BUAN 6357 (Johnston)
Homework 4A (20230330)
Code Due: 8 April 2023 (6PM)

Points available: 90.

This assignment is cross-validation with Ordinary-Least-Squares. Use "data(airquality)" for this analysis. The outcome variable (dependent) is Ozone. Seed the RNG with 504737137 for each analysis section which requires the use of random numbers.

For this assignment you will need the package "tidyverse" and "data.table". You should not use any additional packages. You should use only the "require()" or "library()" statement in your code. Any use of the install.packages() function in submitted code will result in a score of 0 for that submission. Multiple instances of concurrently submitted code will result in only the most recent code being evaluated and scored on the assumption that you realized the earlier submitted code contained a problem which was fixed and then re-submitted.

The first commands of your code submitted for grading to eLearning MUST be:

setwd("c:/data/BUAN6357/HW_4"); source("prep.txt", echo=T)

and the last command of your code (which actually provides the scores) MUST be:

source("validate.txt", echo=T)

Be careful with the quote characters as they must ALL be the same at the beginning and end of a string. (Use the single or double quote character from the key close to "Enter".) Inclusion of these lines is required BEFORE your code will be tested. Submit the code to eLearning as an ASCII file which can be copied directly into R. Files with the ".rtf" extension will not be evaluated.

You may submit HW 4A as many times as needed until you get full credit. (Then you should stop.)

Deliverables:

1. seed (vector) random number generator seed
2. raw (data.table) see Notes
3. base.resid (lm) non-CV OLS model residuals (all observations)
4. tst (vector) 10% validation set index values
5. cv.trn (vector) training set predicted values from OLS simple CV
6. cv.tst (vector) tst set predicted values from simple CV model; see Notes
7. cv.resid (data.table) tst residuals from simple CV; see Notes
8. jk.resid (data.table) validation residuals from LOOCV; see Notes
9. kf.resid (data.table) validation residuals from K-fold; see Notes

HW 4B will direct you to explore both the intermediate results, the deliverables from Part A and answer questions about each of them. You may submit answers to HW 4B as many times as you wish but only the score for the last submitted code will be retained.

Notes:

- raw: retain only complete case observations; drop variable "Day"; convert variable "Month" to factor
- residuals: actual – predicted
- cv.resid, jk.resid: two variables, "loc" (index) and "diff" (residuals)
- Use k=10 in K-fold CV.
- kf.resid: three variables, "k" (group), "loc" (index), and "diff" (residuals)
- cv.fitted values are generated using validation data and predict()
- variables "k" and "loc" are for tracking grouping and original observation location to aid in debugging if needed; in jk.resid "loc" is also the group as all groups are of size 1
- Re-start the RNG for each Cross Validation strategy.