**Final Exam 599, 05/09/22**

We will use the ICARE rehabilitation data (Winstein, Jama 2016). The goal is to predict the in an arm and hand impairment score (Upper extremity  Fugl Meyer) after training FM2 as a function of multiple baseline variables including baseline FM1.

Note 1: these data are not publicly available, so PLEASE DO NOT DISTRIBUTE outside of this class.

Note 2: the actual data set has actually many more observations but we deleted the rows with missing data

**Data structure**

Y = FM2 is the dependent variable

X = all baseline/demographic data are the independent variables  - the meaning of the variables are given below

FM1 = hand Fugl Meyer at baseline

FM2 = Fugl Meyer after training

CHAMchallenge = a motivation question in the confidence in arm and hand test

ave\_CAHM =   average confidence in arm and hand test

EQ\_index =     generic health status questionnaire

SIS\_hand = Stroke impact scale of hand function

RNLIadj = A reintegration to normal living index

NIHtot = NIH stroke impact scale ; A brief assessment of physical function post-stroke

log\_mean\_time\_MA\_PA = Time on the Wolf motor function; more affected hand

log\_mean\_time\_LA\_PA = Time on the Wolf motor function; less affected hand

grip\_MA = grip strength, more affected

grip\_LA = grip strength, less affected

dose\_hours  = actual dose of training

onset\_to\_rand = time since stroke at start of trial

age\_at\_rand   = age at start of trial

old\_stroke = whether participants had a stroke or not prior

Q1/ Interpreting and plotting a linear model output (10 points)

Create a binary variable by comparing log\_mean\_time\_MA\_PA to its mean. Make a regression model with two predictors: FM1 and this new binary variable (interactions). Compute and present the two slopes and intercepts from the model output. Using these values draw the regression lines of FM2 as a function of FM1 for the two cases when log\_mean\_time\_MA\_PA is high and low. Add the data for the two cases with different markers. Add a complete legend.

Q2/ Estimating regression coefficients with a manually implemented bootstrap. (30 points:: 10 + 20)

1. Estimate 95% confidence interval via the “typical” least square regression.
   1. Perform a cross-validated Lasso to find the best lambda. Plot the CV MSE across values of lambda. Use the lambda corresponding to 1 SE to fit lasso a model to predict FM2 from the data. Print the coefficients.
   2. Re-run the “typical” regression model (`lm()`) using only the lasso-selected variables with no interactions. Report the 95% confidence interval of the parameters using the R output (which uses the formulas in the text).
2. Estimate 95% confidence interval via the bootstrap.
   1. Now implement a bootstrap “**manually**” (i.e., write the code to do the sampling) without any boot function from R) to estimate the 95% confidence interval of the parameters again with only the lasso-selected variables in the model. Plot the histogram of the parameter distributions for FM1. Use percentiles to get the 95% CIs for all parameters. Make sure your results do not depend (too much) on your number of samples by generating many samples
   2. Compare the two sets of CIs - from R regression and from the bootstrap. Discuss similarities/differences in your R notebook (note that any differences may be understood by performing regression diagnostics – remember the assumptions of regression and notably how are computed the regression coefficient SE for regression using formulas in the text)

Q3/ Trees and random forest. Divide the whole data in a train set and a test-set of 30% of the data. (30 points: 25 + 5)

1. Using the train set:
   1. Predict FM2 from the data using a pruned tree via cross validation. Plot the tree. Discuss your results.
   2. Predict FM2 using bagging. Discuss your results (including variable importance)
   3. Predict FM2 using random forest, using the formula for the number of selected predictors as given in the book. Discuss your results.
   4. Plot the variable importance for random forest. Compare with the pruned tree and discuss your results.
2. Predict FM2 using the lasso model of Q1 with the same test set. Compare the MSE on the test set of the four different methods. Discuss your results.

Q4/ PCA and K-means of the lasso-selected predictor variables. (30 points: 10 + 10 + 10)

1. Perform a PCA on all predictor variables selected by the lasso in Q2. Using a 90% variable accounted for cut-off, how many PCs do you find? Plot the results with a biplot. Discuss your results. What are the most important factors in each component?
2. Now, using K = 2, perform a K-means clustering on the predictor variables selected by the lasso
3. Plot the K-means results in the  PC1/PC2 axes. Discuss your results.