QBUS2810
Statistical Modelling for Business

Semester 1, 2021

Group Assignment Task 3

This group assignment task 3 will contribute 25% towards your final result in the unit. The deadline is Wednesday, June 2^{nd} by 11:59pm. Submission is via Canvas and Turnitin.

This assignment must be completed in groups of 3 (unless special permission is granted, in writing, by Nethal Jajo)

Maximum Length: There is no maximum page length for this assignment. If you have something interesting and worthwhile to include, then please do so without worrying about a page limit. However, irrelevant or overly long-winded material will reduce your overall mark. As a guideline, I expect the typical report to have between 20-25 pages, excluding Python code.

Notes on Marking:

- The assignment will initially be marked out of 55.
- Up to an additional five (5) marks will be awarded based on the overall presentation quality of your report. Thus, you will receive a total mark for this assignment out of 60. You will lose some of these 5 presentation marks for poor, inefficient, unclear and/or unprofessional presentation. You will be rewarded for professional, efficient and clear presentation methods. I expect your final report to be done in a professional editing package and to be submitted in pdf only. Html files of jupyter notebooks are not suitable.
- You must use Python for this assignment. You are being assessed on how well you can use Python to complete the assignment tasks. NB: You can use Excel for simple data manipulations and clean-up; but Python is better at these

tasks too! All plots and statistical output in the assignment must have been produced in Python, though you can of course make nicer tables in a text editor to include in your assignment. Please include an appendix in your assignment that contains the Python code your group used to produce ALL outputs in your assignment. A heavy penalty will apply if the Python code is not supplied (or the code supplied does not run or work when the marker tries to run it).

Pre-analysis Instructions for data:

Please include the python code in the Jupyter notebook file "grp_assnt_gendata.ipynb" in your Jupyter notebook file to input and clean the data . Collect the student ID numbers for the members of your group and then add these numbers together. Input the result into the python code where instructed. Run the subsequent code to generate two datasets: "train" and "test". Most analysis you do will only use the "train"data set. Any forecasting your group does will only use the "test" dataset. The purpose of these commands is to ensure that each group receives different randomly selected datasets for "train"ing and "test"ing purposes. Two other python codes are included in case you need it: forward_selection.py and backword_selection.py

Business problem:

The California Department of Education (www.cde.ca.gov) is interested in the effect of school class sizes on test scores for school age children. They also wish to build a model that can accurately predict test scores for school age children so they can better understand the drivers of student academic performance. Your group has been commissioned to research on and analyse the data provided and then report back to the Department of Education, principally regarding the two major goals they are interested in.

Data and Description:

Please see the file CaliforniaTestScores.pdf for information on the data collected. The data used here are from all 420 K-6 and K-8 districts in California collected over a two year period. The dataset is in the file "caschool.xlsx". Please see CaliforniaTestScores.pdf for description of the variables in the study and for more information. The measure of academic success via test scores to be used is the average of the reading and math scores on the Stanford 9 standardized test, administered to 5th grade students, averaged over each school district. The class size is given by the total number of students in the district, divided by the total number of teachers in that district. There are 420 school districts in the final sample.

Goals and primary questions:
There are three primary goals that the California Department of Education would like your group to focus on:

(a) Develop an optimal model for predicting the average 5^{th} grade reading and math test score;

(b) Understand the relationship between class size (as student teacher ratio) and the average 5^{th} grade reading and math test score;

(c) If the department is awarded some extra state or federal funding, what aspects should they focus on, and spend money on, to best improve the average reading and math score for 5^{th} graders.

Tasks:
1. Conduct a suitable exploratory analysis on this dataset that is relevant to the goals of this study (5 marks).

2. Analyse and test for a relationship between class size and average test scores. Include a discussion of whether the assumptions of your analysis and test could hold for this data and whether and how strongly the data actually fits the model. (5 marks)

3. Discuss which variables in the dataset, could be causing omitted variable bias in your analysis in part 2, and justify clearly why you think that. Include these omitted variables, together with class size, in a multiple regression model. Fit the model, without any transformations or interactions or nonlinear effects. Again, test for a relationship between class size and test score. Also include a discussion of whether the assumptions of your test could hold for this data and whether and how well the data actually fits the model. Also discuss the level and sources of multi-collinearity present in the data and your analysis and whether you think this is problematic, or not, and why; and if so, problematic for what? (10 marks)

4. Conduct a variable and model selection exercise, including at least two potential interaction effects as potential predictors and also at least two transformations/nonlinear effects as potential inclusions. You must properly motivate and discuss your choices here. Then, report a summary of the comparison of fit over at least 8 different models/transformations/variable sets that you tried, all while forcing class size to stay in the model in some form. The goal is to find an optimal model that is highly accurate, but also parsimonious, to predict and explain the average test scores. Finally, fully report and give diagnostics on the final optimal model, as well as briefly discussing any collinearity issues it may have. Also, if there are any nonlinear effects in this model, clearly discuss and illustrate their effects on test score. (15 marks).

5. Discuss your results and conclusions regarding the overall goals of this study, in light of the results from your overall analysis of the "train" dataset. Be technical but clear here. Also, include a prediction of what would result if the student-teacher ratio was reduced by 1 student, on average in each district, using at least the optimal model so far. (5 marks)

6. Using (at least) the 5 best model specifications considered so far (and any others you think relevant), generate forecast predictions in the"test" dataset for the average 5th grade reading and math scores. Present a summary table, and suitable plot(s), of the forecasts and their accuracy for these models, using the forecast measures RMSE, MAD and forecast R^2. Re-discuss your results and conclusions regarding the overall goals of this study, in light of these results and your overall analysis. Be technical but clear here. (10 marks)

7. Write a final report, in as close to plain English as is practical and possible, that discusses and summarises your analysis above and gives conclusions on the overall goals of this study. Address the report to, and write it at a level appropriate for, the Department of Education in California. Include in your report a prediction of what would likely occur if the Department spent money on hiring more teachers, so as to reduce class size, and whether you recommend they take that action, or not; plus any suggestions for how the Department could further assist school age children in improving their average maths and readings scores and also any suggestions, if you have any, for any future studies they should do to facilitate that. (5 marks)