Mark Mersnik

ID: 010455185

EE257 Homework 4

**PROBLEM 1:**

3a) K-fold cross validation is implemented by taking the n observations and splitting them up into k equal sized groups that do not share any observations. One of the k groups is treated as the validation set while the method is fit to the remaining groups. The MSE is then calculated on the validation set. This process is repeated, using each k group as the validation set and fitting the method to the remaining groups. The average of each of the MSEs that are calculated are then taken.

3b) i) The advantage of k-fold over validation set is that the sets aren’t picked at random so there won’t be overlapping data. Also, the results of k-fold won’t vary as much as the results of the validation set approach. Lastly, the validation set approach is split in half, so it has less data points to fit the method to.

ii) The advantage of LOOCV is that each data point is represented which results in less bias. In addition to this, the resulting MSE will not vary as much as the validation set and even the k-fold set. The disadvantage of this approach, however, is that it requires the method to be fit n different times for each observation. This can be computationally intensive when compared to k-fold.

1) a) Best subset selection has the smallest training RSS since it fits a far greater number of models.

b) The most likely model to have the smallest test RSS is best subset selection although any of the models could be the smallest.

c) i) True. Since the model k+1 model is the same as the k model just with an additional predictor.

ii) True. Due to the same reason as the previous question.

iii) False. The predictors in backwards stepwise and forward stepwise are not correlated in any way.

iv) False. Due to the same reason as the previous question.

v) False. The k and k+1 predictors are not guaranteed to contain the same models.

4) a) Steadily increase since the other coefficients will have a smaller affect and the model will become less flexible.

d) Steadily increase since the coefficients are ignored and the value goes further and further away from RSS as lambda increases.

e) Remain constant since irreducible error is always present and constant no matter what changes are made to the model.

**PROBLEM 2:**

**PROBLEM 3:**

1. In the dataset n is 100 and p is 2. The model used to generate the data is the normal or Gaussian distribution with a mean of 0 and a standard deviation of 1. The generated data is then put through the following function in order to generate the Y results where x and n are 100 data points.
2. The plot shows a upside U shape that is to be expected with Gaussian distribution. The mean appears to be right around 0. When examining the standard deviation, the value seems to be a little less than the 1 that we inputted.
3. SEED = 12

The negative MSE of polynomial #1: -13.671850335938538

The negative MSE of polynomial #2: -1.0917916266189267

The negative MSE of polynomial #3: -1.1155769040248986

The negative MSE of polynomial #4: -1.2162271480556512

1. SEED = 22

The negative MSE of polynomial #1: -11.483713574205192

The negative MSE of polynomial #2: -1.1030599565291779

The negative MSE of polynomial #3: -1.1740764516212001

The negative MSE of polynomial #4: -1.2302982602736225

The results are not the same for c and d because the data that is generated by each of the seeds is different. The random seed number that we put in is correlated to its own set of data points.

1. ii, the 2nd degree polynomial had the smallest LOOCV error. I did think that this would be the case since anything more would be overfitting the training data and anything less is underfitting the data.
2. The coefficient estimates do agree with the conclusion that the second-degree polynomial is the best fit. This is because the coefficients that represent the first- and second-degree polynomial carry the most weight, are the furthest away from 0. This means that they are the most important when predicting the result accurately.

**PROBLEM 4:**