Practice questions for Logistic and Linear regression questions:

The dataset (CellDNA.xls) on the Blackboard contains various numeric measurements

(i.e. size, center, etc) from thousands of bacterium under microscope. All the

measurements are in different units. The last column with non-zero values indicate the

bacterium are interesting enough for further study. Otherwise (i.e. last column with zero

values), those bacterium are not interesting candidates for further study.

1. Use **\*\*logistic + lasso regression\*\*** with **\*\*10-fold cross-validation\*\*** to

identify useful predictors. Plot a lasso plot with readable tick labels on the X and

Y coordinates in your plot for easy visualization and verification. Missing clear

and readable tick labels in your plot will cost you significant points for this

assignment.

2. Which top **\*\*Three\*\*** predictors are you going to select to explain why a bacteria

is an “interesting” candidates for further study?

3. What is the lambda (l) value you choose in order to select the top 3 predictors

you identified in the last question?

4. What are the **q** values for the 3 selected predictors at the lambda (l) value you

identified in the last question?