MSDS 6372: Unit 2 HW 2

One of the major advantages of LASSO regression is that the estimate of the regression coefficients (betas) are allowed to be biased whereas the OLS estimates are forced to be unbiased. This highlights the commonly referred to “variance / bias trade-off”. Since MSE = Variance + Bias2, it is easy to see that for the OLS unbiased estimates that the MSE(betas) = Variance(betas) since the Bias is zero. However, with the biased LASSO estimates, one is often able to reduce the variance of the estimate of the betasat the cost of introducing a little bias. Often the reduction in the variance is greater than the increase in the squared bias and we see a reduction in the MSE of the betas.

In short, this means that our estimate of the regression equation with the smaller MSE has greater probability of being close to the equation with the real betas: the real trend. This means that if we cross validate our model on a test set (that maintains the same trend as the training set but with different noise), the model with the smaller MSE has greater probability of capturing more of the true trend. Statistically, this will be reflected in statistics such as the ASE (Test) (Average Squared Error for the model trained on the training set and used to fit the test set) and R squared of the test set (basically any goodness of fit statistic that is **with respect to the model and the test set**.)

We will use the ASE (Test) statistic to provide evidence of preferred models for this data set: LASSO or OLS.

Find the HW 2 Sas Code in the HW 2 folder. This code has the data set from the paper you read as well as some code to divide the data set into a training and test set.

The assignment for this week is simple. Simply run the code and make an argument / discussion as to which model / estimates (LASSO or OLS) will provide better predictive ability.

Make sure and copy and paste all relevant output to support your decision. Don’t overthink this. The answer can reference a single statistic.

Extra credit (5pts): Use the SAS Code from the Unit 2 Power Point and from last semester to provide confidence intervals for all estimates from both the LASSO and OLS models. Also report the margin of error for each interval and comment on which margins of error seem to be smaller. This should be fun! It uses a cutting age function and method that is related to machine learning topics (bootstrapping) called model averaging. If you have time, google search the SAS documentation and read about it … it is a very strait forward yet powerful idea.

When finished simply submit your finished HW on Canvas!