**Week 4 Discussion Board:**

How should we determine the importance of a predictor variable?  One type of importance is "statistical significance".  What does it mean for a predictor variable to be statistically significant (be very specific)?  Does statistical significance translate to predictive accuracy?

The importance of a predictor variable is centered around the goal of reducing the prediction error in the regression model (Ratner, 214). The p-value tests for significance using a hypothesis test, but it should not be assumed that the lower the p-value the greater the variables predictive importance. In fact, the p-value confirms the likelihood that the variable has some predictive value, but it does not answer how much importance. Once statistical significance has been verified through the associated p-value process, statistically analyzing the regression coefficient is the next step in delineating the most important predictor variable. Given that coefficients are often in different units, standardizing the units is important to verify and compare the magnitude of different predictor variables. Standardizing the regression coefficient entails multiplying the conversion factor by the regression coefficient. The conversion factor is often calculated by dividing the Standard Deviation of predictor variable Xi by the Standard Deviation of the response variable. One can also substitute the H-spread for the standard deviation. After all the coefficients have been standardized, the variables can numerically be ranked from largest to smallest and thus one has an ordinal ranking of most-least important predictor variables. This is assuming the data is not collinear.

After reading chapters 12 & 13 of SaMLDM by Ratner, statistical significance goes beyond the p-value for regression coefficients. A statistically significant predictor variable incorporates both a strong p-value and a numerically strong standardized regression coefficient. These two methods evaluate the reliability of a model, but do not confirm predictive accuracy. One would need to evaluate the extent to which the model measures what it is intended to measure using a different criterion than statistical significance (Ratner, 226). Utilizing the average correlation method is a technique that can be used to assess predictive accuracy between models and variables.

When we say that a predictor variable is statistically significant we mean that the predictor variable in question provides reliability and validity (in the words of Ratner in *Statistical and Machine-Learning Data Mining*), that is, we are asking "if it significantly reduces the prediction error of the regression model" (Ratner, 215).   We can assess the significance of a variable first through testing our null/alternative hypothesis with a p-value, and by this we are examining the prospects of our observations in the predictor variable coming about by happenstance.  

It is important to note, however, that determining a variable to be statistically significant is not the same thing as the variable being of predictive importance.  Although a variable may be significant, and significance may well be an indicator of importance,  we should instead rely on the standardized regression coefficient to evaluate its importance to a model.  This measure uses the correlation coefficient and multiplies it by a conversion factor to produce a new score (the standardized regression coefficient) by which predictor variables can be assessed for their impact on the response variable.  Normally distributed variables will use the standard deviation for this conversion factor, while those that are not can employ the H-spread (the difference between the 75th and 25th percentile of the distribution) as a conversion factor.

To determine the importance of the predictor variable, the average correlation should be used. Ranking the regression coefficient from highest to lowest will only work if the predictor variables are uncorrelated. The average correlation is the mean of the absolute values of the pairwise correlation coefficients. As Ratner states, the correlation coefficient is a measure of reliability, or closeness, of the relationship between a pair of variables. This should give understanding that the mean of all pairs should be an honest measure of closeness among the predictor variables in a given model.  To answer the questions, the mean is used with the correlation coefficient to yield the average correlation. The average correlation then can be used to assess the importance of a predictor variable. An indication of the importance of a predictor variable can be understood by where the value falls. Ratner suggest that values of the average correlation that range from 0 to 0.35 are soundly honest assessment of the contribution of the predictor variable in model. Values that range from 0.35 to 0.55 are somewhat honest and values that range from 0.55 and greater are questionably honest. The average correlation value should be under 0.40 to be acceptable.

When building OLS regression models, we typically use data transformations on both the response variable and the predictor variables.  Why do we use data transformations?  
In particular:  
(1) Can data transformations be used on the response variable so that the OLS modeling assumptions are met?  (Hint:  What if Y takes values: (a) in [0,1], (b) only integer values, or (c) very large positive values.)  
(2) Why would we want to transform a predictor variable?  
(3) Should the coding of a dummy variable be considered a data transformation?

RABE opens the chapter with: “Data do not always come in a form that is suitable for analysis”.  As a result, we use data transformations to accomplish certain objectives such as to ensure linearity, to achieve normality, or to stabilize the variance.  When the original variable violates one or more of the standard regression assumptions we can leverage data transformation to attempt to reverse this.     
  
The coding of a dummy variable should be considered a data transformation.  With dummy variables we are converting a nominal or categorical variable into a binary variable.  A variable may already have only two values (think X2 from the assignments) but creating a dummy variable may allow for more accurate use of the variable in the model.

Transformations are sometimes, or depending on your data, often necessary because the data itself is not in a format that facilitates the kind of analysis we ought to do when building a model.  A common example of this is when one of our potential predictor variables is a measure of some qualitative feature that makes it impossible to report that information with a continuous variable.  In other instances, our predictor variables may be nonlinear, but linear after transformation.  Finally, transformations can also help us reduce the error variance   
   
1)  It is my understanding that transformations can be be applied to both predictor and response variables.  Chatterjee and Hadi lay out several methods of transformation in RABE, including logarithmic, power, and weighted least squares, each of which is applicable in certain circumstances.    
   
2)  Transforming a predictor variable is recommended when the relationship is either nonlinear, the data is not normally distributed, or if the analysis of the residuals shows us that the variance in the predictor variable is not constant.  In these instances it is advisable to apply a transformation to the data and coax it into satisfying OLS modeling assumptions

3)  I feel like the creation of a dummy variable is a transformation.  Particularly because during the previous assignment I did observe a decrease in the standard error when comparing the unaltered model to one that included a dummy variable for X2.  This indicated to me that the dummy variable had in fact "improved" the data and made the model more accurate as a result.

R. Bhatti: Coding dummy variables is a transformation.  However, we are not making the data "more normal".  Instead, we are allowing the regression model to capture a nonlinear relationship between the predictor and the response variable.  Also, we do not interpret dummy variables in the same manner as continuous variables.  Dummy variables are interpreted as "intercept adjustments" (as you will see in Assignment #7) and plots of the residuals against the dummy variable are not valid.