**Week 5 Discussion Board:**

Why is fitting a binary response variable with OLS regression statistically incorrect?  What problems would this OLS model exhibit, i.e. how would the fitted model violate the OLS assumptions?

OLS regression depends upon several assumptions which, when violated, will lead to inaccurate or misleading results.  One assumption of OLS regression is that residuals are normally distributed.  However, when modeling a binary response variable, residuals will not be normally distributed they will follow a binomial distribution.  Additionally, when modeling a binary response variable, residuals will not exhibit constant variance (homoscedasticity).  Applying OLS regression techniques to the binary outcome may also result in predictions outside of the bounds of 0 and 1 which will be difficult or impossible to interpret.

Fitting a binary response variable with OLS regression is statistically incorrect because of the violation of assumption of OLS. As others have mentioned, the normality assumption of the error term is violated with a dichotomous dependant variable. Though the consequence of the violating normality is not that severe. The normality assumption is not needed if the sample size is large. The central limit theorem gives the assurance that coefficient estimates will have a distribution that is approximately normal, even if the error is not normally distributed. If the sample size is small, the approximations would not perform as well. The second violation of assumptions of OLS, that others have mentioned as well, is the violation of homoscedasticity. Unlike the normality assumption, violation of the homoscedasticity assumption has two undesirable affects. First, the coefficients are no longer useful, which would mean that there are alternative method of estimation with smaller standard error. Secondly, the standard error estimates are no longer consistent estimates of the true standard error. Estimated standard errors would then by biases to unknown degrees, which eventually cause the test statistics to be difficult to calculate.

What is the response variable in a logistic regression model?  To what distribution does the response variable belong?  What quantity follows a logistic distribution?  
When a logistic regression model is fit, what quantity serves as the left hand side of the regression equation?  Is this value observed?  If not, how can we estimate it?  
Is the logistic regression model a 'linear model'?

In logistic regression, the response variable is categorical rather than continuous.  Often the response variable for logistic regression models is dichotomous meaning it can take on one of two possible values which often indicate the presence or absence of an effect or condition.  For example, we mights score financial transactions to predict whether a specific transaction is fraudulent (1) or is not fraudulent (0).  The response variable belongs to a binomial distribution.  The  conditional mean of Y given X is the quantity that follows a logistic distribution. When a logistic regression model is fit, the left hand side of the equation is the conditional mean of Y given X which is bounded between 0 and 1.  Is this value observed or if not how do we estimate?  To use the Credit Approval dataset from our assignment as an example, once we re-code Variable A16 to our response Variable Y, we do have an observed value of Y (either 0 or 1) for each observation.  We then work through Exploratory Data Analysis and also leverage automated Variable selection techniques to identify the best predictors of Y.  After we fit our model, we can compare estimates of Y generated via our model vs actual values of Y observed in our data and assess model fit.

**What is the response variable in a logistic regression model?**

In logistic regression, the response variable is dichotomous or binary.

**To what distribution does the response variable belong?**

The distribution of the response variable in logistic regression is a conditional distribution that follows a distribution with probability given by the conditional mean.

**What quantity follows a logistic distribution?**

The quantity that follows a logistic distribution is based on the conditional mean, which must be greater than or equal to zero and less than or equal to one.  
**When a logistic regression model is fit, what quantity serves as the left hand side of the regression equation?**

The quantity of a fitted logistic regression model will have the expression of logit or log-odds on the left hand side.

**Is this value observed?  If not, how can we estimate it?**The value of the logit serves as a likelihood function, which expresses the probability of the observed data as a function of unknown parameters. The maximum likelihood estimations of the parameters are chosen to those values that maximize the likelihood function.

**Is the logistic regression model a 'linear model'?**

Dr. Bhatti Logistic regression might have been based on linear regression, but it does not produce a linear model. Instead of achieving linearity, the curve of a dichotomous variable is S-shaped. The "score" or predicted probability from a logistic regression model follows a logistic distribution.