STA 108 --- TERM PROJECT

Due: Friday, Dec 18, 2020 at noon

Select an 80% subset for your analysis.

The data set "countries.csv" contains information on the following variables:

- Country – list of countries in the data set
- Code – three letter code
- LandArea – land area in square kilometers
- Population – population in millions
- Rural – percentage of population living in rural areas
- Health – % of government expenditures directed towards health care
- Internet – % of population with internet access
- BirthRate – Births per 1000 people
- ElderlyPop – % of population at least 65 years old
- LifeExpectancy – Average life expectancy in years
- CO2 – CO2 emissions in metric tons per capita
- GDP – Gross Domestic Product per capita
- Cell – Cell phone subscriptions per 100 people

We are interested in finding a parsimonious model to predict life expectancy. Use the tools we have learned in this course to

1. Build a model with LifeExpectancy as the outcome and any of the remaining variables as predictors.
2. Carry out a residual analysis to identify
    - Deviations from linearity in any of the predictors
    - Possible transformations of predictors
    - Possible transformation of the outcome variable
3. Assess the potential for multicollinearity
4. Identify which variables are predictors of LifeExpectancy using suitable model selection algorithms.

Submit your analysis in the form of a written report that should contain:

1. An introduction
2. A results section that states the model you selected and interprets the results in context, explaining which variables predict life expectancy and how teach affects it.
3. A model building section where you describe your approach to finding your model
4. A brief summary of your project
5. The maximum number of pages allowed is 5, including graphs.
6. Your R code should go into an appendix as should any incidental plots. Plots that are essential to your model justification go into the main report.