SAS – Analytics

Lab Report – Masters Level

December 2018

Instructions

You will be required to submit your lab report as a pdf. Put your matriculation number at the start of your document – **do not** include your name.

For this assessment you must submit your work as a report – there will be questions of interest that you must answer and discuss by conducting statistical analysis in SAS. Start your report with a brief introduction to the problem, clearly outlining the motivation for your analysis (i.e. discuss what the questions of interest are). Next, do an exploratory analysis for each question of interest a) to c) (no formal statistical tests or models at this point) – this might include plotting your data, looking at summary statistics etc in order to get an initial impression as to the conclusions of your questions of interest.

Following this, you will answer each question of interest a) to i) with a formal statistical analysis. Finally, give a conclusion summarising the overall results of your report. Since this is a lab report, ensure that any figures/tables etc are clearly explained/discussed and justify any of your conclusions. Do not leave it to the reader to hunt for information (walk them through any output or conclusion you make), you will be marked on the clarity of your report. You can either split the report into exploratory analysis and formal analysis **or** present the exploratory analysis for each of a)-c) alongside the formal analysis for each of a)-c).

For **each** question of interest, you will submit **two** things:

1) Relevant SAS output. This might include tables and plots from SAS containing the relevant results of whatever code you have written. Make sure your output is clearly titled, labelled etc. This should be done within your SAS code – **do not** edit this with other software.

2) Discussion of whatever analysis you have conducted.

Questions of interest f) and i) ask you to **additionally** submit your code – this will be clearly outlined in the question of interest, **do not** submit code for any other question of interest.

Deadline for submission: **11.55pm on Monday 21st January 2019.** Upload your document to MOODLE.

The **maximum** page limit for your submission is 24 pages.

Setup

The folder P:\SCIENG\MATHS\DATA\SAS Course\Analytics\Assessment contains the following dataset:

- weather.sas7bdat

Set up this folder as a permanent library in your SAS system. Since you are unable to write to the p-drive, ensure you have your own SAS folder assigned as a writeable library.

If you are using SAS online (SAS OnDemand for Academics), then you will see the library name under the course description heading. Since you will be unable to write to that location, ensure you assign a writeable library.

Data

The dataset weather.sas7bdat contains 4657 independent records of 24 weather-related variables for a European country. The variables in the dataset are described in weather_details.pdf.

Questions of Interest

a) A temperature for this country of 25^{o}C or above can alert officials to the possibility of a drought. Is the mean maximum temperature different to this?

b) Investigate whether there is a difference between the mean maximum temperature and the mean minimum temperature.

c) It is quite possible that evaporation is associated with the strongest wind gust direction for a day. Check whether there is a difference in the mean evaporation between the different strongest wind gust directions using an appropriate multiple comparison adjustment.

d) Find the "best" model (as determined by an automated model selection procedure), using evaporation as the response and all other variables in the dataset as potential predictors (do not consider any interaction terms, transformations or higher order terms, and **do not include** the variable "RainTomorrow" in the modelling process). When you come to fit a model for this question of interest, use two different directional automated procedures (fit two different models considering the same proposed predictors, using different directional selection approaches) and examine

whether they result in the same or different final models. If they are different, which do you prefer? Justify your answer.

e) Since the variables “WindGustDir”, “WindDir9am” and “WindDir3pm” have many levels, it is of interest to see whether some levels could be combined. Perform Greenacre’s method, with the variable “RainTomorrow” as the response, for each of these variables in turn. If, for any of the above variables, Greenacre’s method suggests combining levels is reasonable, combine the appropriate levels and use this(these) new variable(s) in any subsequent analysis (give any new variables **new** names).

f) Split your data for model assessment into training data (75% of original dataset), and test data (25% of original dataset), ensuring you have a roughly equal proportion of Yes to No responses for the variable “RainTomorrow” in each dataset. **In addition to output and discussion, include your code in your report for this question of interest**.

g) Using Hoeffding’s and Spearman’s statistics, check whether there is any evidence of non-linearity and/or irrelevant variables for the continuous variables in your dataset (use the training data).

h) Find the “best” model (as determined by an automated model selection procedure), using the variable “RainTomorrow” as the response and all other variables in the dataset as potential predictors. Use the training data and do not consider any interaction terms, transformations or higher order terms.

i) It is of interest to see whether a specific subset of the variables can be used to make accurate predictions about whether it will rain tomorrow. Consider the model that only uses the variables “RainToday” and “MaxTemp” as predictors. Investigate how the level of predictive performance to the test data differs using this model to that of your final model from part h). **In addition to output and discussion, include your code in your report for this question of interest**.