PORTFOLIO ASSIGNMENT 1

Andrew Guo

PROBLEM 1

The accuracy of the prediction $\hat{Y} = \hat{f}(X)$ for Y depends on two quantities. What are these quantities? State if they can be eliminated or reduced, and how would one accomplish that.

The accuracy depends on reducible and irreducible error. Reducible error can be reduced but not eliminated, but irreducible error cannot be reduced nor eliminated. Reducible error can be reduced by using more accurate modelings of f using f(hat), in other words, constructing a model (f(hat)) that more closely captures the nuances and true relationship of f. This could be done by using a higher order, more flexible model (A quadratic or cubic model as opposed to a linear model).

PROBLEM 2

(a) Briefly discuss one advantage of using a more flexible model. What can be gained by increasing flexibility?

Flexible methods, especially in situations with many predictors and a large sample size, are apt at reducing the amount of reducible error for our model are in general more effective at accounting for nonlinearity. Flexible models in general are more effective at predicting the true relationship of f and are said to have low bias.

(b) Briefly discuss one disadvantage of a more flexible model. What can be lost by increasing flexibility?

Flexible models tend to overfit the training data and have high variance, meaning that small changes in the data will lead to huge changes in the model. This means that if a flexible model fits a training data set extremely closely, there is a great possibility that this model will not fit the test data as well; the nuances and specific characteristics of the training data may not be present in the test data, and thus, the model no longer becomes suitable as a model to predict other/future data.

PROBLEM 3

We are given $n = 100$ observations with a single predictor, but are considering using a polynomial model of order p:

$$E(Y) = \beta_0 + \beta_1 X + \beta_1 X^2 + \cdots \beta_p X^p$$

(a) As p increases, how will the MSE change for the training data? For the test data?

As p increase, this means that our model will increase in its degrees, meaning that it will become more flexible. As such, the training MSE will decrease because more flexible models will be more adept at following the shape and nuances of the testing data. For test data, this is unknown because the true relatinoship of Y is not given. If the true relationship of Y is highly flexible, then we can expect the test MSE to be low. However, if the true relationship is more linear, then the test MSE may be higher due to the model overfitting the training data.

(b) As p increases, what will happen to the bias?

As p increase, the bias will decrease because higher order models do a better job than linear models at fitting the true relationship of F better.

(c) As p increases, what will happen to the variance?

Variance increases due to the model following the "noise" of the training model more closely, thus meaning that there will be much more variation due to a small change in the training data leading to a huge change in f.