**Week 1 Discussion Boards:**

Before we start fitting regression models, we begin the model building process by performing an Exploratory Data Analysis (EDA).  Given that we have a response variable Y and four predictor variables X1, X2, X3, and X4, how do we perform an EDA for a simple linear regression model?  Suppose that X1 and X2 are continuous and that X3 and X4 are categorical, does your approach to EDA for X1 and X2 differ from X3 and X4?

Exploratory data analysis will provide useful insights into our data and provide the following benefits:  check our assumptions; assist in identifying the appropriate model(s); identify relationships among various predictor variables and identify and assess the direction and size of relationship between a predictor variable and a response variable.  The first step is to prepare the data for analysis.  While we will follow a common methodology for EDA, regardless of the variable type however, we will use different techniques for different data types. For non-graphical EDA, continuous variables lend themselves to calculation of summary statistics (mean, median, mode,etc) as well as calculation of measures of correlation including Pearson's r.  For the categorical variables, x3 and x4, the concepts of central tendency do not apply however we are interested in the range of values and the frequency for each categorical value.  This information is best summarized and presented via a cross tab or contingency table.  We can also explore data graphically - for categorical data simple bar charts or pie charts would be appropriate while for continuous variables we could generate histograms, box plots, stem-leaf diagrams, etc.

First Question:     
    Exploratory Data Analysis (EDA) is   
based on the premise that the model always follow the data (Ratner   
2012). Flexibility, practicality, innovation, universality, and   
simplicity are all characteristics of EDA (Ratner 2012). EDA for a   
multiple regression model as stated above would follow these steps:

* Statement of the Problem - Perhaps one would like to explore the correlative relationship between X1...X4 and Y.
* Selection of Potentially Relevant Variables - Given this problem, one already has  
  defined the predictor variables, X1...X4, and the response variable, Y.
* Data Collection - Assembling the data consists of structuring the data such   
  that there are a specific number of observations where Y and the   
  predictor variables are measured and recorded.
  + Predictor variables can be both qualitative and quantitative in nature. It will   
    be assumed for this first response that all the predictor variables are   
    continuous. The second response will explain how to handle categorical   
    predictor variables in EDA.
* Model Specification or Analysis - The data needs to be analyzed such that a conclusion can be   
  reached regarding the linearity of the data. Specific models are used   
  once it is determined whether or not the data is linear. Given that this   
  problem is supposed to take four predictor variables and use one,   
  using the model with the highest R-score would be best. One could   
  use the r-selection technique in SAS to quickly discern which model is best.
* Method of Fitting - Before fitting the model, one must estimate the parameters  
  of the model based on the data. Techniques for determining the method   
  of fitting are least squares method, maximum likelihood, ridge   
  regression, and principal components.

Once the above steps are completed using the EDA approach, model fitting would be the   
next step for multiple regression (Chatterjee & Hadi 2012).

Second Question:    
  
    When dealing with continuous and categorical predictor variables   
simultaneously, the EDA changes slightly. The change that occurs   
compared to the previous EDA is an additional step for classifying the   
categorical predictor variables. Categorical variables are often called   
qualitative variables (Chatterjee & Hadi 2012 p129). These variables  
must be classified as indicator or dummy variables and are assigned a 0  
or 1 signifying specific categories. Additionally to this step, one   
must conduct an analysis of covariance. These two additions change the   
EDA slightly, but maintain the overall theme for EDA.

We can build statistical models for inference purposes or predictive purposes.  Imagine that we fit a simple linear regression model Y = b0 + b1X.  If we fit this model for the purpose of statistical inference, what are our primary motivations and what are the types of conclusions that we would like to draw?  What statistical tests are used to help us make these conclusions?  
  
Now, if we are building this model for predictive purposes, is statistical inference important?  What is important?  Are we interested in the results of statistical tests or are there other metrics that of more importance?  How do we evaluate a model for predictive purposes?  What metrics can we use?

Statistical inference pertains to extrapolation and interpretation of a specific statistical model. The primary motivations for statistical inference center on accuracy and precise measurement, and following strict methodology (Ratner 2012 p4). Statistical tests used to demonstrate statistical inference include confidence interval estimation, test of hypothesis, and goodness-of-fit tests (Chatterjee and Hadi 2012 p62). I think of statistical inference as a type “A” person who is not very flexible and conducts life with rigidity. The conclusions of statistical inference are myopically focused on the equation at hand and the exact parameters of the equation.

           Predictive modeling utilizes statistical inference, but is much more flexible in its application of formulas. Confidence interval estimation, test of hypothesis, and goodness-of-fit tests are all valuable in prediction, but the correlation coefficient is more valuable. The traditional inference statistics are a great guide and keep the equation in perspective. When prediction is involved, understanding the strength of relationship between the predictor (s) variable (s), and the response (s) variable is pivotal (Sirkin 2006). Metrics such as the correlation coefficient and coefficient of determination are metrics better suited for predictive purposes. These equations generate a score and percentage, and the higher the number (both positive and negative) the stronger the relationship (Chatterjee & Hadi 2012). When an equation has a strong relationship, it is considered a reliable indicator for predicting other values that fall within a specific range.

If a relationship is found between two variables 2 questions arise

 [1] first what is the probability that the relationship exists? - this is addressed by test of statistical significance. The test answers if the relationship is by chance.

 [2] second, if a relationship exists, how strong is the relationship?  this is addressed by Measures of Association.

 Statistically significant means there is a good chance the relationship exists. But statistical significance DOES NOT tell its predictor power of the variable. We can have a statistically significant finding, but the implication of the that finding may have no practical implication.

 For example, we may find that there is a statistically significant relationship between a citizen's age and satisfaction with city recreation services. It may be that older citizens are 5% less satisfied than younger citizens with city recreation services. But is 5% a large enough difference to be concerned about?

    Often times, when differences are small but statistically significant, it is due to a very large sample size; in a sample of a smaller size, the differences would not be enough to be statistically significant.

When we are fitting a model for statistical inference, our primary motivations are in measuring the relationship of the predictor variable(s) to the response variable.  That is, we are ultimately trying to gauge how much of a response in Y can be explained by the corresponding changes in X.  From this we would hope to be able to describe and quantify the relationship (is it linear or nonlinear?  a strong or weak relationship?  direction of the relationship?)  with measures of covariance and correlation coefficients.  Other tests such as hypothesis testing, , goodness of fit, and constructing a confidence interval can help us verify our findings and any conclusions we may be drawing from our analysis.

When we are using this model for predictive purposes our interests lay in "filling in the blanks", as it were, for data points we do not have.  However, statistical inferences remain important because those same tests provide us with measures for gauging our results.  After all, we cannot hope to produce reasonably accurate predictions if we don't understand the relationships between X and Y variables, how much of a change in Y is explained by a change in X, and how strongly they are related.  By knowing these values, we are in a much better position to make forecasts or fill in for unknown values with your model.

The r-square measure is a measure of interest for predictive accuracy because it measures how much of a change in the response variable is a result of the change in predictor variable.  This is very useful in telling us how directly the two variables are related.

When we build a model for the purpose of statistical inference our focus in on the model's explanatory value... for example, how well does variation in the predictor variable X explain variation in the response variable Y.  As discussed, when we take this approach to modeling we  set up a formal hypothesis test and generate goodness of fit statistics, F statistics, etc to test the explanatory value of our model.  When we build a model for predictive purposes, the focus is on how well the model performs.  Rather than focusing on statistical tests, calculating goodness of fit, etc we apply our model against a holdout sample of data and evaluate its performance in terms of accuracy, model gains, lift, etc