SAS Analytics Assessment

December 2018

Instructions

This assessment has two parts. Part One looks at data from competitive powerlifters. Part Two looks at features of orthopaedic patients.

You will be required to submit your assessment as a word document or pdf. Write your matriculation number at the top of your document and for the document name – do not include your own name. Some questions may ask you to submit code, some may require you to submit plots, some may ask you to comment on your results or some combination of these – make sure you read each question carefully, what you are required to submit will be in bold. The **maximum** page limit for your submission is 15 pages.

Deadline for submission: **11.55pm on Tuesday 18th December 2018.**

Setup

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

- powerlifting.sas7bdat
- orthopaedics.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 names under the course description heading. Since you will be unable to write to that location, ensure you assign a writeable library.

PART ONE: Power Lifting

This part uses the dataset powerlifting.sas7bdat

a) Calculate a new variable, *TotalKg*, which contains the sum of a person's best squat, best bench and best deadlift. Produce descriptive statistics and a histogram of this new variable. **You should submit your code and histogram**.

b) Using an appropriate one-sample test, determine whether there is evidence that the total weight lifted (*TotalKg*) is different to 400kg. You should produce diagnostic plots and a confidence interval plot which includes a reference line at the null hypothesis**. Submit your code and confidence interval plot.**

c) Investigate whether the total weight lifted (*TotalKg*) is different between males and females, using descriptive statistics and an appropriate test, producing a confidence interval plot. **You should submit your code**.

d) From your result from part c), do you reject the null hypothesis or not? **Submit a sentence or two answering this question (and justifying your answer).**

e) Check that the assumptions of your test from part c) are valid. **Submit your plots and relevant tables from your SAS output, and include a sentence or two discussing the validity of the assumptions.**

f) Using an analysis of variance, check if there is a statistically significant difference in the average age of a competitor and their choice of equipment. Ensure you check your modelling assumptions. **Submit your code and plots, and include a few sentences commenting on the conclusion of your test and the validity of the modelling assumptions.**

g) Perform a post-hoc analysis on your model from f), adjusting for multiple comparisons and produce a suitable plot for your chosen multiple comparison method. **Submit your code, plots, and tables and include a few sentences concluding your analysis.**

h) Using PROC REG, fit a linear model with Wilks as the response and age as the predictor. Ensure you check the modelling assumptions. Does it look like Wilks is associated with age? **Submit your code and a sentence or two justifying your answer**.

i) Create a new variable, *Place2*, which is a categorical version of the variable *Place*, where any observed value not equal to 1, 2 or 3 is coded as "Other". Your variable *Place2* should therefore have 4 levels: 1, 2, 3 and Other. **Submit your code.**

j) Fit a linear model with Wilks as the response and the following as explanatory variables; Place2, LiquidConsumed, Sex, Equipment, GymCost, Age, BestSquatKg, BestBenchKg, BestDeadLiftKg, Schedule, and AverageTime. Produce parameter estimates and confidence intervals, and check your modelling assumptions. **Submit your code and plots.**

k) Implementing the method of automated backwards model selection, reduce your model from j) using Schwarz' Bayesian information criterion. Ensure you produce p-values for your parameter estimates of your model and diagnostic plots of how the coefficients, fit criteria and average squared errors behave when adding new variables. **Submit your code and comment on the association(s) between the response and explanatory variable(s) in your final model.**

l) For your model from part k), use an appropriate procedure to produce diagnostic plots to assess your modelling assumptions. **Submit your code and a couple of sentences giving your conclusions.**

PART TWO: Orthopaedics

This part uses the dataset orthopaedics.sas7bdat. The event of interest in this dataset is whether a patient's condition is classified as normal or abnormal (variable *class*).

m) Using an appropriate test, determine whether there is a statistically significant difference in the degree of spondylolisthesis, on average, between those with a condition classified as normal and abnormal. **Submit your code, and a sentence or two answering the question and discussing the validity of the modelling assumptions**.

n) Divide your data into training and test data (75% should go to your training data and 25% into your test data), ensuring you have an equal proportion of events to no events in each. **Submit your code**.

o) Fit two binary logistic regression models on your training data. The first should use the event of interest as a response and include all other variables as explanatory variables. The second should be the same, but do not include the degree of spondylolisthesis as an explanatory variable. In both, ensure you use "Normal" as the reference category for the response variable. Produce confidence intervals and plots of the odds ratios and predicted probabilities from the models. **Submit your code, plots and your comments on the conclusions and validity of each of your models.**

p) Using your test data, overlay two ROC curves, one for each of your binary logistic regressions from part o), and use this to inform whether the degree of spondylolisthesis should be included in your final model. **Submit your code, plot, table and a couple of sentences concluding the question.**

[End of assessment]