DS5110 Homework 4

Kylie Ariel Bemis

7 March 2022

Instructions

Your solutions should include all of the code necessary to answer the problems. All of your code should run (assuming the data is available). All plots should be generated using `ggplot2`. *Make sure that you answer all parts of the problem.*

Submit your solutions on Canvas by the deadline displayed online. For full credit, your submission must include exactly two files:

- R Markdown (.Rmd)
- PDF report (.pdf)

Problems must appear in order, and problem numbers must be clearly marked. Any written responses should appear outside of code blocks and use Markdown for text formatting. Code comments are encouraged, but will be ignored for grading purposes. Solutions that are especially difficult to grade due to poor formatting will not receive full credit.

All solutions to the given problems must be your own work. If you use third-party code for ancillary tasks, you **must** cite them.

Part A

Problems 1–3 use the complete Gapminder dataset. Download the data files from "ddf–gapminder–systema_globalis-master.zip" on Piazza. The original datasets can be found on Github (https://github.com/open-numbers/ddf--gapminder--systema_globalis). The data is divided into tables of "entities" and "datapoints". The "entities" tables provide dictionaries of terms, countries, regions, etc. The "datapoints" tables contain single variables measured for each country and year.

Problem 1

We would like to build a model for predicting life expectancy. Create a data frame that includes only complete cases (no missing values) and includes columns for country code, year, and the following response + predictors:

- Life expectancy (years)
- Infant mortality rate (per 1,000 births)
- Murder (per 100,000 people)
- GDP per capita (US$ adjusted for inflation)
- Medical doctors (per 1,000 people)
- Poverty rate (people below $5.50 a day)

Visualize life expectancy versus the five candidate predictors, transforming variables as necessary, and describe their relationships.

Problem 2

Build a linear regression model for life expectancy using a single predictor, justifying your choice based only on the visualizations from Problem 1. Then use residual plots to perform model diagnostics.

Comment on any outliers or violations of model assumptions you notice in the residual plots. If necessary, fix the issue, re-fit the model, and perform model diagnostics again.

Problem 3

Use residual plots to determine if any other candidate predictors should be added to your model from Problem 2. If so, add up to one additional predictor to the model, and then perform model diagnostics on the new model.

Part B

Problems 4–5 continue to use the same dataset from Part A.

Problem 4

Using the full dataset (minus any outliers you removed), perform reproducible 10-fold cross-validation on your model from Problem 3. Report the cross-validated RMSE, as well as the RMSE of the original model from Problem 3 on the data originally used to train it. Which RMSE is larger? Is this surprising, and why?

Problem 5

Reproducibly partition the dataset (minus any outliers) into a training, validation, and test set using a 50/25/25 split. Keeping any transformations you found to be appropriate in Problem 1, perform stepwise model selection to build a predictive model for life expectancy using RMSE as the selection criterion. Show the RMSEs at each step and note which variable is being added/dropped, and then report the RMSE of the selected model on the test set.