Data Science Toolbox Portfolio Questions

02 Regression and Statistical Testing

Daniel Lawson

Block 1

Portfolio 02

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See https://dsbristol.github.io/dst/assessments.html for advise.

Question R02.1: You are tasked with making a temperature prediction for 2040 based on the Temperature Data used in Workshop 2.3. Design a cross-validation setup that could be used to obtain predictions along with uncertainty quantification, carefully describing its advantages over what is presented above, and its limitations. You may wish to investigate standard forecasting methods.

Question R02.2: It was claimed without proof that the leave-one-out cross validation error can be cheaply computed for linear regression as:

$$CV = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{e_i}{1-h_{ii}}\right]^2,$$

where $e_i = y_i - \hat{y}_i$, $\hat{y}_i = \beta X_i$ and h_{ii} is the diagonal entries of the hat matrix. This also works for penalised regression, to come later. Consider the proof presented in https://robjhyndman.com/hyndsight/loocv-linear-models/ or otherwise, and rewrite this proof with simple annotations for an Undergraduate audience. Briefly discuss the implications of the theorem for both datasets.

Question R02.3: Consider the final non-linear stepwise model that was obtained for the diamond data (the object called `modelcvintstep` and named `intstep`). It has the highest R^2 with the test data, and highest AIC of all models considered. Investigate and discuss the ways that this model may be considered *best* and how it may yet be bettered by other models considering the same model space and data (i.e. all pairwise quantitative features plus the ordinal factors). Discuss what interpretation we can make on the linear and non-linear effects of the parameters.