Homework 3

Connor Johnson

```r
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.0.3
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.4

## Registered S3 methods overwritten by 'tibble':
##   method     from
##   format.tbl pillar
##   print.tbl  pillar
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.4

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(class)
```

Problem 2

(a)

i. In the range [0.05,0.95], 10 percent of the observations are used. In the range [0, 0.05], 100x+5 percent of the observations are used. In the range [0.95, 1], 105-100x percent of the observations are used. So the average percentage used is

$$\int_{0.05}^{0.95} 10dx + \int_{0}^{0.05} 100x + 5dx + \int_{0.95}^{1} 105 - 100xdx = 9.75$$

ii. The average fraction for one feature is 9.75%. So for two, it is $0.0975^2 = 0.00950625$ or .950625%.

iii. $0.0975^{100} \approx 0\%$

iv. The percentage of observations used in a prediction is an exponential function of p. Because of this, the number of test observations needed to make good predictions also increases exponentially as p increases. This is a major drawback because it's pretty much impossible to have enough test observations when p is large.

v.

$$p = 1, l = .1$$
$$p = 2, l = \sqrt{.1} \approx 0.32$$
$$p = 100, l = 0.1^{1/100} \approx 0.98$$

(b) We would like to know what n is. We saw in a that non-paramteric approaches cover a small amount of the data when p is large. This is a problem if the sample size is too small to account for this. With a larger smaple size, a large p would still perform well because the extra data would allow the model to still be accurate.

(c) The point that ISLR Ch. 4 Exercise 4 is trying to illustrate, is that in high dimensional space, you are often forced to overfit.

(d) More data is always a good thing if the added data is just increasing the size of your sample size. A larger sample size always improves the quality of a model as long as the data is taken correctly. More data could not improve the model when adding features. As shown in (a), adding features requires more observations to maintain the predictability of the model. Adding features that may not even be good predictors could greatly reduce the predictability of our model.

Problem 3

Exercise 5

(a) QDA is expected to perform better on the training set and LDA is expected to perform better on the test set because QDA is likely to overfit the model.

(b) QDA is expected to perform better on both the training and test set.

(c) As n increases, we expect the accuracy of QDA to improve relative to LDA. QDA adds more flexibility to the model which gives better results and the potential added variance ca be offset by a larger sample size.

(d) False. QDA is likely to overfit the model with a small sample size which results in more error in the test cases.

Exercise 8

When K=1, the error on the training set is 0%. Since the training and test sets are the same size and the overall error is 18%, the error on the test set for kNN must be 36%. So, the logistic regression has a lower error on the test. Because of this, we prefer to use the logistic regression for classification of new variables.

Problem 4

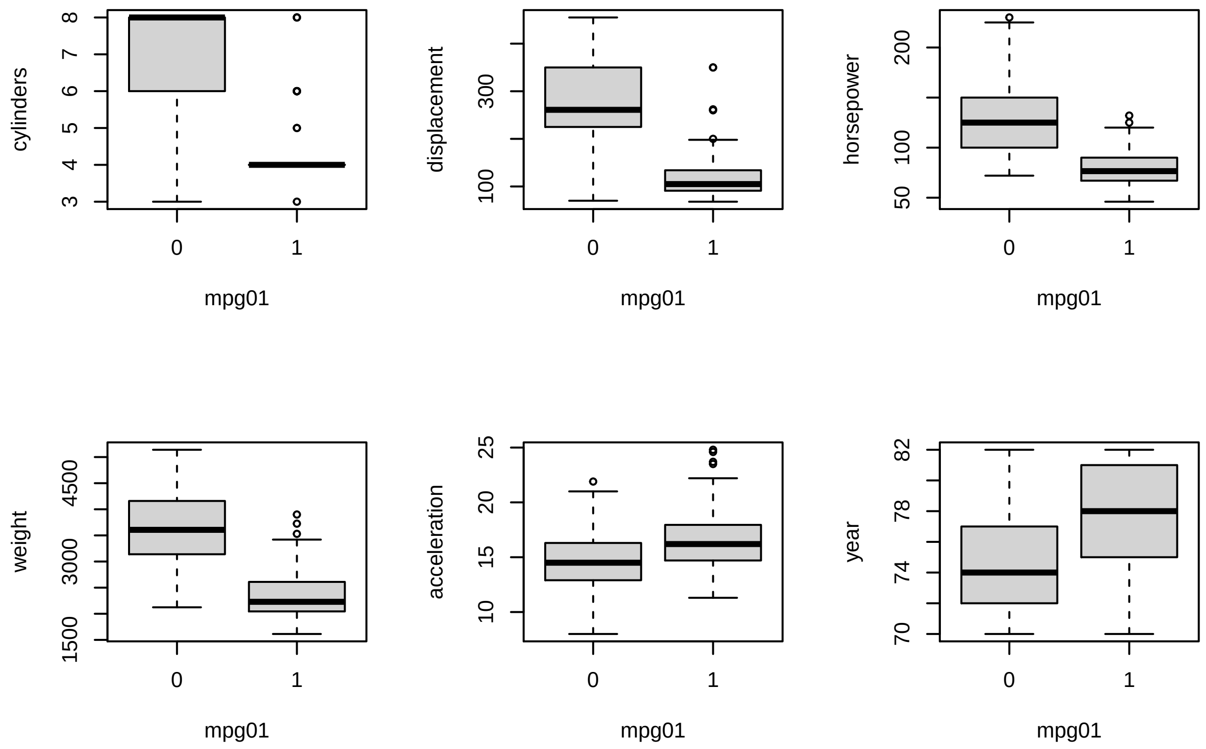
Exercise 11

(a)

```
mpg01 = rep(0, nrow(Auto))
mpg01[Auto$mpg > median(Auto$mpg)] = 1
Auto1 = data.frame(Auto, mpg01)
```

(b)

```
par(mfrow=c(2,3))
boxplot(cylinders~mpg01, data=Auto1)
boxplot(displacement~mpg01, data=Auto1)
boxplot(horsepower~mpg01, data=Auto1)
boxplot(weight~mpg01, data=Auto1)
boxplot(acceleration~mpg01, data=Auto1)
boxplot(year~mpg01, data=Auto1)
```



Cylinders, displacement, horsepower, weight, acceleration, and year all have a correlation with mpg01. Cylinders, displacement, horsepower, and weight have lower medians when mpg is above the median. Acceleration and year have higher medians when mpg is above the median.

(c)

```
set.seed(1)
train = sample.int(n = nrow(Auto1), size = floor(.75*nrow(Auto1)), replace=F)
Auto1.train = Auto1[train,]
Auto1.test = Auto1[-train,]
```

(d)

```
lda.fit = lda(mpg01 ~ cylinders + weight + displacement + horsepower, data = Auto1.train)
lda.pred = predict(lda.fit, Auto1.test)
mean(lda.pred$class != Auto1.test$mpg01)
```

```
## [1] 0.1326531
```

The test error for LDA is 13.3%

(e)

```
qda.fit = qda(mpg01 ~ cylinders + weight + displacement + horsepower, data = Auto1.train)
qda.pred = predict(qda.fit, Auto1.test)
mean(qda.pred$class != Auto1.test$mpg01)
```

```
## [1] 0.122449
```

The test error for QDA is 12.2%

(f)

```
log.fit = glm(mpg01 ~ cylinders + weight + displacement + horsepower, data=Auto1.train)
summary(log.fit)
```

```
##
## Call:
## glm(formula = mpg01 ~ cylinders + weight + displacement + horsepower,
##     data = Auto1.train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.03117  -0.10806   0.08459   0.16257   1.05061
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.730e+00  1.284e-01  13.473  < 2e-16 ***
## cylinders    -9.726e-02  3.598e-02  -2.704 0.007266 **
## weight       -2.104e-04  6.123e-05  -3.436 0.000677 ***
## displacement -9.391e-04  7.887e-04  -1.191 0.234712
## horsepower    1.177e-03  1.116e-03   1.055 0.292436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.09914733)
##
##     Null deviance: 73.446  on 293  degrees of freedom
## Residual deviance: 28.654  on 289  degrees of freedom
## AIC: 161.82
##
## Number of Fisher Scoring iterations: 2
```

```
log.pred = predict(log.fit, Auto1.test, type = "response")
mpg.pred = numeric(length(log.pred))
mpg.pred = log.pred > 0.5
mean(mpg.pred != Auto1.test$mpg01)
```

```
## [1] 0.1326531
```

The test error for the logistic model is 13.3%

```
min = 2
min_k = -1
for (i in 1:100){
  knn.pred = knn(Auto1.train[,2:5], Auto1.test[,2:5], Auto1.train$mpg01, k=i)
  mu = mean(knn.pred != Auto1.test$mpg01)
  if (mu < min){
    min = mu
    min_k = i
  }

}
print(min)
```

```
## [1] 0.1122449
```

```
print(min_k)
```

```
## [1] 4
```

The best k value is k=4 with an error of 11.2%

Exercise 5

(a) The plot shows a bias toward men in student admissions. A higher percentage of men get admitted compared to women. 44.5% of men were accepted while 30.4% of women were accepted.

(b) The plots show that there isn't a bias when admission rates are looked at by department.

(c) The paradox is that the admission rates appear to be biased towards men when the data is all put together. When the data is split into departments, this bias disappears. Using data this generally without considering certain aspects of it can cause results don't accurately represent what is occurring.

(d) The data shows that more women apply to departments that have lower admission rates. This is confounding with the overall admission rates because women would obviously have a lower admission rate if they apply to more difficult departments to get into. Sorting by department makes the bias disappear because the bias didn't actually exist; it was the result of misusing the data that they had.

(e)

```
data(UCBAdmissions)
Adm <- as.integer(UCBAdmissions)[(1:(6*2))*2-1]
Rej <- as.integer(UCBAdmissions)[(1:(6*2))*2]
Dept <- gl(6,2,6*2,labels=c("A","B","C","D","E","F"))
Sex <- gl(2,1,6*2,labels=c("Male","Female"))
Ratio <- Adm/(Rej+Adm)
berk <- data.frame(Adm,Rej,Sex,Dept,Ratio)

LogReg.gender <- glm(cbind(Adm,Rej)~Sex,data=berk,family=binomial("logit"))
summary(LogReg.gender)
```

```
##
## Call:
## glm(formula = cbind(Adm, Rej) ~ Sex, family = binomial("logit"),
##     data = berk)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -16.7915   -4.7613   -0.4365    5.1025   11.2022
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.22013    0.03879  -5.675 1.38e-08 ***
## SexFemale   -0.61035    0.06389  -9.553  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 877.06  on 11  degrees of freedom
## Residual deviance: 783.61  on 10  degrees of freedom
## AIC: 856.55
##
## Number of Fisher Scoring iterations: 4
```

There is a negative coefficient for the female variable. The p-value is also very lowe so it is statistically signinifcant. This shows there is a negative relationship between being female and admission rates.

(f)

```
LogReg.gender2 <- glm(cbind(Adm,Rej)~Sex+Dept,data=berk,family=binomial("logit"))
summary(LogReg.gender2)
```

```
##
## Call:
## glm(formula = cbind(Adm, Rej) ~ Sex + Dept, family = binomial("logit"),
##     data = berk)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
## -1.2487   3.7189  -0.0560   0.2706   1.2533  -0.9243   0.0826  -0.0858
##       9       10       11       12
##  1.2205  -0.8509  -0.2076   0.2052
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.58205    0.06899   8.436   <2e-16 ***
## SexFemale    0.09987    0.08085   1.235    0.217
## DeptB       -0.04340    0.10984  -0.395    0.693
## DeptC       -1.26260    0.10663 -11.841   <2e-16 ***
## DeptD       -1.29461    0.10582 -12.234   <2e-16 ***
## DeptE       -1.73931    0.12611 -13.792   <2e-16 ***
## DeptF       -3.30648    0.16998 -19.452   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 877.056  on 11  degrees of freedom
## Residual deviance:  20.204  on  5  degrees of freedom
## AIC: 103.14
##
## Number of Fisher Scoring iterations: 4
```

When the department is added, the gender coefficient is positive and much smaller. The p-value is also 0.217 which shows it isn't significantly different from zero. The departments, however, are good indicators of admissions with very small p-values. We have shown that confounding variables can lead to results that are misleading. When the department wasn't considered, the data showed a clear significant bias against women. It is important to consider all possible variables that might affect the response variable. Otherwise, some unknown data can confound with the variables used which can lead to statistically significant results that are very misleading.