Data Science Toolbox Portfolio Questions

02 Regression and Statistical Testing

Daniel Lawson — University of Bristol

Block 2

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 The Assessment Page for advice.

These questions may make reference to the content from the current block.

Question R02.1: It was claimed without proof in Lecture 02.1 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 the Temperature and Diamonds datasets from Workshop 2.3.

Question R02.2: Consider the paper Model-agnostic out-of-distribution detection using combined statistical tests. What are the key reasons that justify statistical testing over standard machine-learning, and how does it relate to our course content? Provide an experimental design to test whether these results hold in practice for the experiments performed in Diamonds data from Workshop 2.3.

Question R02.3: Imagine that you are tasked with making a temperature prediction for 2040 based on the Temperature Data used in Workshop 2.3. Design and justify 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.