Multiclass logistic regression

Milica Cudina

For a similar analysis, look at this tutorial from UCLA.

First, we import the data.

```
data<-read.csv("hsbdemo.csv")
data
```

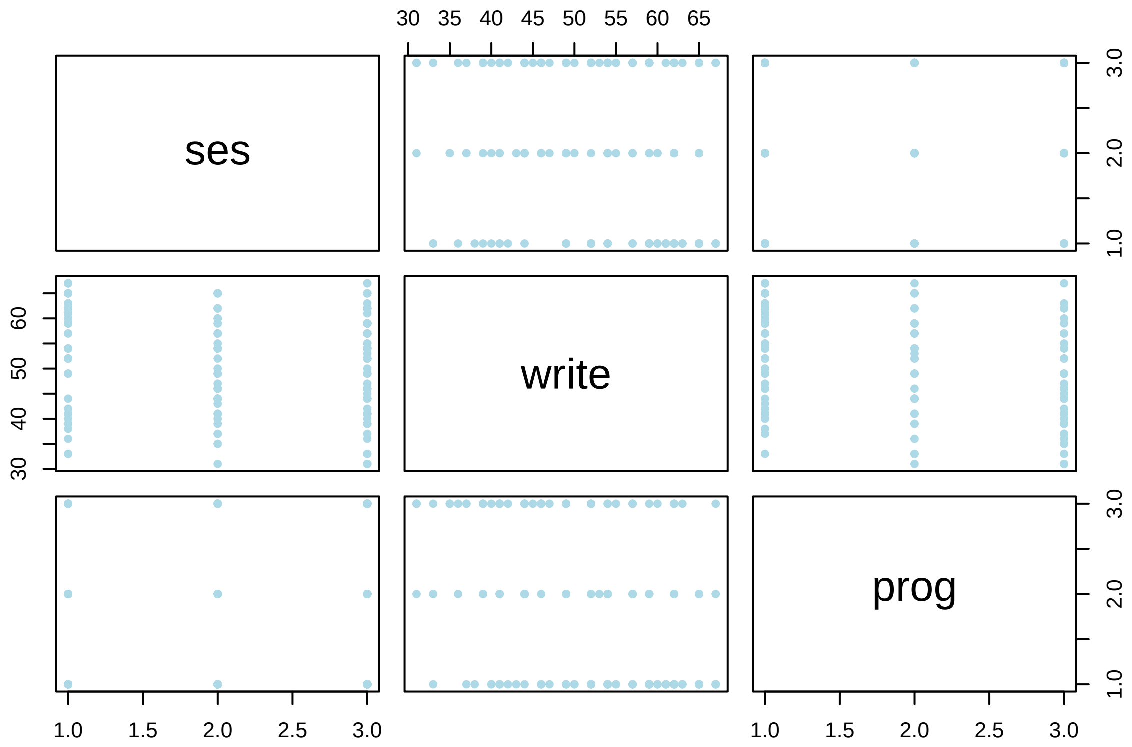
```
##   X  id female    ses schtyp     prog read write math science socst
## 1 1  45 female    low public vocation   34    35   41      29    26
## 2 2 108   male middle public  general   34    33   41      36    36
## 3 3  15   male   high public vocation   39    39   44      26    42
## 4 4  67   male    low public vocation   37    37   42      33    32
## 5 5 153   male middle public vocation   39    31   40      39    51
## 6 6  51 female   high public  general   42    36   42      31    39
## 7 7 164   male middle public vocation   31    36   46      39    46
##         honors awards cid
## 1 not enrolled      0   1
## 2 not enrolled      0   1
## 3 not enrolled      0   1
## 4 not enrolled      0   1
## 5 not enrolled      0   1
## 6 not enrolled      0   1
## 7 not enrolled      0   1
##  [ reached 'max' / getOption("max.print") -- omitted 193 rows ]
```

The data set contains variables on 200 students. We will focus on a small subset. The predictor variables will be social economic status `ses` (a three-level categorical variable) and writing score `write` (a quantitative variable). The outcome variable is program type `prog` (a three-level categorical variable). Since I am interested in just this subset, I will create a smaller data frame to analyze.

```
df=data.frame(data$ses, data$write, data$prog)
colnames(df)<-c("ses", "write", "prog")
attach(df)
```

Some exploratory data analysis is called for. The first idea is, probably to try the `plot` command.

```
plot(df,
     pch=20, col="lightblue")
```



As we can see, this is not too useful. To look at the *association* between `ses` and `prog`, a two-way table might do the trick.

```
tab=table(ses, prog)
tab
```

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

Which test might we use to test the independence hypothesis?

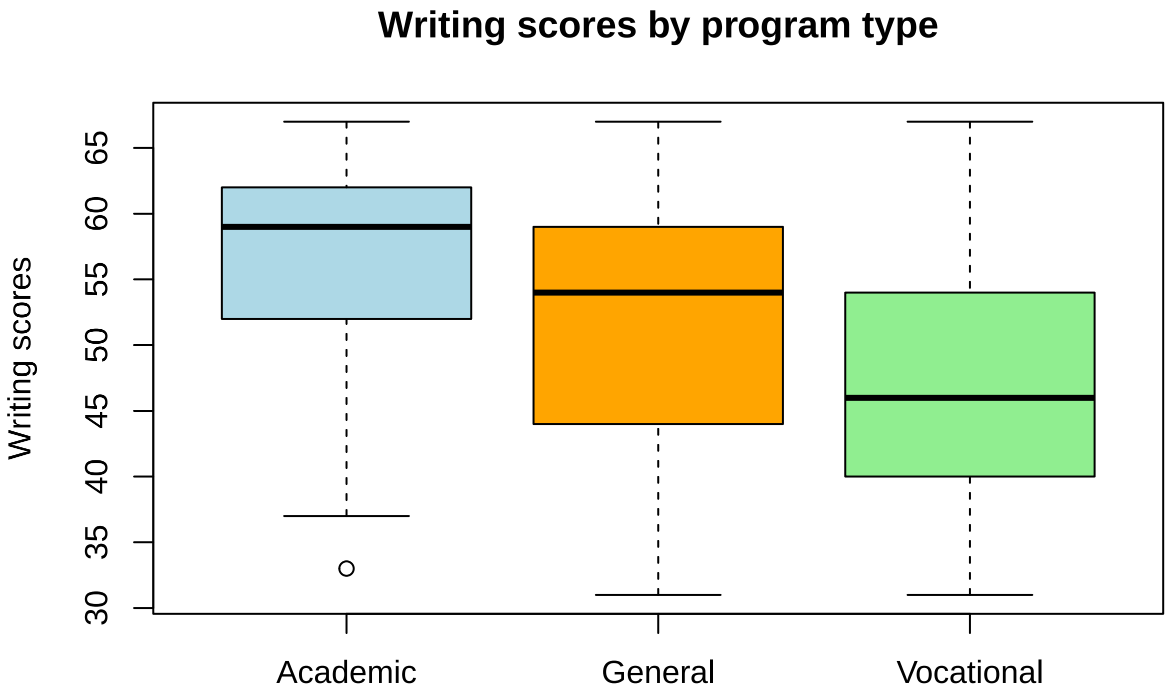
```
chi2<-chisq.test(tab)
chi2
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  tab
## X-squared = 16.604, df = 4, p-value = 0.002307
```

What about the association between the writing score and the program type? Side-by-side boxplots might give insight.

```
write.ac=write[which(prog=="academic")]
write.gen=write[which(prog=="general")]
write.voc=write[which(prog=="vocation")]

boxplot(write.ac,write.gen, write.voc,
  main = "Writing scores by program type",
  ylab = "Writing scores",
  names = c("Academic", "General", "Vocational"),
  col=c("lightblue", "orange", "lightgreen")
)
```



We see that there is an association, but we will learn more about the extent of the effect if we run a multiclass logistic regression.

`multinom`

Out first option is the `multinom` implementation in the `nnet` package.

```
#install.packages("nnet")
```