Authors: Wojciech Ptaś (7042843), Mhd Jawad Al Rahwanji (7038980)

Problem 1 - Cross-Validation

1. The value of *k* dictates the number of times we divide the data and by extension the number of cross validation folds/rounds. 1) A *k* value of 1 is the same as the validation set approach. 2) Whereas, a value of *k = n-1*, where *n* is the number of observations in the dataset, is equivalent to LOOCV. 3) A value of 5 or 10 for example is a common middle ground. In 1) we have high bias low variance, in 2) we have low bias high variance (but averaged) and in 3) we have ok bias and variance and it's less expensive to compute.
2. Leverage can be defined as how different/far sample *i* is from the expected value or other samples, in terms of its independent variables. A high leverage sample has an advantage over the rest, the same way wrenches with long handles have more torque. In terms of regression, it can either be consistent with the overall pattern or not. Assuming it isn't, it is capable of skewing the fit towards it, influencing/changing the prediction and increasing bias, then it's called a high influence point. Removing such a sample would result in a less biased model that better fits the data (i.e. less MSE). To measure the influence we use Cook's D.
3. We have the entire dataset, the responses, the dataset without a sample *i* (LOOCV) and the responses without the corresponding response to sample *i*:

$$X \in R^{n x p}, y \in R^{n}, X_{-i} \in R^{n-1 x p}, y_{-i} \in R^{n-1}$$

Now we fit least squares once on the entire dataset:

$$\widehat{\beta} = (X^{T}X)^{-1}X^{T}y$$

Now what would be our fit had we fit least squares with LOOCV:

$$\widehat{\beta}_{-i} = (X_{-i}^{T}X_{-i})^{-1}X_{-i}^{T}y_{-i}$$

Now we have the hat matrices for both cases:

$$H = X(X^{T}X)^{-1}X^{T}$$

$$H_{-i} = X_{-i}(X_{-i}^{T}X_{-i})^{-1}X_{-i}^{T}$$

Our predictions would now look like this:

$$\widehat{y} = x_{i}^{T}\widehat{\beta}$$

$$\widehat{y}_{-i} = x_{i}^{T}\widehat{\beta}_{-i}$$

Since the prediction in least squares is a projection of the response using the hat matrix. It is possible to use the hat matrix from the fitted least squares over the entire dataset to simulate predictions without a given sample by removing its influence when projecting.

$$\hat{y}_{-i} = \sum_{i \neq j} H_{ii} y_i + H_{ii} \hat{y}_{-i}$$

$$\hat{y}_{-i} = \sum_{j=1}^{n} H_{ij} y_j - H_{ii} y_i + H_{ii} \hat{y}_{-i}$$

$$\hat{y}_{-i} = \hat{y}_i - H_{ii} y_i + H_{ii} \hat{y}_{-i}$$

Now our prediction error:

$$y_i - \hat{y}_{-i} = y_i - (\hat{y}_i - H_{ii} y_i + H_{ii} \hat{y}_{-i})$$

$$y_i - \hat{y}_{-i} - H_{ii} y_i + H_{ii} \hat{y}_{-i} = y_i - \hat{y}_i$$

$$MSE_i = y_i - \hat{y}_{-i} = \frac{y_i - \hat{y}_i}{1 - H_{ii}}$$

Now going back to the original formula for calculating the average MSE in CV:

$$CV_{(n)} = \frac{1}{n} \sum_{i}^{n} MSE_i$$

$$CV_{(n)} = \frac{1}{n} \sum_{i}^{n} \frac{y_i - \hat{y}_i}{1 - H_{ii}}$$

This proof was heavily inspired by [1].

References:

1. User "krisrs1128" 2012, *Linear Regression LOOCV Trick*, Notes of a Statistics Watcher, accessed 5 December 2022, <notesofastatisticswatcher.wordpress.com/2012/12/18/linear-regression-loocv-trick>