Homework 2

Jacob Thielemier

7 February 2024

Question 1

4.17 is

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{1}{2\sigma^2}(x-\mu_k)^2)}{\sum_{l=1}^{k} \pi_l \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{1}{2\sigma^2}(x-\mu_l)^2)}$$

and the discriminant function is

$$\delta_k(x) = x.\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma_2} + \log(\pi_k)$$

σ^2 is constant

$$p_k(x) = \frac{\pi_k \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{k} \pi_l \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

Maximizing $p_k(x)$ also maximizes $p_k(X)$, so maximize $\log(p_K(X))$

$$\log(p_k(x)) = \log(\pi_k) - \frac{1}{2\sigma^2}(x-\mu_k)^2 - \log\left(\sum_{l=1}^{k} \pi_l \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)\right)$$

Maximize over k, the last term does not vary with k so it can be ignored. Now maximize

$$\begin{aligned} f &= \log(\pi_k) - \frac{1}{2\sigma^2}(x^2 - 2x\mu_k + \mu_k^2) \\ &= \log(\pi_k) - \frac{x^2}{2\sigma^2} + \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} \end{aligned}$$

$\frac{x^2}{2\sigma^2}$ is independent of k

$$\log(\pi_k) + \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2}$$

Question 2

Same as last question

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma_k}} \exp(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)}{\sum_{l=1}^{k} \pi_l \frac{1}{\sqrt{2\pi\sigma_l}} \exp(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2)}$$

Derive the Bayes classifier, without assuming $\sigma_1^2 = ... = \sigma_K^2$

Maximizing $p_k(x)$ also maximizes $p_k(X)$, so maximize $\log(p_K(X))$

$$\log(p_k(x)) = \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi\sigma_k}}\right) - \frac{1}{2\sigma_k^2}(x-\mu_k)^2 - \log\left(\sum_{l=1}^{k} \frac{1}{\sqrt{2\pi\sigma_l}} \pi_l \exp\left(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\right)\right)$$

Maximizing over k, and since the last term does not vary with k it can be ignored. So maximize

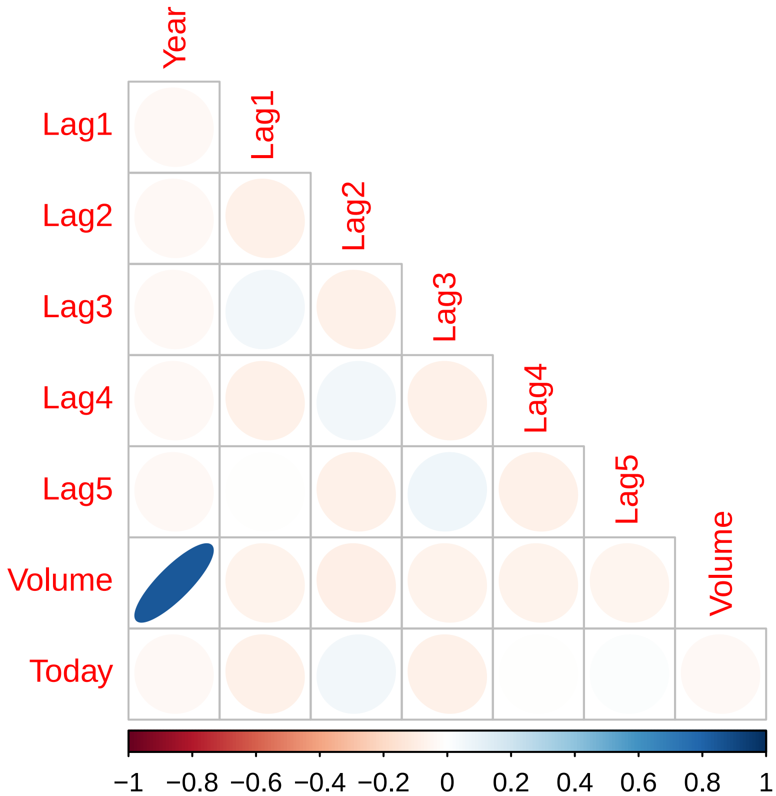
$$\begin{aligned} f &= \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi\sigma_k}}\right) - \frac{1}{2\sigma_k^2}(x-\mu_k)^2 \\ &= \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi\sigma_k}}\right) - \frac{x^2}{2\sigma_k^2} + \frac{x\mu_k}{\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} \end{aligned}$$

Now $\frac{x^2}{2\sigma_k^2}$ is not independent of k, so retain the term with x^2, therefore f, the Bayes' classifier, is a quadratic function of x.

Question 3

(a)

```
##       Year           Lag1               Lag2               Lag3
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260
##       Lag4               Lag5              Volume              Today
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540
##  Median :  0.2380   Median :  0.2340   Median :1.00268   Median :  0.2410
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260
##  Direction
##  Down:484
##  Up  :605
##
##
##
##
```



Volume is strongly positively correlated with Year. Other correlations are week, but Lag1 is negatively correlated with Lag2 but positively correlated with Lag3.

(b)

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
```

```
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

Lag2 is significant.

(c)

```
##      Up
## Down  0
## Up    1
```

```
##
##               Down  Up
##   Down (pred)   54  48
##   Up (pred)    430 557
```

```
## [1] 0.5610652
```

The overall fraction of correct predictions is 0.56. Although logistic regression correctly predicts upwards movements well, it incorrectly predicts most downwards movements as up.

(d)

```
##
##               Down Up
##   Down (pred)    9  5
##   Up (pred)     34 56
```

```
## [1] 0.625
```

(e)

```
##
## pred   Down Up
##   Down    9  5
##   Up     34 56
```

```
## [1] 0.625
```

(f)

```
##
## pred   Down Up
##   Down    0  0
##   Up     43 61
```

```
## [1] 0.5865385
```

(g)

```
##
## fit    Down Up
##   Down   21 30
##   Up     22 31
```

```
## [1] 0.5
```

(h)

```
##
## pred   Down Up
##   Down   27 29
##   Up     16 32
```

```
## [1] 0.5673077
```

(i) Logistic regression and LDA are the best performing.

(j)

```
## [1] 0.5673077
```

```
## [1] 0.5865385
```

```
## [1] 0.5865385
```

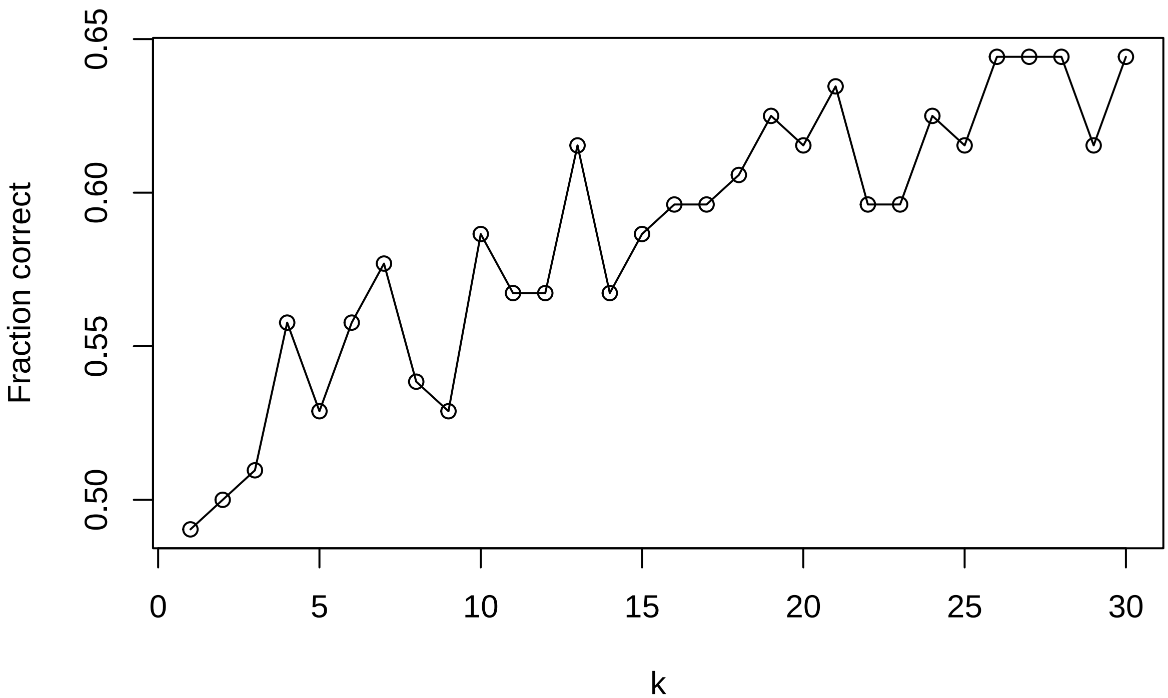
```
## [1] 0.5865385
```

```
## [1] 0.5961538
```

```
## [1] 0.5769231
```

```
## [1] 0.5192308
```

```
## [1] 0.5096154
```




```
## [1] 26
```

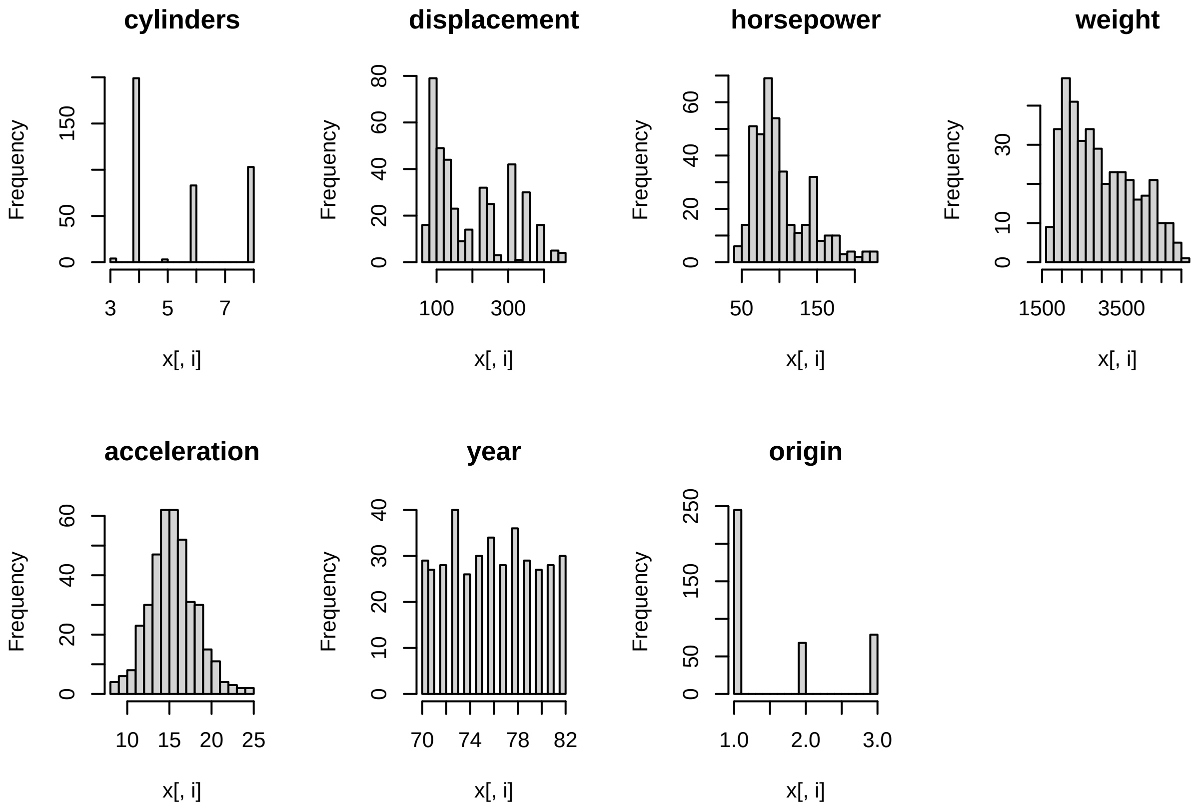
```
##
## fit    Down Up
##   Down   23 18
##   Up     20 43
```

```
## [1] 0.6346154
```

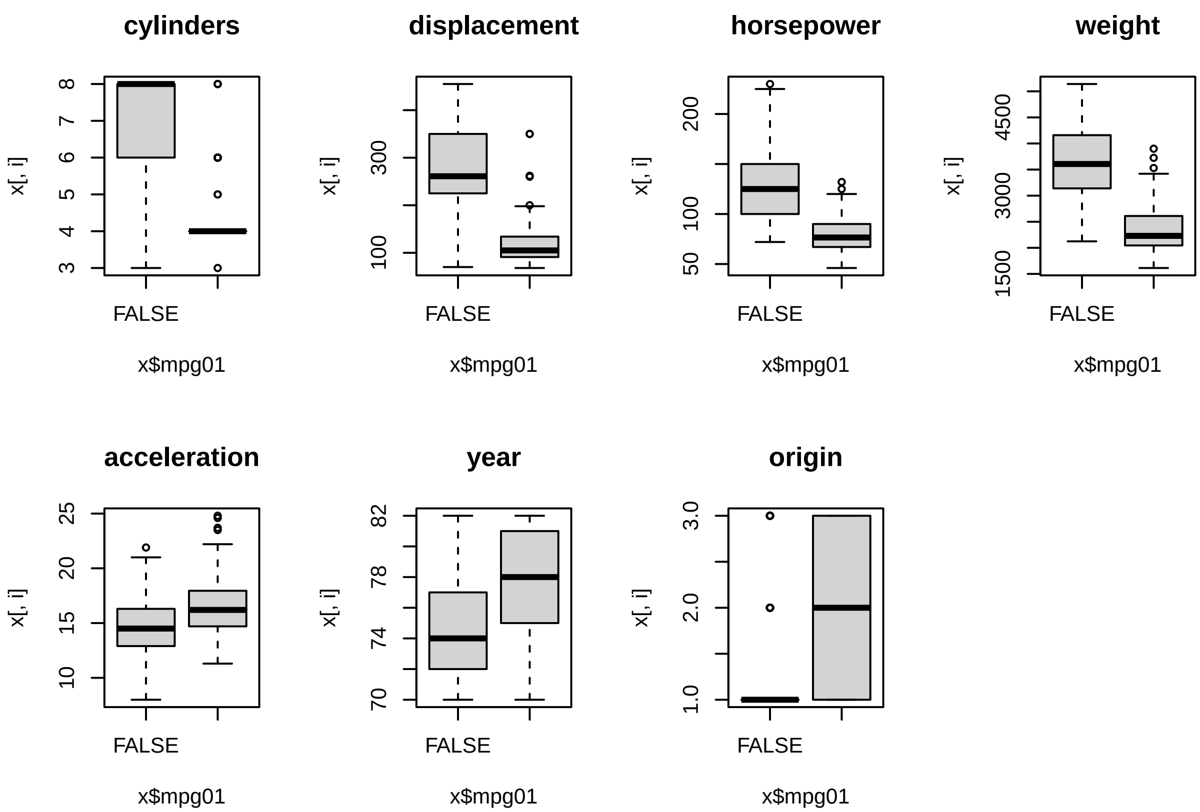
KNN using the first 3 Lag variables performs marginally better than logistic regression with `Lag2` if we tune k to be $k = 26$.

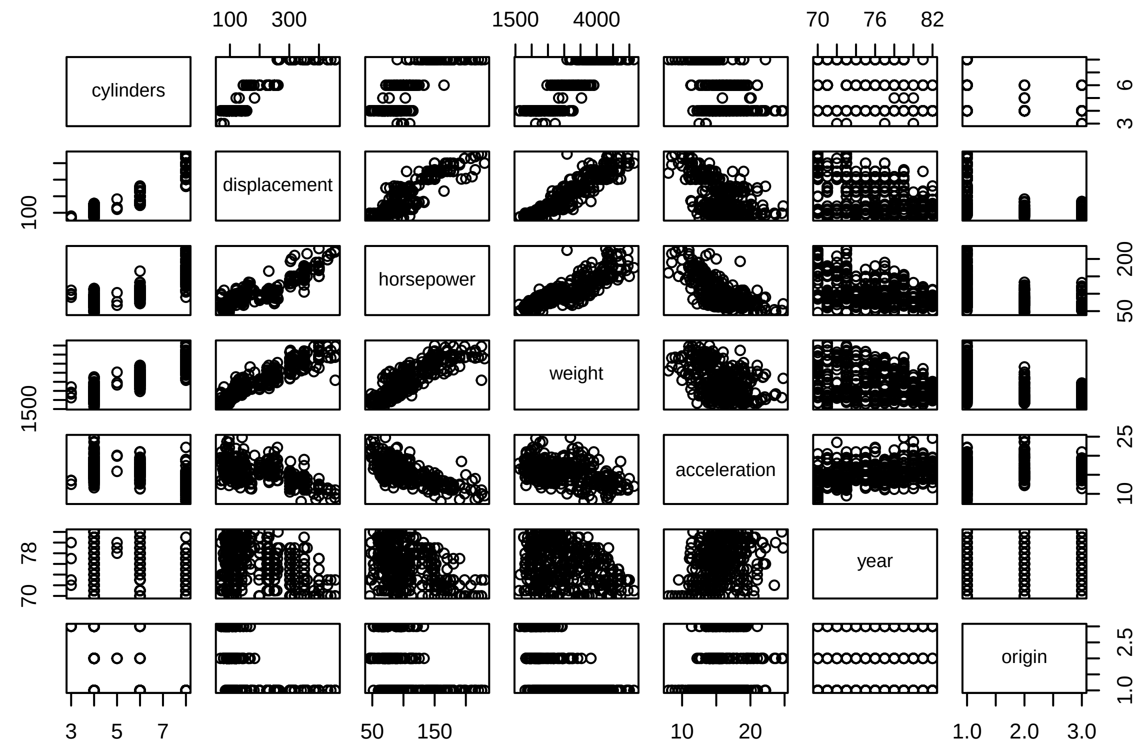
Question 4

(a)


cylinders
displacement
horsepower
weight
acceleration
year
origin
Frequency
x[, i]

(b)


cylinders
displacement
horsepower
weight
acceleration
year
origin
x[, i]
FALSE
x$mpg01



Most variables show an association with `mpg01` category, and several variables are colinear.

(c)

(d)

```
## acceleration         year       origin   horsepower displacement       weight
##     7.302430     9.403221    11.824099    17.681939    22.632004    22.932777
##    cylinders
##    23.035328
```

```
## [1] 0.1068702
```

(e)

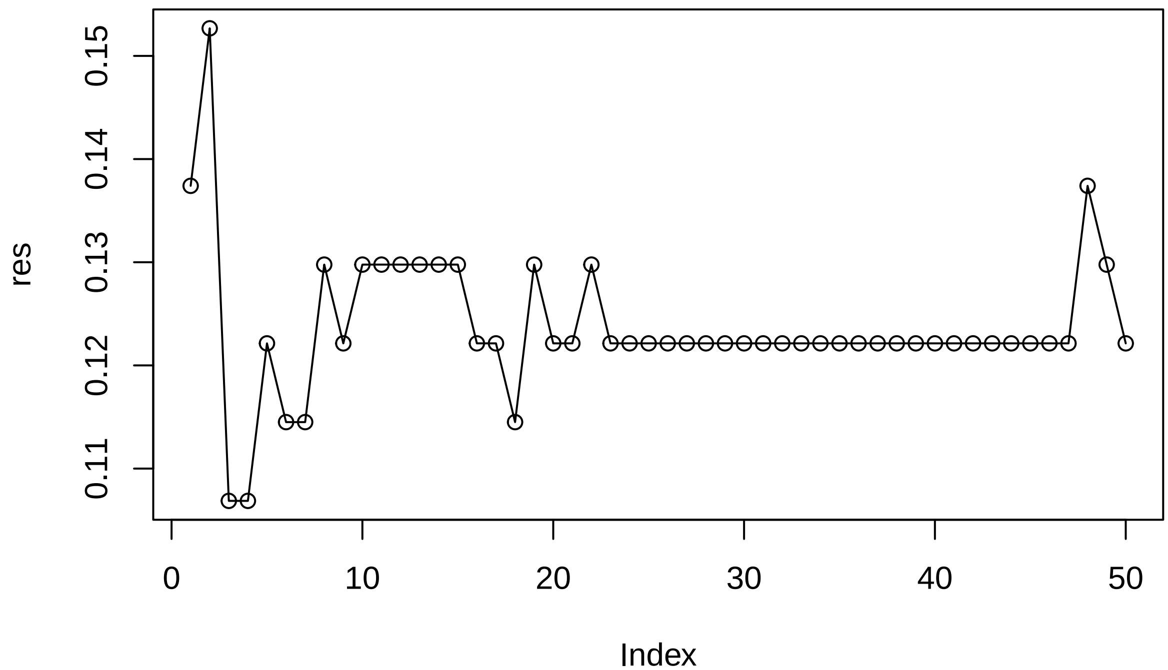
```
## [1] 0.09923664
```

(f)

```
## [1] 0.1145038
```

(g)

```
## [1] 0.09923664
```



(h)

```
##         3
## 0.1068702
```

For the models tested here, $k = 32$ appears to perform best. QDA has a lower error rate overall, performing slightly better than LDA.