University of Glasgow

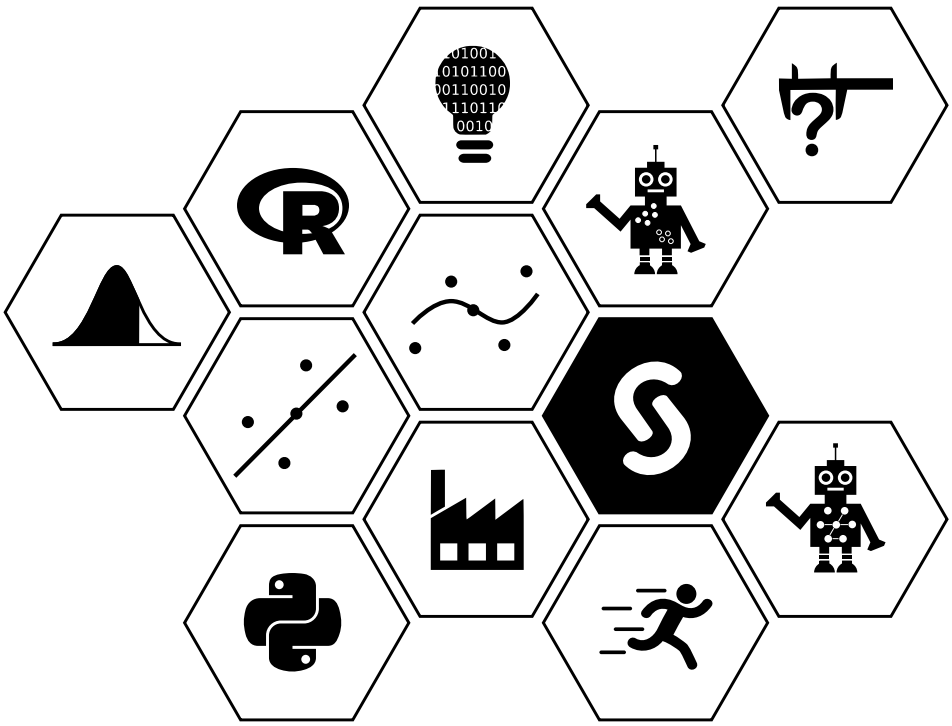
Data Management and Analytics using SAS

Benn Macdonald and Michael Waltenberger

Academic Year 2020-21

DMAS Assessment 4:

Timed Assessment



DATA ANALYTICS GLASGOW

Timed Assessment

Instructions

You should produce the code in a SAS editor to answer the questions in this assessment (you can submit a .txt file instead, if you prefer). Some questions may ask you to state some information instead/in addition and you should include this as a SAS comment in your editor (or .txt file). Questions will clearly indicate if you are to submit code, a comment or both.

Your submission should consist of 1 file only (either a .sas or .txt file), with all your work contained therein. Please ensure that you submit before the timer for the assessment finishes - you will not be able to submit after the timer expires. The timer will be 2 hours - the exam time is originally 1 hour, which is doubled to 2 hours to account for downloading the questions, uploading your submission and dealing with any technical issues that may arise. Please don't leave it to the last minute to upload your submission, give yourself plenty of time to do this.

For ease of marking, write in your SAS editor (or .txt file) in a SAS comment the number of the question you are currently answering.

You may use your DMAS notes and/or SAS help documentation.

Save your script regularly throughout your assessment - SAS might timeout (online version) or could potentially crash.

Data

You have been given one dataset for this assessment: power.sas7bdat

This dataset consists of 1470 independent records of variables related to competitive powerlifters. The variables in your dataset are given in the following table:

Variable Name	Type	Description
ID		Participant ID
Winner	Categorical (2 levels)	Whether a competitor won their last event: Y = Yes, N = No
Equipment	Categorical (4 levels)	Hand covering type: Single-ply, Multi-ply, Wraps, Raw
Sex	Categorical (2 levels)	Sex assigned at birth: M = Male, F = Female
AverageTime	Assume continuous	Average time (minutes) per training session
Age	Assume continuous	Age (years)
LiquidConsumed	Assume continuous	Average liquid consumed (ml) per training session
BodyweightKg	Assume continuous	Bodyweight in kilograms
BestSquatKg	Assume continuous	A competitor's best squat in kilograms
BestBenchKg	Assume continuous	A competitor's best bench press in kilograms
BestDeadliftKg	Assume continuous	A competitor's best deadlift in kilograms
GymCost	Assume continuous	Monthly amount paid ($) for all gym costs (including membership, products etc)
Displacement	Assume continuous	Distance (cm) competitors pushed a weighted object (positive or negative depending on direction)
Wilks	Assume continuous	Wilks score (strength adjusted for body mass)

You should place this dataset in a location that the version of SAS you are using has access to. If you are using SAS OnDemand for Academics, you should upload this to a permanent folder **you have created** in "Server Files and Folders" (right click on the folder you want to upload to and select "Upload Files…"). Once you have uploaded the dataset, you can get the path needed for your libname by right clicking your folder and selecting "Properties" (see "Location").

Questions

Note: all questions refer to the dataset power.sas7bdat

1) Find out if there is a statistically significant difference in the average bodyweight between those competitors that won their last event and those that did not and produce a corresponding confidence interval plot. Submit your code used. **(2 marks)**

2) Interpret the results from 1) and comment on the validity of your modelling assumptions. You should submit a comment in your editor answering this question. **(2 marks)**

3) Fit a binary logistic regression with Winner as the response and all other variables (except ID and Wilks) as explanatory variables. Submit the code you used to fit this model. **(2 marks)**

4) Now perform stepwise automated model selection with significance levels of 0.05 used as the selection criteria, using your model from 3). Submit both your code used to do this and a comment in your editor interpreting the outcome of your final model (you can assume that your modelling assumptions are valid). **(2 marks)**

5) Using a PROC SQL step, find out how many of the competitors were female. Submit your code you used to find out this information. **(1 mark)**

6) Without using a %let statement, put the number of females in the dataset into a macro variable (this step should not print anything to the results window). Now produce only one figure containing boxplots (only) of gym costs paid by females separated by their choice of equipment type. The y-axis should be called "Gym Cost ($)" and the x-axis should be called "Equipment Type". The values on the y-axis should be displayed with dollar symbols ($) next to them. Include an appropriate title that also uses the macro variable you produced to state the number of females (ensure there is no unnecessary whitespace). Submit all your code used to answer this question. **(6 marks)**

7) Use a PROC other than PROC MEANS to find out the highest Wilks score of the competitors. Submit your code you used to do this and state the Wilks score in a comment in your editor. **(2 marks)**

8) Produce a plot of Spearman vs Hoeffding ranks using Wilks score as a response variable and all other continuous variables in your dataset as explanatory variables (AverageTime, Age, LiquidConsumed, BodyweightKg, BestSquatKg, BestBenchKg, BestDeadliftKg, GymCost, Displacement). Submit all your code you used. **(6 marks)**

9) Using your plot in 8), which (if any) variables show evidence of a non-linear relationship with the response variable? Submit a comment in your editor answering this question and ensure you justify your answer. **(2 marks)**

10) Write one data step that converts age to character. The entries of this character variable should read "Grp 1", "Grp 2" or "Grp 3" (no leading or trailing whitespace), if the numeric variable was less than 30, between 30 and 50, or over 50, respectively. This character variable should be called age. Submit your code (note: it should consist of only one data step and nothing more). **(3 marks)**

Total: 28 marks