Biost/Stat 571: Homework # 1

Tentatively Due 5pm, Wed January 18 via Canvas

Note: Homework should be submitted in as a PDF document with problems clearly labeled and in order. Use of Latex is preferred, but hand-written math is acceptable. In the case of the latter, handwriting must be in clearly legible print. Illegible (unreadable by the TAs and instructors) and unclear work will not receive credit. Clarity in derivations and exposition count as these are essential in any professional setting.

Note: **Problems 1-3 are intentionally open ended.** You do not need to provide your code, just your results.

Problem 1. Investigate the impact of correlation (among observations) on the performance of linear regression. Specifically, conduct a series of simulation experiments in which there are m individuals who each have n observations. For simplicity, you may make the outcomes normally distributed with common correlation ρ (i.e. compound symmetry under a random intercept model) and assume that there is only a single independent variable. Things to consider include different values of range of m, n, and ρ values. Then assess:

1. Type I error (you can use the usual wald p-values): note that this model should be under the null

2. Bias and coverage of 95% confidence intervals under some alternative model

Note that the mvtnorm package can be used to simulate multivariate normal data. Please summarize your findings using appropriate figures or tables with accompanying paragraph describing results. This should be similar to the simulation section of a research paper (albeit shorter and simpler).

Problem 2. Repeat the previous exercise except using a dichotomous outcome and using logistic regression. Note that instead of ρ you may use the probability that any pair of variables are equal to 1 or any other analogous quantity (e.g. ORs). Several different packages can be used to generate multivariate dichotomous outcomes.

Problem 3. The problem with dealing with multivariate data is that there are multiple outcomes. If there were only one outcome variable per subject, then we can just directly use the methods from 570. One easy way to make the data univariate is to simply take a weighted (either equal or unequal weights) average of the outcome measures. In many fields and application areas, this is what analysts (including some statisticians) do. For now, let's just assume that we use flat weights. Write a coherent discussion of the pros and cons of this mode of analysis from the perspective of practicality, validity, and power. Consider what situations it would be appropriate to use this type of analysis and under what situations such an analysis might even be nearly optimal. Justify your assertions and explain how you come to your conclusions.