**Week 6 Discussion Board:**

How do we perform an Exploratory Data Analysis for logistic regression?  Do we use the same tools that we used for ordinary logistic regression?

Whether we are performing Exploratory Data Analysis in support of a ordinary least squares regression problem or a logistic regression problem, there are similarities in the approach we will take.  Whether our outcome variable is continuous or categorical, we begin by exploring data, understanding the values different variables can take and generating hypotheses.  In ordinary least squares regression, we are interested in how well each potential predictor variable explains variation in our response variable.  R2 is a useful metric which we can use to evaluate the explanatory power of each candidate variable, given the assumed linear relationship between the candidate variable and the response variable holds.  We can validate the linear assumption by reviewing a scatter plot of candidate and response variable values.

For Logistic regression, given the categorical (often dichotomous) nature of the response variable, reviewing a scatter plot of values often does not provide a clear picture of the relationship between a given candidate variable and the response variable.  Also, our focus is different.  We are not looking for a candidate variable that best explains variation in a continuous response variable, we are looking for the candidate variables that can best classify a given observation into a specific response category.  Thus, rather than focusing on R2 and scatter plots, for logistic regression we utilize tools such as a frequency distribution for the candidate variable by the response variable.   As we have seen, if a given candidate variable almost always takes on a value of 1 when our response variable Y equals 1 (and rarely takes on a value of 0 when Y = 1), this is a candidate variable that we should retain.

βen: Like OLS, the purpose of performing an EDA for logistic regression is to identify and select the optimal variables to include in the model, as well as to ensure the combination of variables selected are ideal in terms of the models overall fit.     
  
To begin an EDA we first prepare our variables for appropriate use in logistic regression.   We create dummy variables for categorical variables and discretize continuous variables.  The latter discretization of the continuous variables is an iterative process in which we try to construct categories (cut-points) that create imbalance in the response variable.  The relationship of each predictive variable to the response variable can be examined by its mean distribution (against the response variable).  A higher variation in means indicates strong potential predictive power.     
  
Once the data has been transformed and we understand the relationship of each predictor variable to the response variable, we continue the EDA by selecting the optimal variables for the model.  To do this we can leverage the selection=score option in SAS to identify the best single variable logistic model (as indicated by the largest Chi-Square statistic). As we do with OLS regression, we can utilize automated variable selection (forward, backward, stepwise) to identify the optimal predictor variable(s) for the model.  Summary statistics such as the Wald Chi-Square for the intercept measures the significance of the model as a whole.     
The tools are similar in that for both linear and logistic regression uses statistics, or tools, to understand the relationship between the response and predictor variables. Though the measurements of the relationship will defer, in that logistics will use the Chi distribution and linear will use the coefficient of determination, they still measure the relationship between predictor and response variables  
The overall goal of an EDA is the same for OLS regression as it is for logistic regression: prepare and select the best variables for the model.  The steps we employ to do this are also similar, however the statistics (tools) used assess the strength of each predictor variable, and the model as a whole, are different.  Additionally, logistic regression models do not need to meet linearity requirements, as is the case with OLS regression.

How do we evaluate the goodness-of-fit for logistic regression?  Is the procedure the same as analyzing the goodness-of-fit for ordinary least squares regression?

Finally, ROC curves are useful visual representations of how well a logistic regression model fits the data and the area under the curve is a useful measure of model performance. Because in sample goodness of fit statistics ,which are derived from the same data used to train the model, often present an overly optimistic assessment of model performance, we should also validate the model against out of sample data.  This is especially true if our purpose in building the model is to use it to predict the occurrence (or lack of occurrence) of a specified event.    Performing out of sample tests will provide insight into how reliable our model might be in making predictions contrasted with the in sample tests focus on model accuracy.  One method of performing out of sample checks on our model is to partition our source data into training and testing subsets.  The model is trained using a majority of the data and the holdout sample (testing partition) is then used to assess model performance

We examine overall measures of fit including the Pearson Chi-Square, a measure of the difference between observed and fitted values (y and yhat) that “establish whether or not an observed frequency distribution differs from a theoretical distribution”. (Wiki, 2012)  Additionally, the Hosmer-Lemeshow test(s) can be used to assess “whether or not the observed event rates match expected event rates in subgroups of the model population”. (Wiki, 2012)     
  
We also make graphical interpretations of individual components of the summary statistics.  This includes classification tables, which are often more intuitive to understand.  These visual cross-tables produce sensitivity and specificity results that provide an additional measure of goodness-of-fit.  However, an even better technique to measuring sensitivity and specificity are ROC curves, which summarize the area under the curve.  A higher value indicates higher discrimination, and therefore higher predictive power in the model.  On page 162 of ALR we see that if the area under the ROC curve is greater than 0.9 then the model is considered to have “outstanding discrimination”.  This area under the ROC curve is also generated numerically in the standard SAS output (denoted as ‘c’).  Additional summary statistics include Sommers D, Gamma, and Tau-a, all of which range in absolute value from 0 to 1, with a higher value indicating a stronger relationship.