group assignment

Student ID: 474084,476397,474869,457942,473928

```
#11.(e)
Auto=read.csv("/Users/chloe/Desktop/R/Auto.csv",header=T, na.string="?")
Auto=na.omit(Auto)
attach(Auto)

median = median(mpg)
mpg01 = rep(1, length(mpg))
mpg01[mpg < median] = 0
Auto = data.frame(cbind(mpg01, Auto))

set.seed(100)
train = sample(nrow(Auto), nrow(Auto)/2, replace = FALSE)
train.x = Auto[train,]
test.x = Auto[-train,]

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

```
## [1] 0.1173469
```

We use weight,displacement,cylinders and horsepower as predictores with training data in the qda model. The error rate is 0.1173.

```
#5(a)
library(ISLR)
```

```
##
## Attaching package: 'ISLR'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     Auto
```

```
detach(Auto)
attach(Default)
class(default)
```

```
## [1] "factor"
```

```
glm.fit=glm(default~income+balance,family=binomial)
```

We use income and balance as predictors in the logistic model.

```
#5(b)
set.seed(1)

train=sample(nrow(Default),nrow(Default)/2,replace=FALSE)
train.x=Default[train,]
test.x=Default[-train,]
contrasts(default)
```

```
##     Yes
## No    0
## Yes   1
```

```
glm.fit2=glm(default~income+balance,family=binomial,data=train.x)
default.prob=predict(glm.fit2,test.x,type="response")
default.pred=rep("Yes",nrow(test.x))
default.pred[default.prob<0.5]="No"

mean(default.pred!=test.x$default)
```

```
## [1] 0.0254
```

We use validation set approach to separate the whole data as training data and a validation set. And then use logistic regression with predictors of income and balance. The error rate is 0.0254.

```
#5(c)_1
set.seed(40)

train=sample(nrow(Default),nrow(Default)/2,replace=FALSE)
train.x=Default[train,]
test.x=Default[-train,]
contrasts(default)
```

```
##     Yes
## No    0
## Yes   1
```

```
glm.fit2=glm(default~income+balance,family=binomial,data=train.x)
default.prob=predict(glm.fit2,test.x,type="response")
default.pred=rep("Yes",nrow(test.x))
default.pred[default.prob<0.5]="No"
mean(default.pred!=test.x$default)
```

```
## [1] 0.026
```

We change the seed and repeat the regression process. And the error rate, when the seed is equal to 40, is 0.026.

```
#5(c)_2
set.seed(66)

train=sample(nrow(Default),nrow(Default)/2,replace=FALSE)
train.x=Default[train,]
test.x=Default[-train,]
contrasts(default)
```

```
##     Yes
## No    0
## Yes   1
```

```
glm.fit2=glm(default~income+balance,family=binomial,data=train.x)
default.prob=predict(glm.fit2,test.x,type="response")
default.pred=rep("Yes",nrow(test.x))
default.pred[default.prob<0.5]="No"
mean(default.pred!=test.x$default)
```

```
## [1] 0.0276
```

We change the seed and repeat the regression process. And the error rate, when the seed is equal to 66, is 0.0276.

```
#5(c)_3
set.seed(88)

train=sample(nrow(Default),nrow(Default)/2,replace=FALSE)
train.x=Default[train,]
test.x=Default[-train,]
contrasts(default)
```

```
##     Yes
## No    0
## Yes   1
```

```
glm.fit2=glm(default~income+balance,family=binomial,data=train.x)
default.prob=predict(glm.fit2,test.x,type="response")
default.pred=rep("Yes",nrow(test.x))
default.pred[default.prob<0.5]="No"
mean(default.pred!=test.x$default)
```

```
## [1] 0.0264
```

We change the seed and repeat the regression process. And the error rate, when the seed is equal to 88, is 0.0264. With the different seed, we will get totally different data in the traning and test data set and then the validation error is changing as well.

```
#5(d)
set.seed(1)
train=sample(nrow(Default),nrow(Default)/2,replace=FALSE)
train.x=Default[train,]
test.x=Default[-train,]
contrasts(default)
```

```
##     Yes
## No    0
## Yes   1
```

```
glm.fit3=glm(default~income+balance+student,family=binomial,data=train.x)
default.prob=predict(glm.fit3,test.x,type="response")
default.pred=rep("Yes",nrow(test.x))
default.pred[default.prob<0.5]="No"
mean(default.pred!=test.x$default)
```

```
## [1] 0.026
```

According to the results, although the validation error decreases a little, there seems no significant improvement of error rate when we add dummy variable of student in the logistic model.