* 1. **– Introduction to Supervised Learning for a Numeric Response**

The general model for predicting a numeric response () given a set of predictors can be represented as:

Here is a function that we need to estimate from our available data and  
 is the set of potential predictors (or features) available to help us predict the value of the numeric response . The predictors () may be a mixture of numeric and categorical/nominal/ordinal variables.

There are two main reasons we may wish to estimate : **(1)** ***prediction*** (the focus of this course) and **(2)** ***inference*** (the primary focus of STAT 360). In prediction we are really only interested in predicting the response accurately. Thus the estimated function is treated as a “black box” that hopefully accurately predicts the response (see diagram below). How each of the inputs (i.e. predictors) to the black box are used is not really of interest, although we may wish to measure in some sense which of inputs are most important. This is sometimes referred to as ***feature selection***. From a practical standpoint for future data collection and storage we would certainly like to eliminate the need to retain unimportant .

**Black Box**

**Output**

Predicted Value  
of Response

**Inputs**

(Predictors or Features)

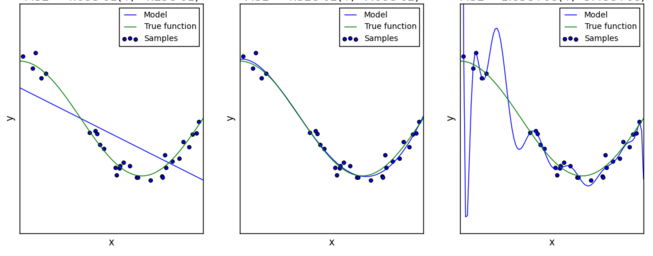
**Feedback about performance**

In this course we will examine numerous methods for estimating . We will start with the simplest methods and move towards increasing complex/flexible methods as the semester progresses. Many of these methods, or a variation of these methods, can used for classification problems as well where the response is non-numeric.

As an example of a method for estimating consider multiple linear regression (MLR) where the function is specified using a function that can be represented as a linear combination of functions of the predictors specified by the model ***parameters*** , i.e.

The functions are called ***terms***. A few examples of terms commonly used in MLR are given below:

MLR is one “simple” method for estimating and is an example of what is referred to as a ***parametric*** linear model. In Section 2 we will “review” MLR regression in much more detail. Again MLR was the main topic in STAT 360 – Regression Analysis. Other methods we will examine in this course are ***nonparametric*** in nature. Nonparametric methods do not make explicit assumptions about the functional form of . Instead they seek an estimate, **,** that gets as close the target response without being too rough or noisy or wiggly (see simple diagram below).



(Quartic Polynomial)

**1.2 – Prediction and the Mean Squared Error (MSE)**

The quality of given estimate is usually measured using the mean squared error (MSE) or Root Mean Squared Error (RMSE) which are given by:

However we need to be very careful here! The summation in MSE formula above was called the Residual Sum of Squares (RSS) in regression. As we learned in regression whenever we increase the complexity of a model the RSS, and hence the MSE, decreases. For prediction purposes overly complicated models will NOT perform well. This is because they fit the data used to develop model too well and thus will not generally perform well when making predictions for future observations that were not used in the model development process. ***Cross-validation***, which we will discuss a bit later, is used to develop models with “good” performance for the predictive problem at hand.

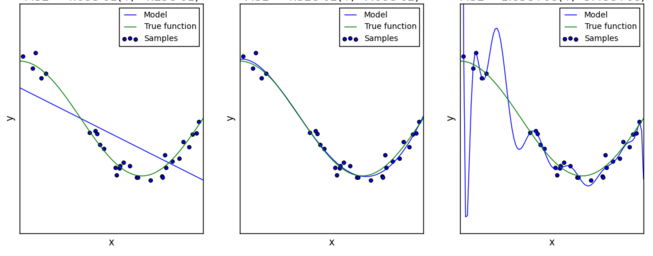
More or on the MSE

The general predictive/regression model for a numeric response is given by,

The MSE above is a sample based estimate of the population MSE which is given by,

Further expanding the right hand side of the expression above we obtain,

Thus the fitted model is balance between the squared bias and the variance of fit. In general more complex models will have smaller bias but will have larger variance. Choosing the “best” model requires striking a balance between the bias and variance. Consider again the diagram above.



(Quartic Polynomial)

Again to truly measure predictive performance we need to use observations that are NOT used in the model development process. Cross-validation methods will help us achieve this.

As stated above in order to obtain a reasonable (i.e. not overly optimistic) measure of predictive performance we need to use observations that were **not used in** **any way** to obtain the estimated model . Suppose we have a set of *m* complete observations that were not used in the model development process. Then we can calculate the Mean Squared Error for Prediction (MSEP) and the Root Mean Squared Error for Prediction (RMSEP) as follows,

We could also express our prediction accuracy as a mean absolute error (MAE) or as the mean absolute percentage error (MAPE).

and

**1.3 – Role of Inference**

In parametric modeling inference is usually of interest. For example when fitting a MLR model

determining which terms to include in the final model is usually based on t-statistics and p-values associated with the estimated coefficients. Also interpretation of the estimated parameters is of interest, particularly when the terms are the predictors themselves (i.e. ). In predictive modeling, accurately predicting the response is our really are only interest, so inferential analyses (i.e. p-values and confidence intervals) will not be seen much in this course. Although in the methods that are extensions of MLR (e.g. ridge, MARS, LASSO, and Elastic Net regression) we will still have the inferential tools used as part of MLR at our disposal.

In the next section we will review MLR with particular emphasis on the concepts of predictors & terms, transformations, model building, model selection, cross-validation, and discuss extensions of the MLR model.