Part A:

“Batch effects” can be examined by evaluating whether there is an association between the “batch” variable, which is “phase” in this case, and other variables of interest, such as the ROI biomarkers, of which there are 108. One way to evaluate batch effects graphically would be to create box plots (as each ROI biomarker is a continuous variable) across phases where one might be able to see a difference in means and standard deviations for each ROI biomarker based on phase. I’m not sure the question was asking for 108 figures of boxplots. I’m guessing the best way to evaluate batch effects across all ROI biomarkers was to have one figure that summarized the data. I was wondering if batch effects by phase could be evaluated using principal components analysis, as in the graph below:



Since the question is worded “evidence of strong batch effects in any image ROI biomarkers” I wasn’t sure if we in fact should be looking at all 108 individually. Would using PCA be an acceptable way to address this question? If so, other than noting the physical separation of red and blue dots on this graph, how do we give a more technical explanation of what the PCA plot is showing?

Part D:

I don’t recall discussion of pseudo R squared statistics in any of our classes. I looked up this topic on the internet and there appears to be many different kinds. Are any pseudo r squared statistics more widely used? Is there a good way choose from the many options? Not all information on the internet can be reliable – do you have a good resource for this topic?

Part F:

I am only partially familiar with the terminology “loss function.” I interpret part F to be asking the following:

1. Use subjects from phase 1 as a training dataset
2. Use subjects from phases 2 and 3 as a test dataset
3. Perform model selection such as to minimize the number of misclassified subjects in the test dataset.

I think the interpretation of this question may not be correct because the question says to use the variables selected in part E, where the model uses the full dataset.

Also, I don’t understand what is being asked in the following sentence: “Write down the objective function of your prediction model in detail.”

Parts E & G:

I am familiar with how to do modeling for polytomous outcomes in SAS, as we have done in several courses, with lower-dimensional data. Can you recommend resources for machine learning techniques using SAS with proportional odds and generalized logits models, or resources for how to model polytomous variables in R? I’ve tried using the polr() function for proportional odds and the multinom() function for generalized logits but I get very different answers from when I run the same model in SAS. The documentation in R is very poor and I don’t really know or trust what it’s doing for these kinds of models. I’ve actually been using R for the last few years for genetic data and trying to make sense of much of their documentation without being able to ask someone for help with it has been an absolute nightmare. Any help or suggestions on this would be appreciated.