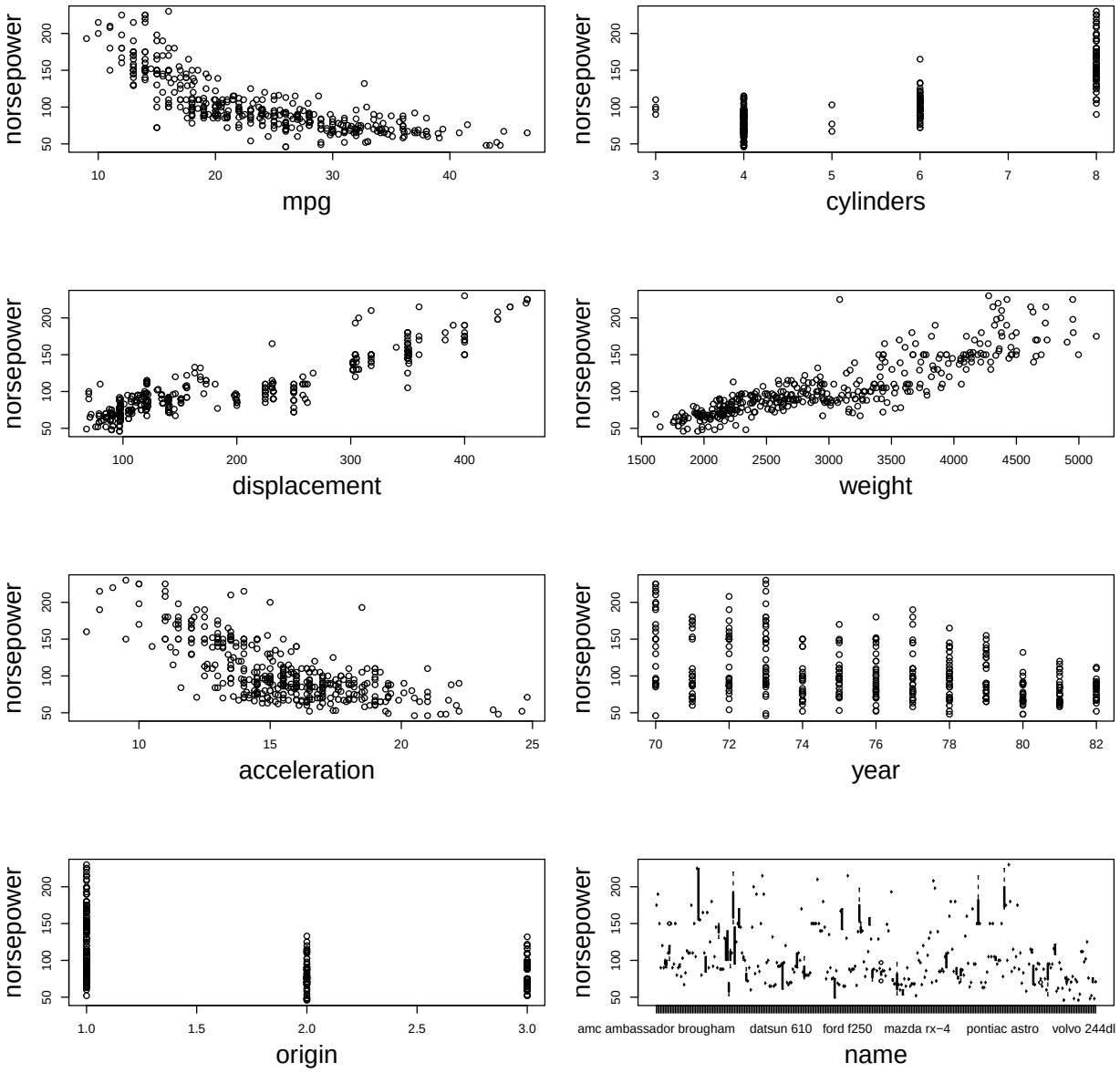
Quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration, year

Qualitative: origin, name

(b) What are the range, median and variance of each quantitative predictor?

```
##                 range    median  variance
## mpg             37.60     22.75     60.92
## cylinders        5.00      4.00      2.91
## displacement   387.00    151.00  10950.37
## horsepower     184.00     93.50   1481.57
## weight        3527.00   2803.50 721484.71
## acceleration    16.80     15.50      7.61
## year            12.00     76.00     13.57
```

(c) Using the full data set, investigate the relationship between individual predictors with the target response engine horsepower (`horsepower`) graphically. Comment on your findings.



horsepower vs. mpg Negative relationship. As mpg increases, horspower approachesb 0 (possibly exponential decrease)

horsepower vs. cylinders No clear relationship. More cylinders does not necessarily mean more horsepower. However, the median horsepower of cars with 8 cylinders is larger than 3 to 6.

horsepower vs. displacment Positive, linear looking relationship. As displacement increases, horsepower also increases.

horsepower vs. weight Positive overall relationship. However the variance in horsepower does increase with weight.

horsepower vs. acceleration Negative relationship. Horsepower shows a decreasing trend with increasing acceleration, which tends to low values at large accelerations.

horsepower vs. year The variance in horsepower appears to be reduced in later rather earlier years. The mean horsepower may also have a negative trend with years.

horsepower vs. origin Cars come from location 1 more than any other. Location 1 also has the largest variance in horsepower. The other two locations are quite similar.

horsepower vs. name Other analysis required

(d) Suppose that we wish to predict `horsepower` on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting `horsepower`? Justify your answer.

Yes: mpg, displacement, weight and acceleration seem to be the most useful attributes for predicting horsepower. From (c), these predictors showed a clear correlation with horsepower. Therefore, they might be useful in predicting horsepower.

4. Linear regression (24% | 24%**)** This question involves the use of simple linear regression on the `Auto` dataset.

(a) Use the `lm()` function to perform a simple linear regression with `acceleration` as the response and `horsepower` as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:

```
##
## Call:
## lm(formula = acceleration ~ horsepower, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9947 -1.2913 -0.1748  1.1229  7.6053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.70193    0.29274   70.72   <2e-16 ***
## horsepower  -0.04940    0.00263  -18.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.002 on 390 degrees of freedom
## Multiple R-squared:  0.475,  Adjusted R-squared:  0.4736
## F-statistic: 352.8 on 1 and 390 DF,  p-value: < 2.2e-16
```

i. Is there a relationship between the predictor and the response?

The magnitude of the t-value is large (18.78) and the p-value is small ($<$ 2.2e-16). We can therefore reject the null hypothesis. There is a relationship between the predictor and the response.

ii. How strong is the relationship between the predictor and the response?

The adjusted R-squared is 0.4736. This implies that only about 47.36% of the sample variation in fire `horsepower` is explained by `acceleration`.

iii. Is the relationship between the predictor and the response positive or negative?

The relationship is negative with slope -9.6155 (0.5119).

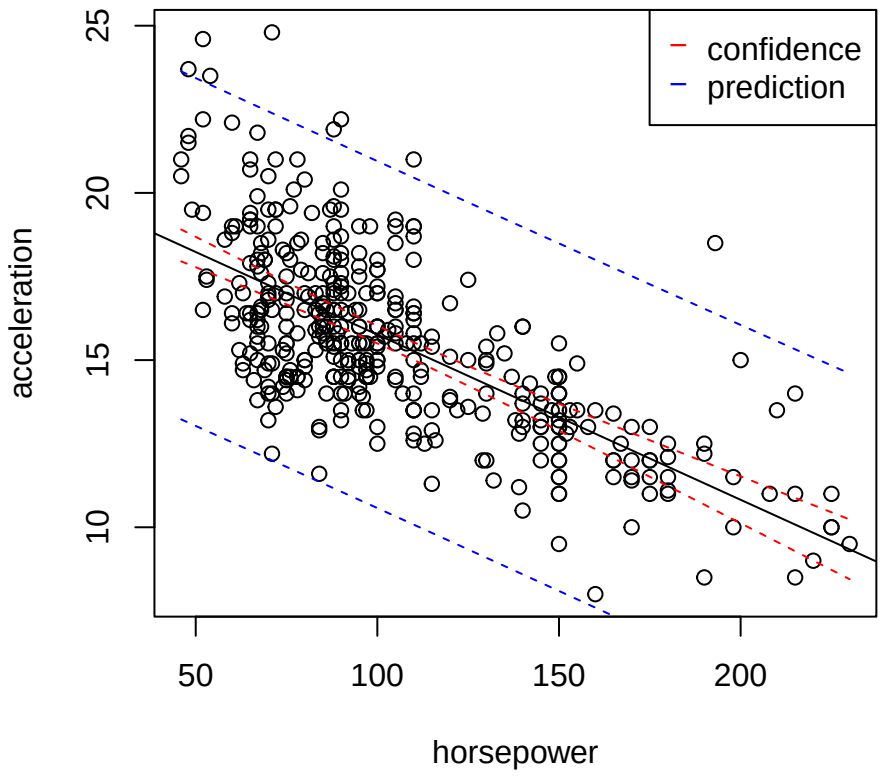
iv. What is the predicted `acceleration` associated with a `horsepower` of 93.5? What are the associated 99% confidence and prediction intervals?

```
##                 fit      lwr      upr
## Confidence 16.08320 15.81107 16.35532
## Prediction 16.08320 10.89501 21.27138
```

The predicted acceleration is 16.0832s. The associated 99% confidence interval is between 15.81107 and 16.35532. The associated 99% prediction interval is between 10.89501 and 21.27138.

(b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

(c) Plot the 99% confidence interval and prediction interval in the same plot as (b) using different colours and legends.



5. Logistic regression and cross validation (30% | 30%)

Marking scheme:
- MSc: (a) 5%, (d) 10%, (e) 15%.
- BSc: (a) 5%, (d) 15%, (e) 10%.

A recent study has shown that the accurate prediction of the office room occupancy leads to potential energy savings of 30%. In this question, you are required to build logistic regression models by using different environmental measurements as predictors (features), such as temperature, humidity, light, CO_2 and humidity ratio, to predict the office room occupancy. The provided training dataset consists of 1,000 observations, whilst the testing dataset consists of 300 observations.

(a) Load the training and testing datasets from corresponding files, and display the statistics about different predictors in the training dataset.

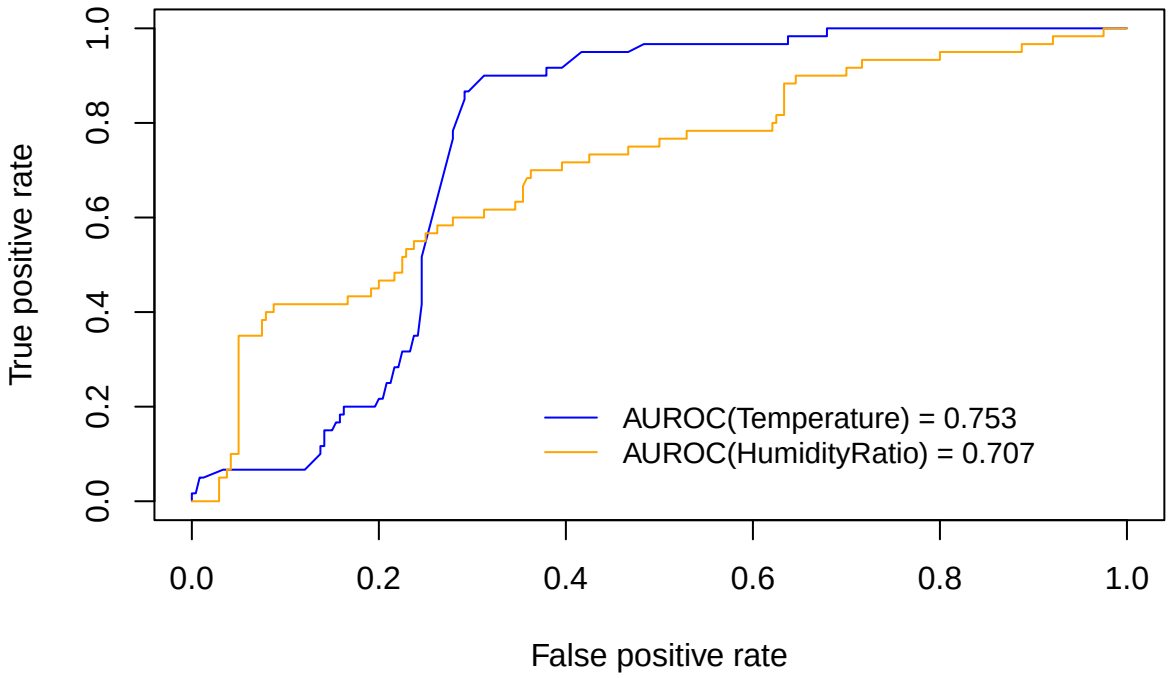
```
##   Temperature       Humidity         Light          CO2
##  Min.   :20.79   Min.   :22.84   Min.   :  0   Min.   : 428.0
##  1st Qu.:21.00   1st Qu.:24.89   1st Qu.:  0   1st Qu.: 447.4
##  Median :21.50   Median :25.86   Median :  0   Median : 498.0
##  Mean   :21.67   Mean   :25.75   Mean   :141   Mean   : 643.9
##  3rd Qu.:22.29   3rd Qu.:26.95   3rd Qu.:438   3rd Qu.: 841.0
##  Max.   :23.18   Max.   :28.50   Max.   :744   Max.   :1139.0
##  HumidityRatio        Occupancy
##  Min.   :0.003462   Min.   :0.000
##  1st Qu.:0.003824   1st Qu.:0.000
##  Median :0.004104   Median :0.000
##  Mean   :0.004134   Mean   :0.242
##  3rd Qu.:0.004472   3rd Qu.:0.000
##  Max.   :0.004793   Max.   :1.000
```

(b) Conduct a 5-fold cross validation to compare the predictive performance of two different predictors, i.e. Humidity and HumidityRatio by using logistic regression method. Report the average accuracies and AUROC values obtained over the 5-fold cross validation. Set the value of random seed as "100" when generating fold indices. Consider the predictive label equals to 1, if the predictive probability is greater than 0.5.

```
## Loading required package: ggplot2

## Loading required package: lattice

## [1] 0.73

## [1] 0.8319416

## [1] 0.784

## [1] 0.8873837
```

HumidityRatio is the best predictor since it gives higher average accuracy and AUROC values from the 5-fold CV.

(c) Predict the class labels of the testing observations by using two different logistic regression models, i.e. one model is trained by using the best predictor obtained in Step (b), whilst the other model is trained by using the Temperature predictor. Compare the performance of those two models by drawing ROC curves and calculating corresponding AUROC values. Discuss the results.



The AUROC value for HumidityRatio is greater than for Temperature. Therefore, the model using HumidityRatio as a predictor is more accurate and better at classifying the Occupancy with a lower error rate.