Multivariate Statistics: Exercise 3

October 24, 2018

Cluster analysis:

Use the data `olives` from the package `classifly` for cluster analysis – see last exercise.

- Apply fuzzy-k-means clustering with the function `fanny()` from `library(cluster)`, and with the function `cmeans()` from `library(e1071)`. In both cases, the result objects contain the list element `$cluster` with the cluster assignments, and `$membership` with the cluster memberships as proportions between 0 and 1. Use the function `ternary()` from the package `StatDA` to visualize the memberships. Which algorithm works better?

Regression analysis:

Use the data *schooldata.csv* (exercise web page and TUWEL) for regression. These are 70 observations for the following 8 variables:

education	education level of mother as measured in terms of percentage of high school graduates among female parents
occupation	highest occupation of a family member according to a pre-arranged rating scale
visit	parental visits index representing the number of visits to the school site
counseling	parent counseling index calculated from data on time spent with child on school-related topics such as reading together, etc.
teacher	number of teachers at a given site
reading	total reading score as measured by the Metropolitan Achievement Test
mathematics	total mathematics score as measured by the Metropolitan Achievement Test
selfesteem	Coopersmith Self-Esteem Inventory, intended as a measure of self-esteem

The aim is to explain scores on 3 different tests (reading, mathematics, selfesteem) from 70 school sites by means of the (remaining) 5 explanatory variables.

- Start with fitting the a linear model for each response separately. Example code looks as follows:

  ```
  m1 <- lm(reading~education+occupation+visit+counseling+teacher,data=d)
  ```

- Then fit a model to all 3 responses jointly. In the R code, the responses need to be joined using `cbind()`. Compare the outcomes with `summary()` applied on the result objects. What do you conclude? Look at disgnostics plots, if possible.

- Replace for the multivariate model the command `lm()` by `manova()`. Apply `summary()` to the result. What do you conclude?

- Eliminate from the 3 separate models step-by-step variables to reach an “optimal” model. This can be done with the function `step()` on the results of the full models. What is the criterion to eliminate variables? Which models do you obtain?

- How could stepwise variable selection be done for the multivariate model?

- Load the package `cvTools`. Suppose your multivariate model is stored in `mod`. Perform:

  ```
  plot(cvFit(mod,data=d,y=cbind(d$reading,d$mathematics,d$selfesteem),R=100))
  ```

  What is this command doing? Use this idea to do model selection.

- Look at plots y versus $\hat{y}$ for each response. The predictions are obtained with the `predict()` command. Compute as a measure of prediction quality the correlations between these quantities. Does it make sense?

Save your (successful) R code together with short documentations and interpretations of results in a text file (= R script file), named as *Matrikelnummer_3.R* (no word document, no plots). Submit this file to Exercise 3 of our tuwel course (deadline October 23).