Statistical Modeling with R - Fall 2016
Homework 4

DUE IN: Tuesday, 11.10.2016 at 11:59 a.m.,

HOW: electronically in pdf-format via submission to `www.turnitin.com`

Class id: 13494794

enrollment password: Ti20Ta16Nic

Please register for the class on turnitin ahead of time.

GROUP WORK: is allowed with a maximum of 3 persons per group. PLEASE stay within the same group throughout the semester. Only one solution is accepted and graded per group. Please include the names of all group members on each assignment.

HOW MANY: There will be a total of six homework assignments in this semester. We will do a random selection of questions to be graded. Each week a total of eight points can be gained. Only the five best homeworks will be counted.

DUE DATES: 20.09., 27.09., 04.10., 11.10., 18.10., 25.10. (tentatively, subject to change)

FORMAT: Please do the required analyses and provide answers in complete sentences. **Provide the R syntax for the commands.** Just report those statistics that are relevant; do not copy complete R output. Integrate requested figures or tables into your document and give a brief verbal comment/caption on them.

Study programs

Entering high school students make program choices among general program, vocational program and academic program. Their choice might be modeled by using their social economic status and various test scores. The data set, which is in STATA format, contains variables on 200 students. The outcome variable is `prog`, program type.

Data `hsbdemo.dta`

Source `http://www.ats.ucla.edu/stat/data/hsbdemo.dta`

Variables Description of variables:

id identifier

female student's gender (`female` or `male`)

ses socio-economic status (three categories: `low, middle` and `high`

schtyp school type (`private` or `public`)

prog program type: `general`, `vocation` or `academic`

read reading score (numeric)

write writing score (numeric)

math mathematics score (numeric)

science science score (numeric)

honors indicator whether student is enrolled in honors program

awards number of awards received

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 socio-economic 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 `oddsratio` in the package `epitools`. 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`.

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

   (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!

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.

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

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

   (c) (half a point) What is the loglikelihood score of this model?

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.

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.