**Math 537, Summer 2020 – Final Exam**

In this exam we will be exploring a dataset called college.csv. In this dataset you will find the following variables:

ID:

X: Name of college (Neither a predictor nor a response)

Potential Responses:

Apps: Number of applications this last year (Potential Response)

Accept: Number of accepted applications (Potential Response)

Enroll: Number of enrollees (Potential Response)

Potential Predictors:

Private: Whether or not the University is private

Top10perc: % of applications that are top 10% students

Top25perc: % of applications that are top 25% students

F.Undergrad: Full time undergrads enrolled

P.Undergrad: Part time undergrads enrolled

Outstate: Out of stats tuition fee

Room.Board: Estimate for room and board

Books: Cost of books annually

Personal: Annual personal expenses

PhD: Number of PhDs granted last year

S.F. Ratio: Student to Faculty Ratio

Perc.alumni: Percentage of alumni who donate

Expend: Cost of education per year on average

Grad.Rate: Graduation Rate

You are to model the exclusivity of a University using the potential predictor variables.

Compute exclusivity as 100\*(Apps – Accept)/(Apps) + 100\*(Enroll/Accept), which is % applications denied + % of accepted applications attending.

You must use some form of cross validation to compare your models. Does not have to be comprehensive (Example, 10 fold cross validation where you rotate which subsets are testing 10 separate times), you can simply do one partitioning of the data into training and testing if you don’t want to loop overall all potential predictor subsets. Long Beach students, if you don’t know how cross validation works, let me know and Ill get you another mini lecture and some code.

Your two goals for the exam are simple.

1. Determine which of the following six models are best (as measured by your choice of cross validated objective fit.)
2. A simple linear model using least squares and all of the variables.
3. A ridge regression model with lambda = best.lam.ridge\*
4. A lasso regression model with lambda = best.lam.lasso\*
5. An elastic.net regression model with lambda1 = lambda2 = best.lam.elasticnet\*
6. PCR. Be sure to document and or discuss the number of components you selected and why.
7. A PLSR. Be sure to document and or discuss the number of components you selected and why.
8. **Very Briefly** (You don’t have time for lengthy discussion) discuss why you think the model that was most competitive won out over the others?

Please submit a single report.

best.lam\* Note, depending on how you treat your data I can’t really provide a one size fits all lambda. I suggest using the glmnet package to run your ridge, lasso and elastic net (alpha = 0, alpha = 1, alpha = .5) respectively. In order to use find your best.lam for each scenario use the following code:

cv.glmnet(data.matrix(x\_train),y\_train,alpha=,nfolds=)$lambda.min

Where x\_train are just the predictor values from your training data as a data.frame and y\_train are the response values from your training data. Again, alpha = 0,1,.5 creates ridge, lasso, elasticnet respectively. nfolds is the number of folds for the cross-validation within the training set, you can choose whatever you feel comfortable with.