High-dimensional statistics

Academic Year 2023–2024

Project n°2 : Dependence structure and linear modeling

1 Preliminary comment

This project may be done individually or in pairs (in the latter case, a unique project needs to be handed in, mentioning the two names). It is not compulsory to keep the same team as for project 1 and/or to keep working alone if that was the case for that project. When working in team, it is expected that all parts of the project have been developed in collaboration between the members of the team.

The project, written in English, is due on Wednesday 13 December 2023 (23h59) and needs to be submitted via eCampus. In the main body of the report (8 pages max), only the results, graphics and **interpretations** must be supplied and discussed (additional graphics or tables may be included in an annex). The `R` script used to compute the outputs of the analyses has to be submitted also as a separate file on eCampus. It is compulsory to use `R` but it is not compulsory to use the commands suggested/illustrated during the lectures. However, if other commands/libraries are used, some information on the inputs/outputs relative to the commands as well as some details on the advantages of this other approach have to be described in the `R-code`.

2 Data

For this project, the same data set as the one used for project 1 has to be used. In case some new teams are created, one of the data sets received individually for project 1 can be chosen. In case a team is split, each member of the team may keep the data.

In case the original data contained some missing values, it is suggested to work on the complete observations only. The application of the function `na.omit` to the data set allows to keep only the complete rows.

3 Graphical model of the dependence structure between the score variables

In project 1, the correlation matrix computed on all the score variables and estimated both in the classic and robust way has been discussed and interpreted. Here the scores will be divided into two groups according to the corresponding test and a graphical model based on a L_1-regularized covariance matrix will be represented in order to visualize the dependencies between the scores within each test.

1. Assuming that the multivariate normality assumption holds for the data base restricted to the scores of each test separately (the 5 scores of test `Saber11` on one side and the 6 scores of test `SaberPro` on the other side), provide plots of the number of edges that

would be included in each graph with respect to the value of the L_1-regularization parameter.

2. Fixing to 5 the number of edges to represent in each graph, represent the two graphical models and interpret, trying to put forward similarities/differences and providing some links, if possible, with the comments derived from the PCA results of Project 1.

3. The graphical models represented in question 2 are based on the multivariate normality assumption. Discuss the relevance of that assumption, using, as justification, some graphics or summary statistics maybe already provided in Project 1.

4 Linear modeling of the `SaberPro` scores

In this project, the linear dependence of the `SaberPro` test with respect to the initial score information and some "background" variables will be investigated. Specialists of educational sciences do not always agree on what might be expected : (i) either the initial and background variables are assumed to be highly explanatory on future university results or (ii) university programs tend to develop other skills implying that scores derived after such a training might not be influenced by previous results. The idea is to see whether the collected data support one out of these two theories.

1. It is planned to use the 5 score variables of test `Saber11` as well as the background variables `COMPUTER`, `TV`, `INTERNET`, `ACADEMIC_PROGRAM` and `SCHOOL_NAT` as explanatory variables of some linear models. Discuss about the potential issue of multicolinearity when including all these explanatory variables in the same model.

2. As there are 6 scores relative to the `SaberPro` test, only the two most promising and the least promising results will be investigated in detail. In order to find the score variables to keep in the analysis, derive[1] the R^2 values of the full linear models (one for each of the 6 score variables, "full" meaning that all the explanatory variables listed in question 1 have to be included). Keep the score variables with the two highest R^2 measures as well as the variable with the lowest R^2 measure.

3. For each of the three selected variables, provide a simplified model with justification (i.e. specify which selection technique has been used in order to derive the simplified models) and try to interpret the results.

4. Check the hypotheses required for the residuals derived from the simplified models.

5. Conclude with some overall comments on the analysis.

[1] The R^2 values have to be reported in the report.