Shuting Chen

HW5

```
> library(faraway)
> d0=data.frame(hsb)
> d0 = hsb[,-1]
> str(d0)
'data.frame':   200 obs. of  10 variables:
 $ gender : Factor w/ 2 levels "female","male": 2 1 2 2 2 2 2 2 2 2 ...
 $ race   : Factor w/ 4 levels "african-amer",..: 4 4 4 4 4 4 1 3 4 1 ...
 $ ses    : Factor w/ 3 levels "high","low","middle": 2 3 1 1 3 3 3 3 3 3 ...
 $ schtyp : Factor w/ 2 levels "private","public": 2 2 2 2 2 2 2 2 2 2 ...
 $ prog   : Factor w/ 3 levels "academic","general",..: 2 3 2 3 1 1 2 1 2 1 ...
 $ read   : int  57 68 44 63 47 44 50 34 63 57 ...
 $ write  : int  52 59 33 44 52 52 59 46 57 55 ...
 $ math   : int  41 53 54 47 57 51 42 45 54 52 ...
 $ science: int  47 63 58 53 53 63 53 39 58 50 ...
 $ socst  : int  57 61 31 56 61 61 61 36 51 51 ...
```

> #a)

```
> table(d0$gender,d0$prog)

        academic general vocation
  female       58      24       27
  male         47      21       23
> round(prop.table(table(d0$gender,d0$prog),2)*100,2)

        academic general vocation
  female    55.24   53.33    54.00
  male      44.76   46.67    46.00
> # The proportion of females choosing the three different programs is almost similar,
55.24%，53.33%，54.00%
> # Likewise,the proportion of males choosing the three different programs is almost similar,
44.76%，46.67%，46.00%
> # Therefore, for different levels of program, proportions of female and male is close to
proportion of female and male in total population
> # Therefore, gender is not a good predictor.
>
> table(d0$ses,d0$prog)

        academic general vocation
  high         42       9        7
  low          19      16       12
  middle       44      20       31
```

```
> round(prop.table(table(d0$ses,d0$prog),2)*100,2)

        academic general vocation
  high     40.00   20.00    14.00
  low      18.10   35.56    24.00
  middle   41.90   44.44    62.00
> # The proportion of high SES choosing the three different program is different,40.00%, 20.00%, 14.00%
> #Likewise, the proportion of low and middle SES choosing the three different program are different
> # Therefore, SES is a good predictor to predict the program they choose.
>
```

> #b)

```
> library(nnet)
> m1=multinom(prog~.,data=d0)
# weights:  42 (26 variable)
initial  value 219.722458
iter  10 value 171.814970
iter  20 value 153.793692
iter  30 value 152.935260
final  value 152.935256
converged
> summary(m1)
Call:
multinom(formula = prog ~ ., data = d0)

Coefficients:
         (Intercept)  gendermale raceasian racehispanic racewhite     seslow
general     3.631901 -0.09264717  1.352739   -0.6322019 0.2965156 1.09864111
vocation    7.481381 -0.32104341 -0.700070   -0.1993556 0.3358881 0.04747323
         sesmiddle schtyppublic        read       write       math    science
general  0.7029621    0.5845405 -0.04418353 -0.03627381 -0.1092888 0.10193746
vocation 1.1815808    2.0553336 -0.03481202 -0.03166001 -0.1139877 0.05229938
               socst
general  -0.01976995
vocation -0.08040129

Std. Errors:
         (Intercept) gendermale raceasian racehispanic racewhite    seslow
general     1.823452  0.4548778  1.058754    0.8935504 0.7354829 0.6066763
vocation    2.104698  0.5021132  1.470176    0.8393676 0.7480573 0.7045772
         sesmiddle schtyppublic       read      write       math    science
```

general 0.5045938   0.5642925 0.03103707 0.03381324 0.03522441 0.03274038
vocation 0.5700833   0.8348229 0.03422409 0.03585729 0.03885131 0.03424763
            socst
general 0.02712589
vocation 0.02938212

Residual Deviance: 305.8705
AIC: 357.8705

#Coefficients read   write   math   science   socst
#general -0.05445264 -0.03716360 -0.1037470 0.1065258 -0.01786542
#vocation -0.04078359 -0.03220268 -0.1099712 0.0537472 -0.07959798
#the unexpected coefficients is the what science subject have
#which is the only one subject have positive influence
#to choose general or vocation program rather than academic program
#(Since the base level of program is academic, so when other predictors hold
#the score of subject science is high, the probability of choosing general and vocation program will increase
#And because the sum of probability of three program is 1
#the probability of choosing academic program will decreases.)

> #c)

> n1=nrow(d0)
> step1=stepAIC(m1,k=log(n1))
Start: AIC=443.63
prog ~ gender + race + ses + schtyp + read + write + math + science +
    socst

weights: 39 (24 variable)
initial value 219.722458
iter 10 value 171.468391
iter 20 value 153.592758
final value 153.142827
converged
weights: 33 (20 variable)
initial value 219.722458
iter 10 value 172.925326
iter 20 value 156.065379
final value 155.776076

#didn’t show all

Call:

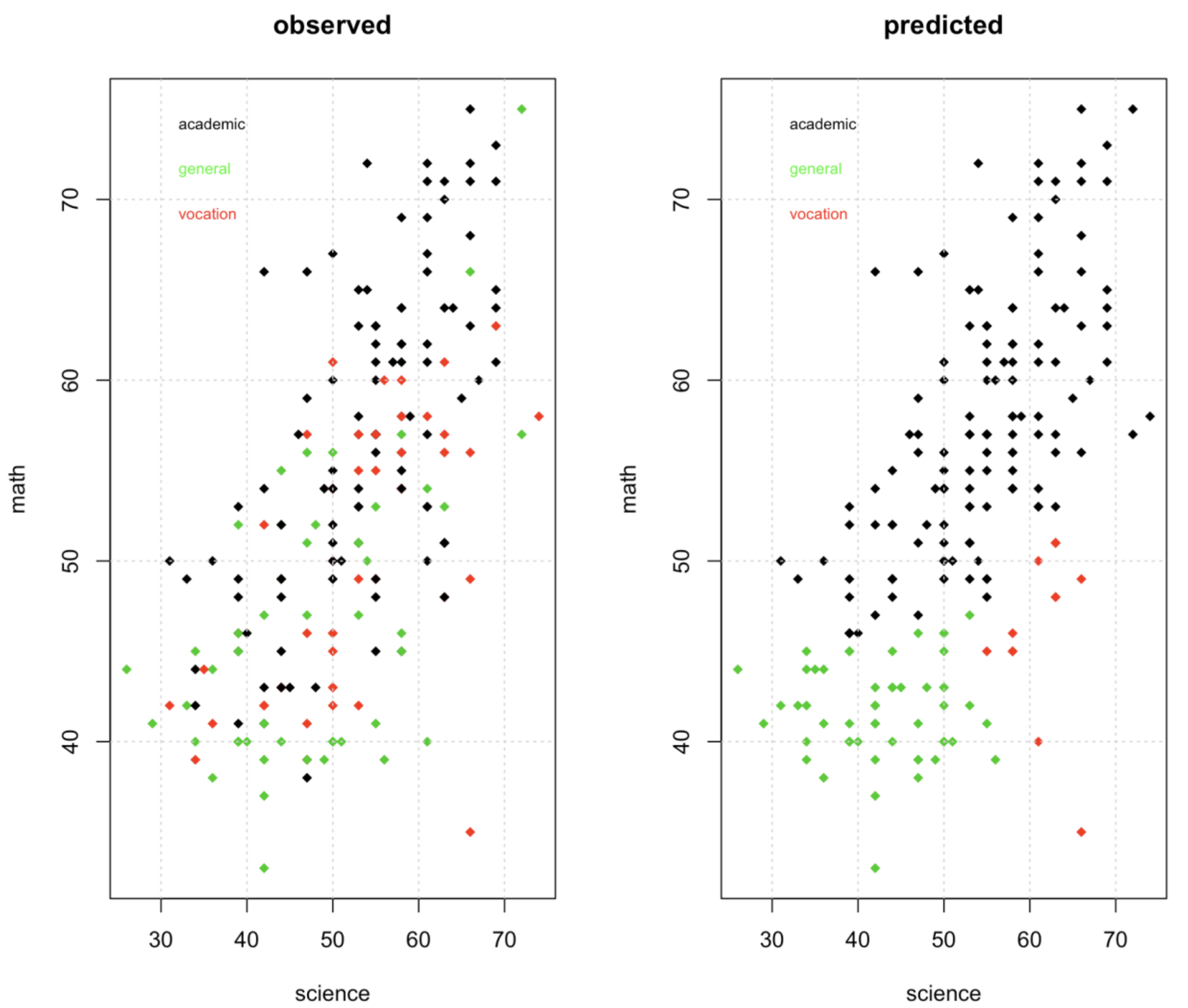
```
multinom(formula = prog ~ ., data = d0)

Step:  AIC=374.52
prog ~ math + socst

# weights:  9 (4 variable)
initial  value 219.722458
final  value 181.550192
converged
# weights:  9 (4 variable)
initial  value 219.722458
final  value 178.114227
converged
       Df    AIC
<none>     374.52
- socst  2 377.42
- math   2 384.29
>
#iteration stop when BIC=374.52, the variables I choose is math and socst
```

> #d)

```
> m3=multinom(prog~math+science,data=d0)
# weights:  12 (6 variable)
initial  value 219.722458
iter  10 value 175.074512
final  value 175.074335
converged
> par(mfrow=c(1,2))
> #predict probabilities
> pi.hat=predict(m3,newdata = d0,type="probs")
> ypred = apply(pi.hat,1,which.max)
> labels=c("academic","general","vocation")
> colors = c("black","green","red")
> plot(math~science,d0,col=d0$prog,pch=18,main="observed")
> legend("topleft",legend=labels,bty="n",text.col = colors,cex = 0.7)
> grid()
> plot(math~science,d0,col=ypred,pch=18,main="predicted")
> legend("topleft",legend=labels,bty="n",text.col = colors,cex = 0.7)
> grid()
>
```




```
> #error rate
> 1-sum(diag(prop.table(table(d0$prog,ypred))))
[1] 0.425
> #so the error rate is 42.5%
>
```

> #e)

```
> library(class)
> y = d0$prog
> str(d0)
'data.frame':   200 obs. of  10 variables:
 $ gender : Factor w/ 2 levels "female","male": 2 1 2 2 2 2 2 2 2 2 ...
 $ race   : Factor w/ 4 levels "african-amer",..: 4 4 4 4 4 4 1 3 4 1 ...
 $ ses    : Factor w/ 3 levels "high","low","middle": 2 3 1 1 3 3 3 3 3 3 ...
 $ schtyp : Factor w/ 2 levels "private","public": 2 2 2 2 2 2 2 2 2 2 ...
 $ prog   : Factor w/ 3 levels "academic","general",..: 2 3 2 3 1 1 2 1 2 1 ...
```

```
 $ read   : int  57 68 44 63 47 44 50 34 63 57 ...
 $ write  : int  52 59 33 44 52 52 59 46 57 55 ...
 $ math   : int  41 53 54 47 57 51 42 45 54 52 ...
 $ science: int  47 63 58 53 53 63 53 39 58 50 ...
 $ socst  : int  57 61 31 56 61 61 61 36 51 51 ...
> x = d0[,c(8,9)] #math and science
> str(x)
'data.frame':   200 obs. of  2 variables:
 $ math   : int  41 53 54 47 57 51 42 45 54 52 ...
 $ science: int  47 63 58 53 53 63 53 39 58 50 ...
> x=scale(x)
> head(x)
        math    science
1 -1.24300207 -0.4898549
2  0.03789315  1.1261613
3  0.14463442  0.6211562
4 -0.60255446  0.1161512
5  0.46485822  0.1161512
6 -0.17558939  1.1261613
> #k=3
> set.seed(1)
> ypred3=knn(x,x,y,3)
> par(mfrow=c(1,2))
> plot(math~science,d0,col=d0$prog,pch=18,main="observed")
> legend("topleft",legend=labels,bty="n",text.col = colors,cex = 0.7)
> grid()
> plot(math~science,d0,col=ypred3,pch=18,main="predicted")
> legend("topleft",legend=labels,bty="n",text.col = colors,cex = 0.7)
> grid()
> #error rate
> 1-sum(diag(prop.table(table(y,ypred3))))
[1] 0.31
> #so the error rate is 31%

> #k=5
> set.seed(1)
> ypred5=knn(x,x,y,5)
> plot(math~science,d0,col=d0$prog,pch=18,main="observed")
> legend("topleft",legend=labels,bty="n",text.col = colors,cex = 0.7)
> grid()
> plot(math~science,d0,col=ypred5,pch=18,main="predicted")
> legend("topleft",legend=labels,bty="n",text.col = colors,cex = 0.7)
> grid()
> #error rate
```

```
> 1-sum(diag(prop.table(table(y,ypred5))))
[1] 0.325
> #so the error rate is 32.5%
```