STAT/BIOSTAT 571: Homework 3

To be handed in on Weds January 27th, 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. [**GEE**] This question considers data from the matched pairs design and data-generating mechanism seen in HW1 Q2

   (a) Give formulae for the point estimates obtained using GEE logistic regression of outcome Y_{ij} on binary covariate x_{ij} and an intercept with i) independence working correlation assumptions ii) unstructured working correlation assumptions. Hint: simulating some data where you know the truth should help you check your answers. Also note this question can be done without extensive algebra, and the answers should appear intuitively sensible

   (b) Suppose our data generating mechanism always gives equal numbers of matched pairs where $a_i = \Phi^{-1}((1-0.5)/100), \Phi^{-1}((2-0.5)/100), ...\Phi^{-1}((100-0.5)/100)$, i.e. the values returned by `qnorm(ppoints(100))`. For a range of values of β, describe the value to which the estimates in a) tend, asymptotically. (Numeric values are sufficient here) Hint: as in HW1, note that for each of the 100 types of matched pair there are only 4 possible outcomes, the probabilities of which you can evaluate

   (c) Imagine you show your answers from b) to a non-statistician, with whom you are working on analysis of a pair-matched study. They claim that b) shows that "GEE doesn't work for matched pairs; we must be assuming the wrong mean model". Write a short paragraph that explains why this is not true, and also correctly explains the results

2. [**GEE**] For the dental growth example seen in class, code up your own Fisher scoring algorithm (i.e. no using `glm()` or `gee()` or similar code) to implement GEE for AR-1 and exchangeable working correlation matrices, and compute sandwich-based confidence intervals for all regression parameters

   (a) Check that your code works by comparing the output to `gee()` (Describing your checks is sufficient)

   (b) To get convergence of all estimates ($\hat{\beta}_k$ and $\hat{\alpha}$) in their 1st/2nd/3rd... significant figure, how many iterations does your algorithm take?

3. [**GEE**] For your approach to HW2 Q2, use `gee()` to implement an analogous (i.e. as similar as possible) approach using GEE instead of nonparametric sandwich estimates

   (a) State which family, link and working correlation structure you used, and compare the results

   (b) Briefly explain any differences you see, or explain why the same results are produced