Statistical Case Studies — Project 1 part 2 — 2022: Dementia risk prediction

due: Wednesday 30. 11. 2022 16:00h

In this second piece of coursework you will work with the easySHARE data. For information on data and variables, see the easySHARE data guide.

Task

The task is to build a model for dementia risk prediciton and assess the predictive performance of the model using validation.

1. In the first step of this project, we want to extract relevant variables related to dementia and dementia risk factors, along with an exploratory data analysis (EDA) of the variables extracted. The file easySHARE.Rmd contains some code to get you started. EasySHARE does not record diagnosis of dementia (e.g. Alzheimer's disease) in all waves. Instead, we will create a composite cognitive score as a proxy for dementia severity and the file easyShare.Rmd gives a suggestion for doing this. This compositve cognitive score should be your response variable.
2. You should expand the above code by extracting relevant variables related to dementia risk factors and any other variables that you wish to explore, for example, age and gender. You should also do the relevant pre-processing of your data. In your report you should comment on any issues with the data, such as missingness, and how you address them, along with possible limitations.
3. SHARE is a multi-country longitudinal survey, and for building/fitting your predictive demential model your model you should select a single country and only one wave. After leaving out 200 randomly selected observations, this constitutes your **training** data.
4. Produce summary statistics and graphical summaries relevant for the analysis.
5. Fit a model using the composite cognitive score as the response to the training data.
6. You should assess the predictive performance of your model by estimating the root mean squared error. You may also add to this other scores along with explanation and motivation (see your notes from Statistical Computing).
7. For validation you should use the following **test** data:
    - The test data is the same as the training data.
    - The test data are the 200 observations you left out from the initial data (same country and wave as training data).
    - The test data is the same wave as the training data, but a different country.
    - The test data is from a different wave than the training data, but the same country.
8. Can you assess the importance of variables for prediction using the rooth mean squared error?
9. Write up a report of your analysis (in word or latex NOT markdown).

Report format

The report should contain a clearly sign posted executive summary. That is, a section aimed at members of the general public presenting and explaining your results. This should contain graphs highlighting the most interesting finding of your analysis. This section of the report should be no longer than 500 words (not counting graphs) and should be readable without specialist statistical knowledge.

The remaining text in your report should be technical and aimed at statisticians and health researchers, explaining each step and what the conclusions were. **You should write out your model equation including assumptions.**

Your report must contain no more than 3000 words. This word limit is only on the main text including executive summary, and does not include the rest of the report (such as title page, graphs, tables, references, and appendices). Please see Learn for more details on formatting.

Your report should be submitted as a **pdf** file alongside a markdown with the analysis **.Rmd and html** file.The markdown .Rmd file should run with the project data.

Marking Scheme

There is no single correct analysis for this type of project, so you will not be marked on the basis of how close you get to some particular model answer. The marks are not subdivided, but will be allocated on a combination of statistical approach and justification, interpretation of results in context and presentation.

80 – 100% A report that could be presented to the client or collaborator with little or no revision. Analysis is sound so that conclusions are well-supported statistically. Interpretation is reasonably mature. The project should demonstrate a clear overview of the work, without getting lost in details, and be free of all but minor statistical errors. The work is to a publishable standard.

70-79% A report that could be presented to the client or collaborator with little or no revision. Analysis is sound so that conclusions are well-supported statistically. Interpretation is reasonably mature. The project should demonstrate a clear overview of the work, without getting lost in details, and be free of all but minor statistical errors.

60 – 69% A project that could be presented after a round of revision, but without having to redo much of the actual analysis. Some flaws in the analysis or presentation (or minor flaws in both), but basically sound. A good grasp of the statistics and context, so that interpretation is reasonable.

50 - 59% Major re-working required before the project could be presented, but containing some sound statistics demonstrating understanding of statistical modelling and its application. Reasonable presentation and organisation.

40 – 49% Major flaws in analysis and presentation, but demonstrating some understanding of statistics, and a reasonable attempt to present the results.

Fail (below 40%) Flawed analysis demonstrating little or no understanding of statistics, and/or incomprehensible or very badly organised presentation.