Consider the HIV drug resistance mutation data described on coursespace.

1. Load the data using the code given in slides #6.

*You need to download the helperfunction.R file from course space, specify correct work path so that R can find this file on your computer.*

*The code in example of slides #6 load data for only one drug, you need to run it multiple times to load data for all five drugs 3TC, ABC, AZT, D4T, DDI*

1. Compare combination of the following method using 10-fold cross validation, and prepare 3 tables as in excel file.

One table for MSE defined as average of over all isolates and all drugs.

One table for average bias, defined as average of

The last table for variance defined as

is the cross validation predicted values, and is the observed values of IC50 ratio. The index i is for isolate, and index j is for drug.

*See R example/tutorial of how to conduct cross validation in slides #5. You can also refer to the R code provided by owner of data, see the link in data description posted on coursespace.*

1. Comments on comparison results you got. If you could try to explain some finding using theoretical calculations. Theoretical calculation part is bonus.
2. Design additional tables/figures to present your results better when needed.

The assignment will be marked according to “presentation is clear and precise”, “results are correct and reproducible”, “Code are well written”, “Novelty and quality in your optional work”

Here are some suggestions:

1. Describe your findings in the beginning (like abstract of paper), and then give the details.
2. Write your code well (try to use functions to module your code, avoid using global variables if possible) with comments to explain what you are doing.
3. Results are reproducible by just running the your code.
4. I suggest you to use R markdown, but it is optional.

I expect the summary/discussion about your analysis results will separate good grade and excellent grade. To help you summarize your results, I give some questions that you could try to answer from result of your analysis. You could also add discussion of your other findings

1. Which regression methods consistently works better than others?
2. Does stacking really always do something? When not, try to figure out why
3. Does improved stacking really improve performance of predictions? Try to answer why