University of Michigan
Data Mining
Stats415

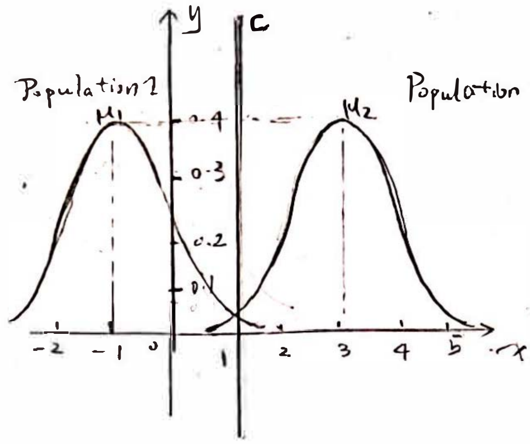
Assignment 4

Author: *Shu* Zhou
ID: 19342932
Lab Section: 001

October 7, 2020

Stats 415 Assignment 3.

1. $\hat{\mu}_1 = -1$, $\hat{\mu}_2 = 3$. $\hat{\sigma}^2 = 1$



(a) Hence $p_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x+1)^2}$

$$p_2(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-3)^2}$$

Since we assume $\pi_1 = \pi_2 = 0.5$

$$\begin{cases} p_1(x) > p_2(x) \text{ when } x < 1. \\ p_1(x) = p_2(x) \text{ when } x = 1 \\ p_1(x) > p_2(x) \text{ when } x > 1 \end{cases}$$

(b) In practice, we do not assume $\pi_1 = \pi_2$.

Assume we still have class mean $\mu_1 = -1$, $\mu_2 = 3$

The Class priors $\begin{cases} \hat{\pi}_1 = \frac{40}{100} = 0.4 \\ \hat{\pi}_2 = \frac{60}{100} = 0.6 \end{cases}$

The discriminant function $\delta_k(x) = x\frac{\mu_k^2}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log\pi_k$ with the term $\log\pi_k$ will be greater for class 2 than class 1.

Hence, the new boundary value $\hat{c}$ will be less than c.

(c) $\begin{cases} \delta_1(x) = -x - \frac{1}{2} + \log 0.4 \\ \delta_2(x) = 3x - \frac{9}{2} + \log 0.6 \end{cases}$

$$\Rightarrow \begin{cases} \delta_1(x) > \delta_2(x) \text{ when } x < 0.8986 \\ \delta_2(x) = \delta_2(x) \text{ when } x = 0.8986 \\ \delta_1(x) < \delta_2(x) \text{ when } x > 0.8986 \end{cases} \Rightarrow \hat{c} = 0.8986 < c$$

(d) I would recommand using QDA, Since Quadratic Discriminant Analysis tends to work better when the variances are very different between classes. Also we have enough observations to accurately estimate the variances.

(e) $\begin{cases} \hat{\mu}_1 = -1; \ \hat{\sigma}_1^2 = 0.25 \\ \hat{\mu}_2 = 3, \ \hat{\sigma}_2^2 = 1.5 \end{cases}$ Also, we assume $\pi_1 = \pi_2 = 0.5$.

Hence $\delta_1(x) = -\frac{1}{2}\log(0.25) - \frac{1}{2}\frac{(x+1)^2}{0.25} + \log 0.5$ $\Rightarrow c = 0.292$

$\delta_2(x) = -\frac{1}{2}\log(1.5) - \frac{1}{2}\frac{(x-3)^2}{1.5} + \log 0.5$

$$\begin{cases} \delta_1(x) > \delta_2(x) \text{ when } x < 0.292 \\ \delta_1(x) = \delta_2(x) \text{ when } x = 0.292 \\ \delta_1(x) < \delta_2(x) \text{ when } x > 0.292 \end{cases}$$

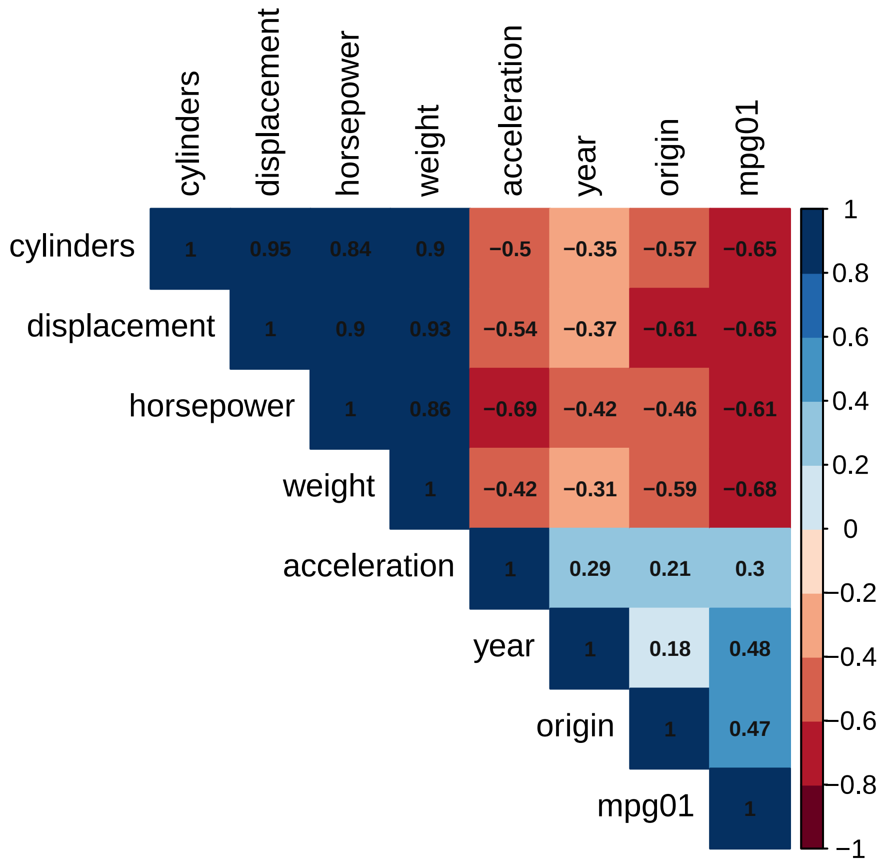
Q2.

(a)

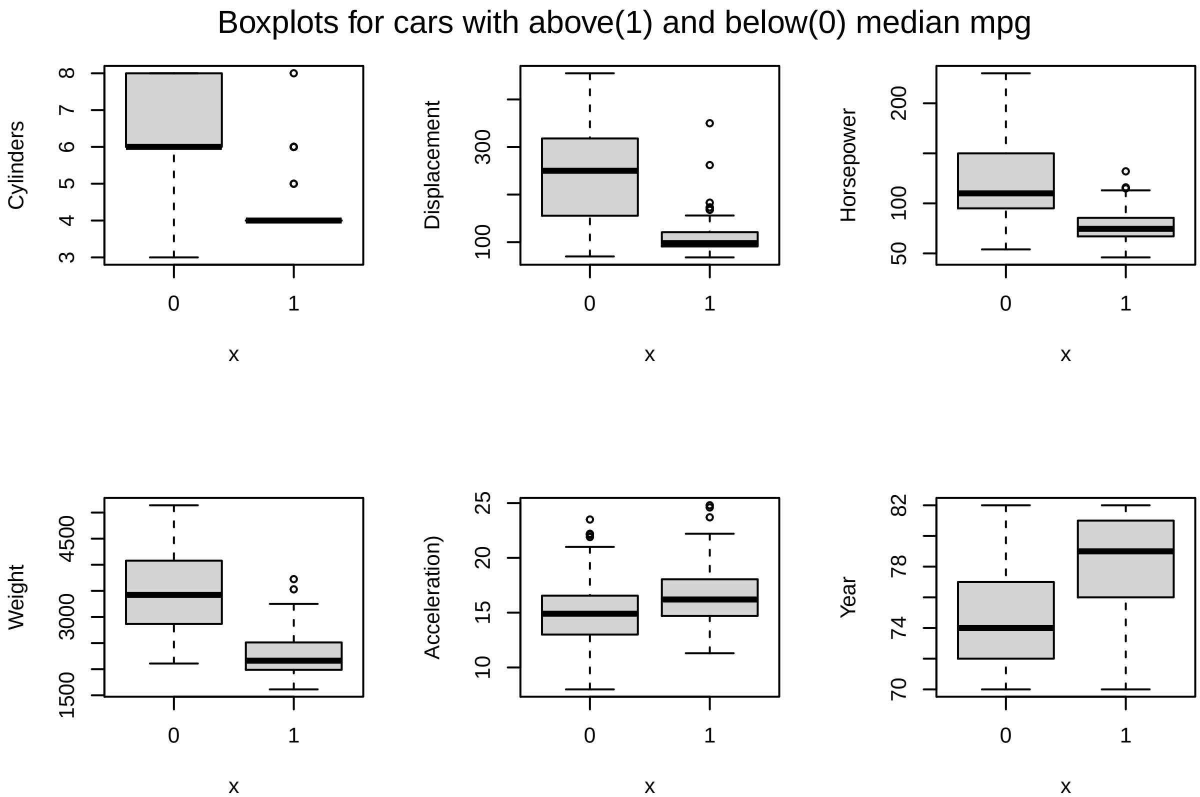
```r
library(ISLR)
data("Auto")
mpg01 <- ifelse(Auto$mpg > 25, 1, 0)
Auto <- data.frame(Auto, mpg01)
```

(b)

```r
corrplot(cor(Auto[,c(2:8,10)]), method="color", type = "upper", col=brewer.pal(n=10, name="RdBu"),
         tl.col="black",tl.srt=90, addCoef.col = "gray8", diag = T, number.cex = 0.65)
```




```r
par(mfrow = c(2, 3))
plot(factor(Auto$mpg01), Auto$cylinders, ylab = "Cylinders")
plot(factor(Auto$mpg01), Auto$displacement, ylab = "Displacement")
plot(factor(Auto$mpg01), Auto$horsepower, ylab = "Horsepower")
plot(factor(Auto$mpg01), Auto$weight, ylab = "Weight")
plot(factor(Auto$mpg01), Auto$acceleration, ylab = "Acceleration)")
plot(factor(Auto$mpg01), Auto$year, ylab = "Year")
mtext("Boxplots for cars with above(1) and below(0) median mpg", outer = TRUE, line = -3)
```



The variables “cylinders”, “displacement”, “horsepower” and “weight” seem to be highly correlated and useful to predict mpg01

(c)

```
set.seed(123)
num_train <- nrow(Auto) * 0.8
inTrain <- sample(nrow(Auto), size = num_train)
training <- Auto[inTrain,]
testing <- Auto[-inTrain,]
```

(d)

```
lda_model <- lda(mpg01 ~ displacement + horsepower + weight + cylinders, data = training)
pred <- predict(lda_model, testing)
table(pred$class, testing$mpg01)
```

```
##
##      0  1
##   0 38  2
##   1 14 25
```

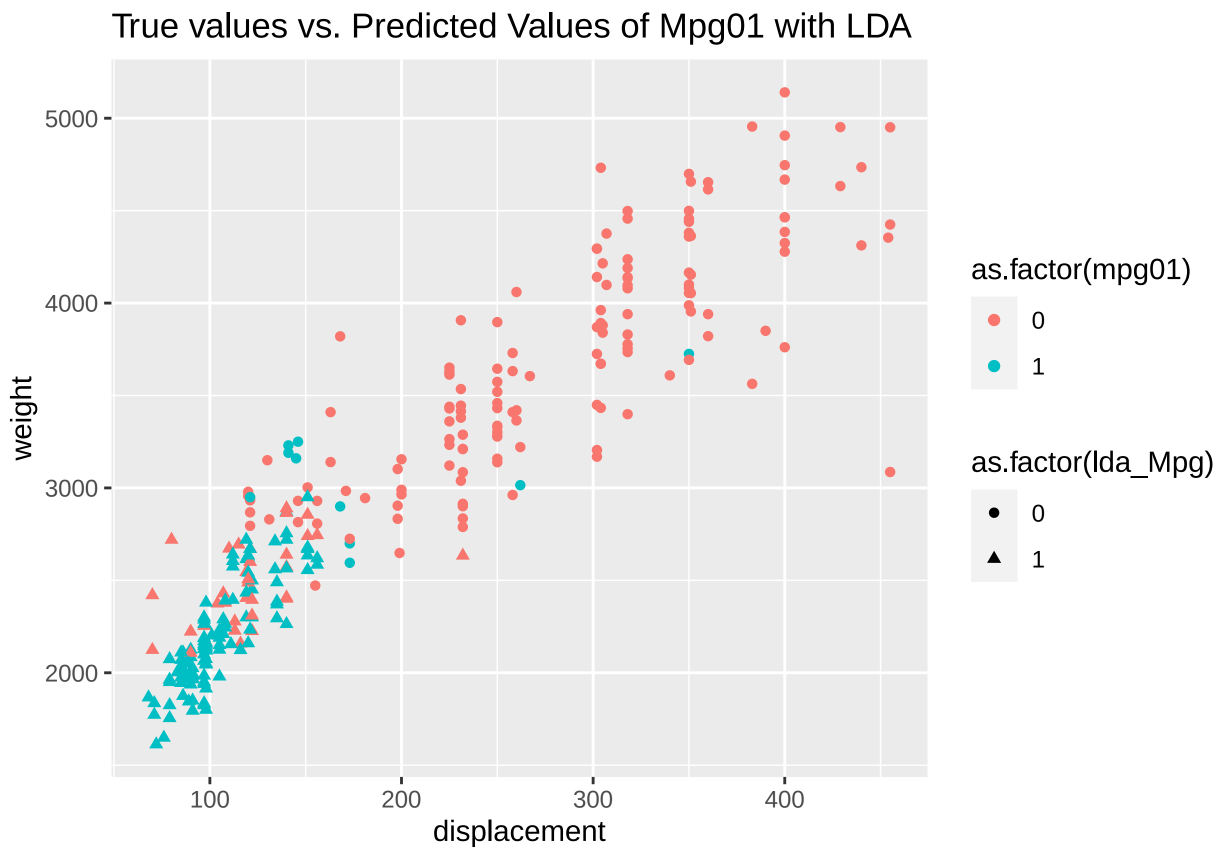
```
1-mean(pred$class == testing$mpg01)
```

```
## [1] 0.2025316
```

The test error is 0.2025316

```
pred_train <- predict(lda_model, training)
training$lda_Mpg <- pred_train$class
```

```r
ggplot(training, aes(x=displacement, y=weight, color = as.factor(mpg01),shape = as.factor(lda_Mpg)))+ge
```



(e)

```r
qda_model <- qda(mpg01 ~ displacement + horsepower + weight + cylinders, data = training)
pred1 <- predict(qda_model, testing)
table(pred1$class, testing$mpg01)
```

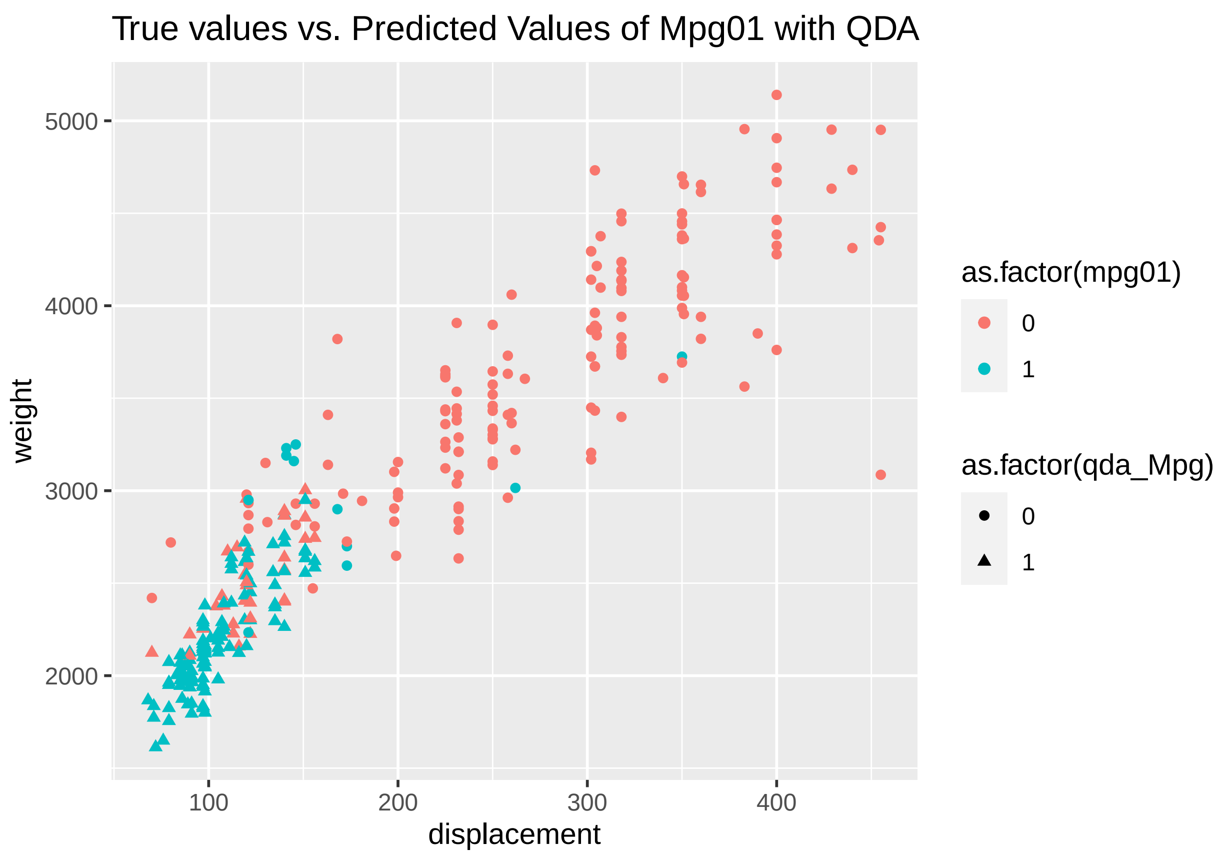
```
##
##      0  1
##   0 41  2
##   1 11 25
```

```r
1-mean(pred1$class == testing$mpg01)
```

```
## [1] 0.164557
```

The test error is 0.164557

```r
pred1_train <- predict(qda_model, training)
training$qda_Mpg <- pred1_train$class
ggplot(training, aes(x=displacement, y=weight, color = as.factor(mpg01),shape = as.factor(qda_Mpg)))+ge
```



(f) The performance of the QDA is better than the performance of LDA in this test. Since we have enough observations to accurately estimate the variances and we have known that the variances are very different between classes, the QDA would perform better as it would take the class-specific covariances into consideration.