UCT MA Psychology

Multivariate Statistics

Test 1, September 2018

============================

**Instructions**

A. Submit answers as html or word doc or pdf, as compiled from R markdown or R script. Submit zip of entire project folder, in the usual way. Upload all to the VULA assignment page. Your report should attempt to attain full reproducibility. Please note that there is an allocation of marks for well formatted reports.

B. The deadline – strict - for submission is 21 September at 12h55.

C. Work entirely on your own. Do not consult other class members or indeed any other person. In all other respects this is an open book test.

D. Page length – strictly a maximum of 15 pages, 1.5 line spacing, font size 12.

**Questions**

The data sets for this test are located on VULA, in the Test 1 Assignment folder. The data set ‘Cambridge delinquency.sav’ file contains most of the data, except for data about convictions of the participants at different ages, which is in ‘conviction data.sav’.

The data come from one of the best known and described prospective longitudinal studies of crime and delinquency. The key authors over the last 57 years have been Farrington and West. Although they have continued to collect data since 1961 on the cohort of working class English participants, the data we have concerns the years 1961 to 1981.

It is a good idea for you to read one or more summaries of the study, and perhaps an overview article on the general problem: what are the early-life predictors of criminality? I have put two articles on VULA in the folder, for you, but you can read more widely on the question if you like.

There is a codebook for the data in the VULA folder. You will see that a great many variables were measured over the first 20 years of the study. The variable names in the data files are not very helpful, but the labels for the variables are.

The questions that follow are very broad and open-ended. This is deliberate. We want you to find your own way into the problem, and to use many of the methods and concepts we have introduced you to in the first six weeks of the course.

You may find the codebook and the sheer number of predictors in the data files a little intimidating. You are free to adopt multiple approaches in terms of which predictors to use for your analysis: you can make composites of the predictor variables, or choose ones that seem important; you can choose variables from one point in time or create variables to include all points in time, and so on. Your decisions could be informed by a little reading on the topic.

Please justify all decisions very clearly so we can follow your logic, especially in relation to model selection.

**Question 1**

Create a data set from the data files containing the variables that you want to use in your analysis. In other words, all the variables you want to use to understand and model the early life predictors of criminality. You must include information about convictions at different ages in your data set i.e., you need to merge the two data files, and select the variables and cases you need for the later analysis. Rename variables so that they are identifiable and usable. If you can do this programmatically i.e., use the tidyverse and other packages to manipulate the data files, and variable names, you will receive higher marks for this question.

10 marks

**Question 2**

Explore the data set, bearing in mind that our key question concerns what the early life determinants of criminality are. What is worth exploring further, do you think? Report several (but not too many!) graphs and tables – choose these prudently, and discuss each graph or table you include.

25 marks

**Question 3**

We wish to model the occurrence and extent of criminality in our participants on the basis of early life events or factors. Use all the considerations at your disposal (the description above, your data exploration in Question 2, your reading of the articles we have provided, etc.) to construct a statistical model, or even more than one model if you wish. Interpret your model(s). Write a function to assess the predictive accuracy of your model(s), test it, and apply it to your data.

(This question is deliberately quite open ended. Here are some things you may want to consider:

Model selection – hierarchical? Or stepwise? Or a penalised method?

Model validation – validation set? Or cross-validation?

Analytic method – OLS regression? Or logistic regression? Or discriminant analysis? Or tree methods?

Reporting of results – normal theory estimates, or bootstrapped estimates? Coefficient meanings?)

50 marks

**Question 4**

Format the final report so that it is clear, easily readable, and contains no (or a minimum of) warning messages etc. from R. The closer you get to <https://rstudio.github.io/tufte/> the higher the mark! (but don’t spend too much time on this, it is not worth a great deal). All your project and output files should also appear on Github, and you need to provide a link to the repository where the project is kept. The project should have received frequent and well annotated commits.

15 marks

======================