MA 542 S' 17 Test 2 Practical

This test is open-book open-note. You may not use a cell phone. You must create and upload a report of your work in Canvas.

Part II: Practical Multiple Regression

The data set World_Demographics (found as the SAS data set sasdata.World_Demographics, or in Canvas as the text file World_Demographics.txt) contains six demographic variables recorded for 232 countries. The demographic variables are

- y: death rate: average number of deaths per year per 1000 population
- x_1: median age of the population
- x_2: health expenditures as percent of GDP
- x_3: birth rate: average number of births per year per 1000 population
- x_4: education expenditures as percent of GDP
- x_5: obesity rate: percent of adult population considered to be obese

The data were randomly divided into a training set of 182 countries and a validation set of 50 countries. The variable *set* indicates to which set a country belongs. (NOTE: if it is more convenient for you, the data sets World_Demographics_t (in sasdata.World_Demographics_t or in Canvas World_Demographics_t.txt) and World_Demographics_v (in sasdata.World_Demographics_v or in Canvas World_Demographics_v.txt) contain the separate training and validation sets.)

Your task in this test is to fit the best regression model you can relating the response, y (death rate), to the five predictors (the other five demographic variables), using the 182 training cases only, and then to assess its performance on the validation set.

1. **(10 points)** Conduct an initial visualization of the data. Describe the patterns you see.
2. **(15 points)** The visualization should suggest a nonlinear relation between death_rate and some of the predictors. To account for this, obtain an initial fit using linear and quadratic terms for all five predictors. Evaluate the multicollinearity and take remedial action if necessary.
3. **(10 points)** Which regressors seem important, and which unimportant in the fit?
4. **(15 points)** Using appropriate diagnostics and measures, assess the fit of the full model.
5. **(10 points)** Using one or more model selection methods, obtain one other model for these data. Evaluate the selected model using the criteria from questions 3 and 4 above. Compare this model with the full model.
6. **(10 points)** Now evaluate the performance of both the original and the full model on the validation set. How does the variation in prediction on the validation set compare with the estimate of error variance for your subset model?