BUAN 6357 (Johnston)
Homework 5 (update)
Code Due: 30 March 2019 (6PM)
Part B Due: 31 March 2019 (11:59PM)

Points available: 80.

This assignment is about generating and evaluating cross-validation error estimates. Use the "Concrete_Data_wj.csv" file from UTDbox>demo as a starting point. (This data file will be available in the working directory when you submit your code to eLearning.) Part A involves data import, modeling, error calculation, and observation tracking across several cross validation (CV) techniques. Part B will require the summary of calculated errors, potentially both across all sub-groups and within each subgroup, for each CV technique. Be prepared to perform both parametric and non-parametric summary operations on these calculated CV error values.

For this assignment you will need the package "data.table". You will not need 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.

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

setwd("c:/data/BUAN6357/HW_5"); source("prep.txt", echo=T)

and the last command of your code 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 next 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.

The Shiny app to evaluate your code is not available for this assignment.

You may submit this assignment as many times as needed until you get full credit.

The model being evaluated for this assignment is an OLS using the variable “strength” as outcome.

Each deliverable is to be a data.table object with the listed components presented in the given order. After reading the data, you should add a tracking value (idx) to the data table in the form: “1:nrow(raw)”. This variable will be part of many deliverables. Any deliverable involving multiple passes of the data will require results from each pass and an additional variable whose name starts with “iter” or “grp”.

Reset the RNG seed before each sampling and/or CV technique. Use 654432970 as your seed value. Use the sample() function for sampling and randomization.

Deliverables (data tables):

1.	raw	original data, as read, with addition of “idx”
2.	base_m	baseline (non-CV) model residuals: resid_base, idx_base
3.	simple_m	simple CV (10%) training results: resid_simple, idx_simple
4.	simple_cv	simple CV (10%) test results: resid_cv, idx_cv
5.	jk_m	LOOCV training results: resid_jk, idx_jk, iter_jk
6.	jk_cv	LOOCV test results: resid_jk, idx_jk, iter_jk
7.	k_m	k=10 training results: resid_k, idx_k, grp_k
8.	k_cv	k=10 test results: resid_cv, idx_cv, grp_cv

Part B of HW 5 will direct you to explore both the intermediate results and the deliverables from Part A and answer questions about each of them. You may submit answers to HW 5 part B as many times as you wish but only the score for the last submitted code will be retained.