Homework 4

Shun-Lung Chang, Dilip Hiremath

```
library(foreign)
hsb <- read.dta('data/hsbdemo.dta')
```

1. Cross-tabulate the variables ses and prog.

```
ct <- table(hsb$prog, hsb$ses)
ct
```

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

(a) (half a point) Which program was chosen by the largest fraction of students with high socio-economic status?

According to the table above, academic program was choosen by the largest fraction of students with high socio-economic status.

(b) (half a point) How many percent of students with low socio-economic status selected the general program?

There is 34.04% students with low socio-economic status selecting the general program.

```
ct[, 1] / sum(ct[, 1])
```

```
  general  academic  vocation
0.3404255 0.4042553 0.2553191
```

(c) (half a point) In the academic program are there more students with middle socioeconomic status than students with high socio-economic status?

Yes, there are more students with middle socio-economic status than that with high socio-economic status from the table.

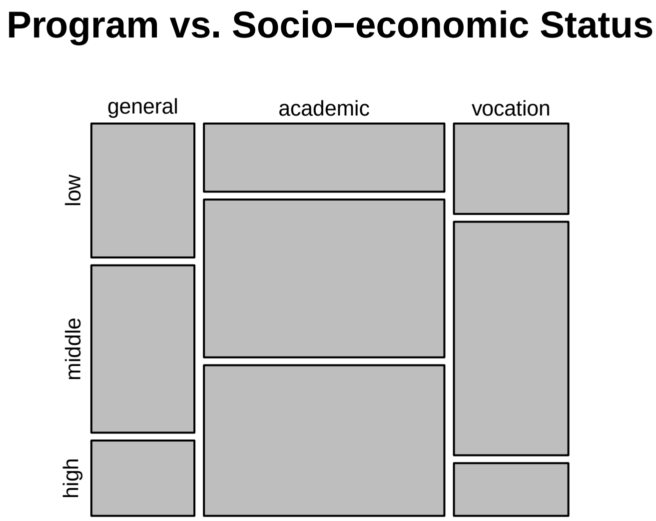
(d) (half a point) What is the least-frequent combination of the two variables?

The least frequest combination of the two variables is high socio-economic status and vocation program.

2. You continue with your analysis of the relationship between ses and prog.

(a) (half a point) Draw a mosaicplot visualising the contingency table of program choice and socio-economic status.

```
mosaicplot(ct, main = 'Program vs. Socio-economic Status')
```



(b) (1.5 points) Are students with low ses less likely (as measured in odds) to choose the academic program than students with higher socio-economic status? Calculate the odds ratios for choosing the academic program comparing students with low ses to students with middle ses and to students with high ses. [hint: use the command loddsratio in the package vcd. First, aggregate the variable prog into a binary variable indicating whether the student has chosen an academic program yes or no.]

Yes, students with low ses less likely (as measured in odds) to choose the academic program than students with higher socio-economic status. The odds ratios for choosing the academic program comparing students with low ses to students with middle ses is 1:1.27 The odds ratios for choosing the academic program comparing students with middle ses to students with high ses is 1:3.0426

```
library(vcd)
hsb$academic_yes <- hsb$prog == 'academic'

ct_2 <- table(hsb$academic_yes, hsb$ses) / nrow(hsb)
oddsratio(ct_2, log = FALSE)
```

```
 odds ratios for  and

 low:middle middle:high
   1.271414    3.042614
```

3. Now, you assess the relationship between prog and ses using the χ^2-statistic.

(a) (1 point) Calculate the χ^2-test to assess the relationship between ses and prog. Is the relationship statistically significant?

The statistic is 16.604 between ses and prog. The p-value of 0.002307 implies that the relationship is significantly dependant.

```
chisq.test(hsb$prog, hsb$ses)
```

```

	Pearson's Chi-squared test

data:  hsb$prog and hsb$ses
X-squared = 16.604, df = 4, p-value = 0.002307
```

(b) (1 point) Calculate the expected frequencies under the assumption that socio-economic status has no effect on program choice. For which cells are expected frequencies higher than the observed ones?

The expected frequencies is shown as the table below.

```
expected <- chisq.test(hsb$prog, hsb$ses)$expected
expected
```

```
          hsb$ses
hsb$prog      low middle  high
  general  10.575 21.375 13.05
  academic 24.675 49.875 30.45
  vocation 11.750 23.750 14.50
```

For the cells low and academic, middle and general, middle and academic, high and general, high and vacation the expected frequencies higher than the observed ones.

```
expected - ct
```

```
          hsb$ses
hsb$prog      low  middle    high
  general  -5.425   1.375   4.050
  academic  5.675   5.875 -11.550
  vocation -0.250  -7.250   7.500
```

4. In the following, perform the last analysis separately for female and male students.

```
hsb_m <- hsb[hsb$female == 'male', ]
hsb_f <- hsb[hsb$female == 'female', ]
```

(a) (half a point) Calculate the χ^2-test to assess the relationship between ses and prog.

The test statistic is 11.557 for male students between ses and prog. And the p-value shows that program choice is dependent of socio-economic status for male students.

```r
chisq.test(hsb_m$prog, hsb_m$ses)
```

```
    Pearson's Chi-squared test

data:  hsb_m$prog and hsb_m$ses
X-squared = 11.557, df = 4, p-value = 0.02097
```

The test statistic is 7.473 for female students between ses and prog. And the p-value shows that program choice is independent of socio-economic status for female students.

```r
chisq.test(hsb_f$prog, hsb_f$ses)
```

```
    Pearson's Chi-squared test

data:  hsb_f$prog and hsb_f$ses
X-squared = 7.473, df = 4, p-value = 0.1129
```

(b) (half a point) Calculate the expected frequencies under the assumption that socio-econmic status has no effect on program choice. For which cells are expected frequencies higher than the observed ones?

For the cells low and academic, middle and general, middle and academic, high and general, high and vacation, the expected frequencies higher than the observed ones for male students.

```r
# male
observed_m <- table(hsb_m$prog, hsb_m$ses)
chisq.test(hsb_m$prog, hsb_m$ses)$expected - observed_m
```

```
          hsb_m$ses
hsb_m$prog        low     middle       high
  general  -3.5384615  0.8461538  2.6923077
  academic  3.7472527  2.2747253 -6.0219780
  vocation -0.2087912 -3.1208791  3.3296703
```

For the cells low and academic, middle and general, middle and academic, high and general, high and vacation the expected frequencies higher than the observed ones for female students.

```r
# female
observed_f <- table(hsb_f$prog, hsb_f$ses)
chisq.test(hsb_f$prog, hsb_f$ses)$expected - observed_f
```

```
          hsb_f$ses
hsb_f$prog        low     middle       high
  general  -1.9541284  0.5688073  1.3853211
  academic  2.0275229  3.5412844 -5.5688073
  vocation -0.0733945 -4.1100917  4.1834862
```

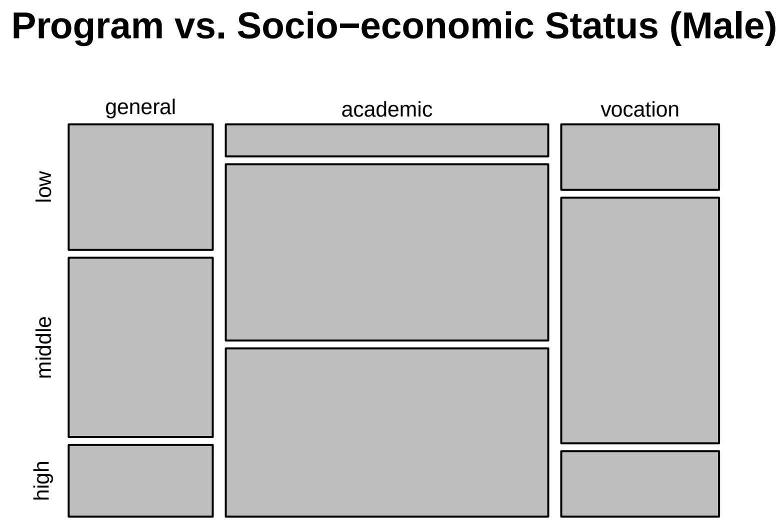
(c) (half a point) Do the results differ for the two sexes?

Yes, as can been seen from 4(a) and (b), the results differ for the two genders.

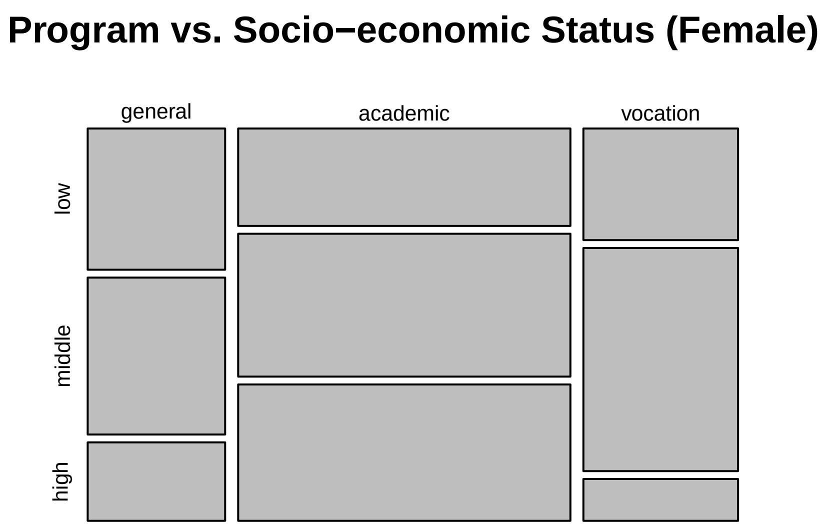
(d) (half a point) Visualise the relationships using mosaicplots. Get any differences between females and males in relation to socio-economic status and program choice visible in the plots?

As can be seen from the plots below, the blocks in the female students are more equally sized than the ones in the male students.

```
mosaicplot(observed_m, 'Program vs. Socio-economic Status (Male)')
```




```
mosaicplot(observed_f, 'Program vs. Socio-economic Status (Female)')
```



5. Create a multinomial logistic regression model using prog as dependent variable and the following predictors: female, ses, schtype, read, write, math, science, honors, awards. [hint: use the function multinom in hte package nnet.]

```
library(nnet)
mod_1 <- multinom(prog ~ . - id - socst - cid,
                  data = hsb)
```

```
# weights:  36 (22 variable)
initial  value 219.722458
```

```
iter  10 value 184.828226
iter  20 value 157.499339
final  value 157.443537
converged
```

```r
summary(mod_1)
```

```
Call:
multinom(formula = prog ~ . - id - socst - cid, data = hsb)

Coefficients:
         (Intercept) femalefemale sesmiddle   seshigh schtypprivate
academic   -5.498379   -0.1522507 0.3751645 1.0802062     0.5372474
vocation    4.163618    0.2804404 1.3235922 0.7959664    -1.2758046
                 read       write        math     science honorsenrolled
academic  0.052955458  0.06429399  0.10006643 -0.10199019      0.5743597
vocation -0.009609139 -0.02166161 -0.02995725 -0.03800852      1.9991952
             awards
academic -0.2703505
vocation -0.3525181

Std. Errors:
         (Intercept) femalefemale sesmiddle   seshigh schtypprivate
academic    2.347466    0.4533749 0.5061641 0.5772063     0.5550686
vocation    2.455693    0.5179411 0.5388177 0.7084789     0.8817585
               read      write       math    science honorsenrolled
academic 0.02919946 0.04927104 0.03480492 0.03090596      0.8644712
vocation 0.03322165 0.05054331 0.03756352 0.03224286      1.1038196
            awards
academic 0.2936491
vocation 0.3856715

Residual Deviance: 314.8871
AIC: 358.8871
```

(a) (half a point) How large is the AIC score for this model?

The AIC score is 358.8871.

```r
AIC(mod_1)
```

```
[1] 358.8871
```

(b) (1.5 points) The default output does not include p-values. Compute p-values based on the Wald-test statistics and determine the coefficients that are statistically significantly different from zero!

As the table below shows, the variables *academic:(Intercept)*, *academic:math*, *academic:science* and *vocation:sesmiddle* are significant under significance level of 5%.

```r
# compute p-values by definition
# z <- summary(mod_1, Wald.ratios = TRUE)$Wald.ratios
# p <- (1 - pnorm(abs(z))) * 2
# p
```

```r
# compute p-values by package function
library(AER)
coeftest(mod_1)[, 4] < 0.05
```

```
   academic:(Intercept)   academic:femalefemale      academic:sesmiddle 
                   TRUE                   FALSE                   FALSE 
       academic:seshigh  academic:schtypprivate           academic:read 
                  FALSE                   FALSE                   FALSE 
         academic:write           academic:math        academic:science 
                  FALSE                    TRUE                    TRUE 
academic:honorsenrolled         academic:awards    vocation:(Intercept) 
                  FALSE                   FALSE                   FALSE 
  vocation:femalefemale      vocation:sesmiddle        vocation:seshigh 
                  FALSE                    TRUE                   FALSE 
 vocation:schtypprivate           vocation:read          vocation:write 
                  FALSE                   FALSE                   FALSE 
          vocation:math        vocation:science vocation:honorsenrolled 
                  FALSE                   FALSE                   FALSE 
        vocation:awards 
                  FALSE 
```

6. Using the model from the previous question and the backward strategy with criterion AIC for variable selection, determine the significant coefficients in the resulting model.

```r
mod_2 <- step(mod_1, direction = 'backward')
```

```r
summary(mod_2)
```

```
Call:
multinom(formula = prog ~ ses + schtyp + read + math + science, 
    data = hsb)

Coefficients:
         (Intercept) sesmiddle   seshigh schtypprivate        read
academic   -3.745688  0.323115 1.0358034      0.608257  0.05912408
vocation    3.907946  1.183126 0.7014962     -1.408038 -0.01218565
                math     science
academic  0.10745053 -0.09076914
vocation -0.03078266 -0.04770648

Std. Errors:
         (Intercept) sesmiddle   seshigh schtypprivate       read
academic    1.401302 0.4876102 0.5648570     0.5484718 0.02807034
vocation    1.564001 0.5201513 0.6880376     0.8662165 0.03127526
               math    science
academic 0.03270710 0.02859475
vocation 0.03535214 0.03005018

Residual Deviance: 322.9919 
AIC: 350.9919 
```

(a) (1 point) Which predictors are included in the resulting model?

According to the table above, the variables *ses*, *schtyp*, *read*, *math* and *science* are included in the final model.

(b) (half a point) What is the BIC score of the resulting model?

The BIC score is 397.1684.

```
BIC(mod_2)
```

```
[1] 397.1684
```

(c) (half a point) What is the log-likelihood score of this model?

The log-likelihood score is -161.496.

```
logLik(mod_2)
```

```
'log Lik.' -161.496 (df=14)
```

7. (2 points) Using the final model that resulted in Question 6 predict the probabilities for the three program types for the combination of all factor levels and the average score of numeric predictors in the model.

The table below indicates academic is the most possible outcome in all combinations. But the students in the private school are more likely to take an academic program than that in the public school.

```
d <- expand.grid(ses = c('low', 'middle', 'high'),
                 schtyp = c('private', 'public'),
                 read = mean(hsb$read),
                 math = mean(hsb$math),
                 science = mean(hsb$science))
```

```
pred <- predict(mod_2, newdata = d, type = 'probs', se = TRUE)
cbind(d, pred)
```

```
     ses  schtyp  read   math science   general  academic   vocation
1    low private 52.23 52.645   51.85 0.2801551 0.6897408 0.03010409
2 middle private 52.23 52.645   51.85 0.2104441 0.7157334 0.07382253
3   high private 52.23 52.645   51.85 0.1226533 0.8507663 0.02658044
4    low  public 52.23 52.645   51.85 0.3597989 0.4821528 0.15804831
5 middle  public 52.23 52.645   51.85 0.2333606 0.4319957 0.33464371
6   high  public 52.23 52.645   51.85 0.1766363 0.6668812 0.15648254
```

8. (2 points) Again using the final model that resulted in Question 6, we now want to investigate the specific dependency on the math score. Generate new data such that you have for each combination of factor levels a total of 51 math scores running from 30 to 80 in increments of one. The other numeric predictors enter again with their mean score into the prediction. Compute the predictions and average them for each level of socio-economic status.

The table below shows the average probabilities of taking different programs with respect to different socio-economic levels. As can be seen from the table, taking an academic program is the most probable in three socio-economic levels. Furthermore, people with a high socio-economic status is more likely to have an academic program.

```
d <- expand.grid(ses = c('low', 'middle', 'high'),
                 schtyp = c('private', 'public'),
                 read = mean(hsb$read),
                 math = 30:80,
                 science = mean(hsb$science))

pred <-  predict(mod_2, newdata = d, type = 'probs', se = TRUE)

bind_d <- cbind(d, pred)
aggregate(bind_d[, 6:8], by = list(bind_d$ses), FUN = 'mean')
```

```
  Group.1   general  academic  vocation
1     low 0.2978392 0.5956783 0.1064825
2  middle 0.1993257 0.5887810 0.2118934
3    high 0.1629723 0.7138563 0.1231714
```