**Week 11 Problem Set**

1. There are two .csv spreadsheets attached to this assignment on canvas. The spreadsheets pertain to a single sample of X and Y that has been split into two sets“Training.csv” represents the training set; “Test.csv” represents the test set. Use that data to answer the following questions. If you forget how to calculate the regression parameters, look up the slides from “Inference for linear ergression”

1. What is the correlation between X and Y in the training set? What about the test set?
2. What are the standard errors of X and Y in the training set? What about the test set?
3. Compute the in-sample error
4. Compute the out-of-sample error

2. Answer true or false for the following questions:

1. Variance of your model is not important because it will eventually converge on the right answer
2. Error should be judged only on whether or not your estimated value is different from the actual value
3. If an estimator is MVUE, then it is the best estimator we can find
4. External validity refers to how well your model reduces out-of-sample error
5. Non-random samples produce excessive variance in your estimator
6. Non-random samples produce bias in your estimator
7. If an estimator’s bias is reduced, its variance is automatically reduced as well

3. Say we want to use cross-validation to calculate our prediction error. We have separate our data into four folds. Write out the different combinations for training and test sets, using those folds.