Homework 4

Shun-Lung Chang, Dilip Hiremath

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

1. Cross-tabulate the variables ses and prog.

(a) (half a point) Which program was chosen 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?

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

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

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.

(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.]

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?

(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?

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

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

(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?

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

(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?

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.]

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

```
# weights:  39 (24 variable)
initial  value 219.722458
iter  10 value 183.152926
iter  20 value 158.073364
iter  30 value 157.310926
final  value 157.310831
converged
```

```r
summary(mod_1)
```

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

Coefficients:
         (Intercept)           id femalefemale sesmiddle   seshigh
academic   -5.420908 0.0009199278   -0.1464439  0.359010 1.0674796
vocation    4.242615 0.0024070518    0.2820687  1.287898 0.7653309
         schtypprivate         read       write        math     science
academic     0.4756503  0.053786496  0.06269320  0.09960699 -0.10390083
vocation    -1.4593116 -0.008368905 -0.02330601 -0.03026180 -0.04271602
         honorsenrolled     awards
academic      0.5774178 -0.2621733
vocation      2.0055515 -0.3441895

Std. Errors:
         (Intercept)          id femalefemale sesmiddle   seshigh
academic    2.377938 0.004340101    0.4538012 0.5098168 0.5795898
vocation    2.481514 0.004700197    0.5186325 0.5430944 0.7118088
         schtypprivate       read      write       math    science
academic     0.6290497 0.02947293 0.04982956 0.03484101 0.03227723
vocation     0.9545994 0.03331620 0.05098022 0.03760150 0.03357679
         honorsenrolled    awards
academic       0.865841 0.2963429
vocation       1.102180 0.3886943

Residual Deviance: 314.6217
AIC: 362.6217
```

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

```r
AIC(mod_1)
```

```
[1] 362.6217
```

(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!

```
# compute p-values by definition
# z <- summary(mod_1)$coefficients/summary(mod_1)$standard.errors
# p <- (1 - pnorm(abs(z))) * 2
# p

# compute p-values by package function
library(AER)
coeftest(mod_1)
```

```
z test of coefficients:

                          Estimate  Std. Error z value Pr(>|z|)
academic:(Intercept)    -5.42090774  2.37793825 -2.2797 0.022627
academic:id              0.00091993  0.00434010  0.2120 0.832138
academic:femalefemale   -0.14644387  0.45380120 -0.3227 0.746919
academic:sesmiddle       0.35901000  0.50981680  0.7042 0.481312
academic:seshigh         1.06747955  0.57958983  1.8418 0.065507
academic:schtypprivate   0.47565029  0.62904966  0.7561 0.449565
academic:read            0.05378650  0.02947293  1.8249 0.068009
academic:write           0.06269320  0.04982956  1.2582 0.208336
academic:math            0.09960699  0.03484101  2.8589 0.004251
academic:science        -0.10390083  0.03227723 -3.2190 0.001286
academic:honorsenrolled  0.57741781  0.86584103  0.6669 0.504845
academic:awards         -0.26217327  0.29634290 -0.8847 0.376321
vocation:(Intercept)     4.24261532  2.48151427  1.7097 0.087324
vocation:id              0.00240705  0.00470020  0.5121 0.608569
vocation:femalefemale    0.28206873  0.51863249  0.5439 0.586531
vocation:sesmiddle       1.28789751  0.54309441  2.3714 0.017721
vocation:seshigh         0.76533093  0.71180878  1.0752 0.282289
vocation:schtypprivate  -1.45931165  0.95459944 -1.5287 0.126335
vocation:read           -0.00836890  0.03331620 -0.2512 0.801662
vocation:write          -0.02330601  0.05098022 -0.4572 0.647558
vocation:math           -0.03026180  0.03760150 -0.8048 0.420933
vocation:science        -0.04271602  0.03357679 -1.2722 0.203306
vocation:honorsenrolled  2.00555154  1.10218031  1.8196 0.068817
vocation:awards         -0.34418946  0.38869428 -0.8855 0.375886
```

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.

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

```
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?

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

```r
BIC(mod_2)
```

```
[1] 397.1684
```

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

```r
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.

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

```
     ses  schtyp  read   math science
1    low private 52.23 52.645   51.85
2 middle private 52.23 52.645   51.85
3   high private 52.23 52.645   51.85
4    low  public 52.23 52.645   51.85
5 middle  public 52.23 52.645   51.85
```

```
6   high  public 52.23 52.645   51.85
```

```r
predict(mod_2, newdata = d, type = 'probs', se = TRUE)
```

```
    general  academic   vocation
1 0.2801551 0.6897408 0.03010409
2 0.2104441 0.7157334 0.07382253
3 0.1226533 0.8507663 0.02658044
4 0.3597989 0.4821528 0.15804831
5 0.2333606 0.4319957 0.33464371
6 0.1766363 0.6668812 0.15648254
```

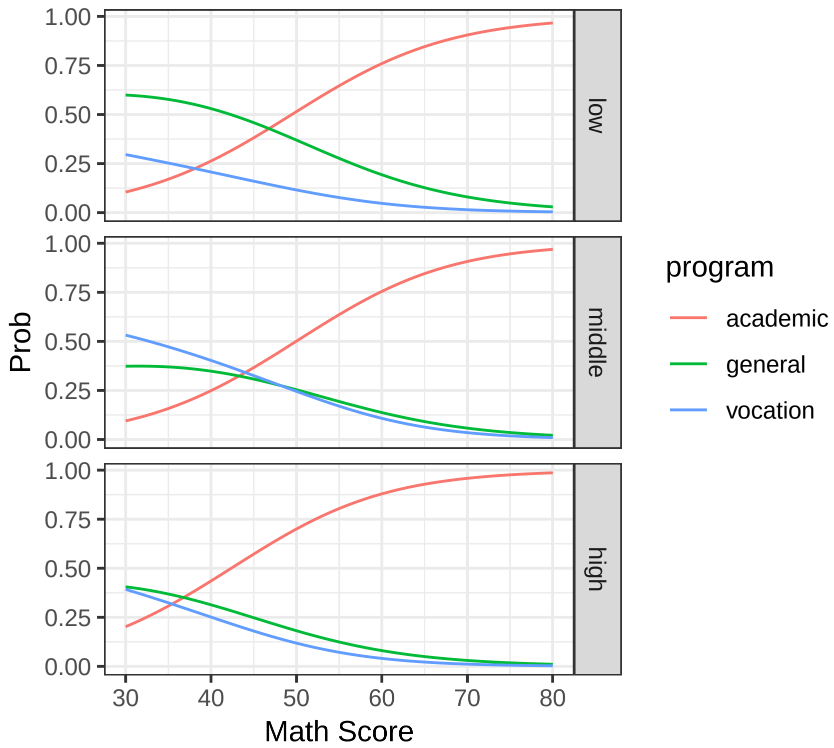
8. (2 points) Again using the final model that resulted in Question 6, we now want to investigate the specific dependency on the math score. Generate new data such that you have for each combination of factor levels a total of 51 math scores running from 30 to 80 in increments of one. The other numeric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

```r
d <- expand.grid(ses = c('low', 'middle', 'high'),
                 schtyp = c('private', 'public'),
                 read = mean(hsb$read),
                 math = 0,
                 science = mean(hsb$science))

preds <- lapply(30:80, function(x) {
    d$math <- x
    pred <- predict(mod_2, newdata = d, type = 'probs', se = TRUE)
    aggregate(pred, list(d$ses), mean)
})

# list to data.frame
library(tidyr)
df <- do.call(rbind, preds)
df$score <- rep(30:80, each = 3)
# transpose
df_long <- gather(df, key = program, value = prob, -c(score, Group.1))

library(ggplot2)
ggplot(df_long, aes(x = score, y = prob, color = program)) +
    geom_line() +
    facet_grid(rows = vars(Group.1)) +
    labs(x = "Math Score", y = "Prob") +
    theme_bw()
```


Prob
Math Score
low
middle
high
program
academic
general
vocation
1.00
0.75
0.50
0.25
0.00
30
40
50
60
70
80