**Week 7 Discussion Board:**

Is the concept of multicollinearity still important in logistic regression problems.  Why or why not?

Multicollinearity is a problem in logistic regression that yields output that is worthless. Detecting collinearity amongst variates is not clear-cut. There are some hints to finding collinearity like very large estimated slopes, and huge standard errors. From researching the SAS website, one can also utilize a collinearity output  option that analyzes values associated with collinear variables.  These values and techniques are very similar to OLS regression.

Separation and  collinearity can be confused with each other, because both situations can cause large parameter estimates and standard errors. PROC LOGISTIC in SAS will display output if separation is an issue. If collinearity is found, rescaling the predictors can be one solution to eliminate collinearity.

As others have stated, multicollinearity is very much an important issue in logistic regression. Essentially we are still presented with a dilemma wherein interactions between predictor variables causes erratic behavior in the overall model when we begin to eliminate predictor variables from the model.  This can be especially problematic for linear or logistic regression because, when left in place, the collinear variables can still provide predictive power to the overall model, but as we attempt to reduce the number of variables in our model we then run the risk of removing one of the collinear predictor variables and consequently making the remaining collinear predictor variables behave very unpredictably.

Tell-tale signs of collinear variables can include wild swings in regression coefficients of variables after other predictor variables have been removed, or conflicting F-test and correlation coefficient statistics.  Though collinearity is not an issue if all variables are left in the model, collinearity can be corrected by ensuring you haven't erroneously used a dummy variable transformation, dropping any collinear variables, or as Daniel pointed out, by re-scaling the offending variables.

Yes, as others have noted, multicollinearity is still important in logistic regression problems and will have the same consequences seen in OLS regression.  First, it will be difficult to properly evaluate the effect that a given predictor has upon our response variable - if that predictor has a strong linear relationship with another predictor.  Second, the instability in model coefficients which results from multicollinearity can lead to highly improbable results.   We can use similar techniques to identify the presence including VIF scores - although we may need to use a lower VIF threshold then we would use for OLS regression.

What is meant by a "classification problem" in statistics?  Logistic regression is a statistical method that can be employed in the case of a binary classification problem.  In this context it is often referred to as "logistic discrimination", as the term "discrimination" is frequently used as a synonym for classification in the statistics literature.  There are two types of classification problems: supervised and unsupervised.  What are the differences between these two classification problems?  What is an example of a statistical technique for each type of classification problem?

classification problem in statistics is “the problem of identifying to which of a set of categories (sub-populations) a new observation belongs on the basis of a training set of data containing observations whose category membership is known.” (Statistical Classification, 2012)     
  
In the case of supervised learning you use labeled training data, whereas with unsupervised learning you are leveraging unlabeled data. (Unsupervised Learning, 2012)  An example method of supervised learning technique is regression.  An example of an unsupervised learning technique would be cluster analysis.     
  
I found a YouTube video *somewhat* helpful in distinguishing supervised from unsupervised learning.  I have no clue who this guy is, but he says there are also techniques that do both supervised and unsupervised learning.  He also notes that unsupervised learning is “the method of the future” because we are trying to make sense out of large amounts of unlabeled data. (YouTube, 2011)

In supervised classification, we have a specific target in mind.  For example, our source data may consist of transactions which have already been classified into two groups:  fraudulent and non-fraudulent (legitimate).  A logistic regression model can be trained to identify the important predictors of fraud and the resulting model could then be applied to score new data for a propensity of fraud. In unsupervised classification, we do not have a specific target field in our data.   Unsupervised techniques such as clustering can be leveraged to find subsets of observations which share common characteristics and thus are grouped together in distinct segments. For example, applying clustering algorithms to customer data may identify distinct customer segments and help a retail organization better tailor its offers to each segment.