ALY 6015 Assignment 3

For the following assignment, please submit your R code, the output, graphs and figures (if any), and your interpretation in a document that has your name, course information, instructor name, and the assignment number.

1. We are going to use the same dataset to try and find the best-fitting model to predict our outcome variable, Y, using different approaches to model selection. It will ultimately be up to you to determine which predictor (X) variables to keep in the model for each model selection approach.
   1. Import the “assignment3.csv” dataset into your R Studio. Split the data into a training dataset and a test dataset, with 75% of observations randomly going to the training data and 25% randomly going to the test data. Please use set.seed(12345) for this step so that we will have reproducible results.
   2. Let’s start with manual feature reduction. Use linear regression to predict EP (energy production) from the other variables in the dataset (using your training data). Start with a saturated model (all X variables in the dataset), and get the Adjusted R2 and F-statistic from the model output. Next, remove the predictors that are weakly correlated with the outcome variable (Y) based on significance level (p-values) and re-fit the model including only significant variables. Record the adjusted R2 and F-statistic for this model as well. Use the anova() function to see if the saturated and reduced models are significantly different from one another. Record results of this test. Using your saturated model and your reduced model (separately) predict values of Y using your test data. Calculate R2 for each model to assess prediction error. For question 1b please report which models you used (write them out as Y = β0 + β1X1 + β2X2 + ε, where Y is your outcome, β0 is the intercept, X1 is your first predictor variable, X2 is your second predictor variable (and so on), and ε is your error term), the parameter estimates, SEs, and p-values (or \*s) for each X variable, model fit statistics for each model from training data (just adjusted R2 is fine), prediction error for each model from test data (just R2 is fine, but you can calculate additional metrics if you want), and your interpretation of these results. You may use tables if you find it easier to present your results that way.
   3. Use the stepAIC() function to implement backward selection (starting with the full model). Does this approach give you the same reduced model you found above?
2. Using the same dataset from above, we’re now going to utilize Lasso to fit a model with all X variables to predict our outcome variable, energy production.
   1. Transform your training and testing datasets into a matrix of Xs and vector for Y, such that you get the following four objects: training matrix of Xs, training vector of Y, test matrix of Xs, test vector of Y.
   2. Using the training matrix, training vector, and the cv.glmnet () function, fit a Lasso model in order to find the best (minimum) Lambda. Using the training objects from above, also fit a Lasso model using glmnet() function and name it. Then use the predict function to assign this model and the minimum lambda to the test data, using something like: predict(ridge.mod, s = best.lambda, newx = x\_test).
   3. Compare predicted Ys and observed Ys using R2 (and/or your favorite prediction error metric). How well does the Lasso approach perform, and compare to the approaches from above?
3. Lastly, let’s try PCR with the same training and testing datasets from the energy dataset used above.
   1. Fit a PCR model to your training dataset using the pcr() function. Use the validationplot() function either set to RMSEP, MSEP, and/or R2 to determine how many principle components should be used in the final model. How many principal components did you choose?
   2. Use the predict() function on your test data to apply your fitted pcr model with the number of principle components you chose: predict(your\_pcr\_model, test, ncomp=number of principle components you chose). Next calculate the R2 to compare your predicted values from the pcr model to the observed values of Y. How did the PCR model do relative to the approaches above?

As always, please email me or post to the Discussion Board if you have any questions regarding the homework assignment.