STAT/BIOSTAT 571: Homework 6

To be handed in on Weds February 17th, in class. Please see 'Chapter 0' of the slides for a summary of how to answer questions appropriately, and the guidelines from 570. Where solutions require use of `R`, summarize your findings in a written answer, and append your **annotated** code, to show what you did. For each question, write up your solution on your own, using **full sentences**.

1. [**Missing data and multiple imputation**] Consider the data-generating mechanism where

$$\begin{aligned} b_i &\sim N(0, \sigma_b^2) \\ X_i &\sim Bern(0.5) \\ Y_{it} | b_i, X_i = x &\sim N(\beta_0 + b_i + \beta_1 t + \beta_2 x, \sigma_Y^2), \text{for } t = 1, 2, 3 \end{aligned}$$

where, as usual, all variables should be considered independent except where stated otherwise.

Suppose that data are missing according to a simple rule: if observation $Y_{it} < 0$ for some t, then all subsequent $Y_{i(t+1)}$ are missing. (For example, patients drop out, as we considered in class.)

   (a) Generate data from this mechanism, and implement GEE linear regression of Y on X and t, with the independence working correlation matrix. Using simulations (that you should describe) verify that complete-case analysis is not valid for inference on $\beta_0, \beta_1, \beta_2$, here.

   (b) Writing for a non-statistician, explain why the problem occurs. You may, optionally, use graphics to help you.

   (c) Implement multiple imputation, and again using simulation show how it improves inference for $\beta_0, \beta_1, \beta_2$. Your imputation model (and how it was fit) should be described carefully, and particularly its congeniality with the GEE analysis of interest

   (d) For keen people: (i.e. optional and earns no extra credit) What happens if you use the exchangeable working correlation matrix? Can you explain this?

   Throughout: use $\sigma_b = 1, \sigma_Y = 0.5, \beta_0 = 1, \beta_1 = -1, \beta_2 = 0.5$. The choice of sample size (n) and number of multiple imputations to use (K) and number of simulations is left to you; say what value(s) you used.

   Note: As mentioned in class, coding the imputation process may take much of the effort in your simulation work. It may help to use the `reshape()` command, so you can switch between datasets in long format for analysis and wide format for imputation.

2. [**Mixed models**] Consider the classic Neyman-Scott problem, where for clusters of size $n_i = 2$,

$$\mathrm{E}[\,Y_{ij}\,] = \mu_i, \mathrm{Var}[\,Y_{ij}\,] = \tau^2,$$

and all observations are independent. Interest lies in estimating τ^2. Suppose you fit this data with a mixed model, where following the notation in the slides

$$\begin{aligned} Y_{ij} | b_i &= \beta_0 + b_i + \epsilon_{ij} \\ \mathrm{E}[\,b_i\,] &= 0 \\ \mathrm{E}[\,\epsilon_{ij}\,] &= 0 \\ \mathrm{Var}[\,b_i\,] &= \sigma_b^2 \\ \mathrm{Var}[\,\epsilon_{ij}\,] &= \sigma_Y^2, \end{aligned}$$

and all ϵ_{ij} and b_i are independent. Using inverse-variance weighted linear regression with the standard plug-in estimates for the intra-cluster correlation familiar from GEE, for what values are $\hat{\beta}_0$, $\hat{\sigma}_Y^2$ and $\hat{\alpha}$ consistent? If your result requires any conditions on the (fixed, unknown) values of the b_i, say what they are.