The hemophilia data set contains two measured variables on 75 women, belonging to two groups: n1=30 of them are non-carriers (normal group) and n2=45 are known hemophilia A carriers (obligatory carriers). MANOVA and LDA are linked together in terms of what is going on under the hood mathematically. MANOVA though is geared towards doing a simple ANOVA model with multiple response variables. LDA on the other hand, says hey, “using the response variables, can we predict what factor level the observation is coming from?”.

Discussion 1.

For each of the two variables do the follow preliminary summary work.

1. Calculate summary statistics per response variable by group.
2. Run single One Way ANOVA models for each variable. Perform residual diagnostics and make appropriate transformations if necessary to ensure each set of residuals meet the ANOVA assumption
3. Provide a scatterplot for Carrier vs Non for your final variables used in #2 and color code the points by group. Since we know each variable satisfies normality from #2, the only additional question is if the data are following multivariate normality with a constant covariance matrix for each group. Try to explain what visual properties you see in the plot would lead you for or against the assumption of constant covariance matrix. Recall covariance matrix is simply book keeping variance and correlation estimates for the two variables. A good search may help here.

Discussion 2

The back end of the SAS code performs LDA. I included a small data set to predict new observations. You will see a section of the output that provides the prediction for these new observations.

1. Take a look at the "Linear Discriminant Function for Group" table. For each group there is a column of numbers that are the estimates that used for predicting new observations. The calculation happens just like multiple regression B0+B1\*X1+B2\*X2. I want you guys to manually calculate the discriminate values for our 3 new observations by plugging them to both functions and reporting the values. Verify that the larger value corresponds to the predicted values in the SAS output table that provides the predictions of the 3 observations.