**Week 2 Discussion Boards:**

What actions constitute an analysis of goodness-of-fit for a regression model?  Why do we assess the goodness-of-fit of a regression model?  What can go wrong in the interpretation of the results and the use of a regression model that would be deemed to "fit poorly"?

Completion of a goodness-of-fit analysis for a regression model involves both statistical tests and graphical inferences. In a goodness-of-fit analysis we are trying to confirm that the model, and the variables selected for it, are truly appropriate given high correlations to the response variable. The actions are somewhat redundant in they all may tell the same story, however execution of these steps in critical to ensure proper model selection. After executing standard steps such as a hypothesis test and analyzing resulting statistics such as t-scores a p-values, we also examine scatterplots to provide visual representations of the relationships between variables.

In the case of multiple regression, an additional step includes producing a scatterplot of the mean of the response variable (Y) versus the mean of the fitted values. As noted in RABE “the graph is very useful because it is used to assess the strength of the relationship between the response variable and the set of predictor variables.” (Chateerjee and Hadi, 2012)

A simple and effective method for detecting model deficiencies in regression analysis is the examination of residual plots. Residual plots will point to serious violations in one or more of the standard assumptions when they exist. Of more importance, the analysis of residuals may point to information in the data that might be missed or overlooked if the analysis is based only on summary statistics.” (Chatterjee and Hadi, 2012, p. 96)

We need to better develop the concept of cross-validation.  What is it?  How do we use it?  Why do we use it?  No one has adequately addressed cross-validation.  Let's use this thread to have that conversation.

The concept of cross-validation deals with how well a model will be able to estimate a prediction. Model-fit statistics, like correlation coefficient or coefficient of determination, indicate how well the model describes the data, not how well it predicts the data. The model will understand, or know, the relationship in the data between the response and predictor variables. This leads the model-fit statistics to overestimate, or over fit, the predictive skill of the model. To estimate the predictive skill of the model, a separate sample is used instead of the sample that was used to construct the model. Specifically, the testing sample should be completely distinct from the training sample.

In simple terms, cross-validation is when a model is formulated with one sample of the population and is then checked for validity by using another sample from the same population.

If the data has already been extracted from the warehouse and is well conditioned (ready for modeling), I would tend to use a larger proportion of the data as incremental benefits will be realized with little or no incremental costs.  While I am not concerned with the size of this dataset in terms of number of records, as 1 million records is relatively small in my experience, the presence of 100 different potential predictor variables does pose a challenge as models which use a larger number of predictors are generally harder to interpret and explain.  In addition to calculating correlation coefficients and ranking potential predictor variables and visually exploring fit of predictor variables via scatterplots, we may also want to use stepwise techniques such as forward regression to test the value added by each potential predictor variable.  We should use cross-validation, which involves dividing our data into training and testing partitions and applying our trained regression model to the holdout (testing) data, to control for overfitting.