

## Math 505 Lab 8

1. Import the data set `Lab8Data.txt`, which contains data for 3146 Biol 120 students, including the following variables
  - Grade: 1 = A, B, or C, and 0 = all other grades.
  - Rank: Percentile rank represented as a decimal between 0 and 1, with values close to one corresponding to higher ranked students.
  - Math and Verbal: Math and Verbal SAT scores
  - Prev: 1 = student has taken Biol 120 before and 0 = student has not.
  - Rdg: Status of student regarding the remedial course Reading 100. Possible levels are Never Taken, Concurrently Enrolled, Passed, and Failed.
  - Father's and Mother's education levels.
  - Gender
2. Build the best possible logistic regression model for predicting Grade based on the other variables.
  - (a) Divide the data set into two parts for the purpose of cross-validation.
  - (b) Fit a univariate model regressing Grade onto each of the other variables. For the quantitative variables, attempt to determine if higher order terms are needed using the `groupplot` function (see `LogisticRegressionFunctions.txt` for some helpful functions). As with linear regression, you can use a likelihood ratio test to formally test whether these terms are needed (`LRtest` function).

For the categorical variables, a univariate model can help to determine if some of the levels can be grouped together to create a variable with fewer levels. This is essential for the `father` and `mother` variables, which have eight levels. It's possible that the groupings used in Lab 4 will be effective here as well. It is likely that a stepwise regression will eliminate one of the parent's education variables, since they are highly correlated and have a large number of parameters.
  - (c) Use stepwise and best subsets methods to narrow down the list of predictor variables. Given the small number of predictor variables, you can also adopt a manual selection approach to select the variables or to modify the results of the stepwise/best subsets procedures.
  - (d) Fit a tentative final model. The quantitative variables should be checked again for functional form and categorical variables should be checked for groupings. You can also consider adding interaction terms.
  - (e) Assess the performance of the model by determining its classification accuracy using a cutoff probability of 0.5 and finding the area under the ROC curve. Each of these metrics can be calculated from the training sample using leave-one-out or delete-*d* cross-validation, and they can be calculating using the validation sample.