STAT/BIOSTAT 571: Homework 5 — Midterm Exam

To be handed in on Weds February 10th, in class. Please note this is an exam so **no conferring or discussion with other students is allowed**. You may ask the instructor and/or TAs for clarifications.

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. [**Vector outcomes, 6 points**]

   In this question we will directly motivate 'sandwich' variance estimates, for 'simple' linear regression, i.e. regression of outcome Y on one covariate X and an intercept, allowing for clustering. For simplicity, we will assume the covariates X are fixed.

   (a) Show that, in the usual notation from class, the unweighted OLS estimate of the X coefficient can be written as

   $$\hat{\beta} = \frac{1}{\sum_{i=1}^n n_i} \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x})}{V} Y_{ij},$$

   where $\bar{x} = \frac{1}{\sum_{i=1}^n n_i} \sum_{i=1}^n \sum_{j=1}^{n_i} x_{ij}$ is the 'grand mean' of the covariate values, and $V = \frac{1}{\sum_{i=1}^n n_i} \sum_{i=1}^n \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$ is the sample variability of the covariate values

   (b) Give an expression for the variance of $\hat{\beta}$, in terms of the notation above and the covariances $\text{Cov}[\,Y_{ij}, Y_{ij'}\,]$ of pairs of outcomes in the same cluster. What terms would you omit from this expression if the outcomes were not clustered?

   (c) By using a 'plug-in' value for these covariance terms in b) (that you should state) give an estimate for the variance of $\hat{\beta}$ that has a 'sandwich' structure, i.e. $\widehat{\text{Var}}[\,\hat{\beta}\,] = (\hat{\mathbf{A}})^{-1}\hat{\mathbf{B}}(\hat{\mathbf{A}})^{-1}$ for 'bread' and 'meat' components that you should describe

   Hint: check your calculations by (correctly!) coding up the answers for examples – such as those in the slides – where you already have valid estimates available.

2. [**GEE choice of mean model, 4 points**] In studies of different outcomes on the same individuals, it is often reasonable to believe that two different outcomes may be correlated despite being measured on very different scales. For example, individual students math and English scores may be correlated due to some common cause (e.g. parental involvement) but the scores themselves may be measured on totally different scales (e.g. tests out of 20 vs out of 100, or scores given as percentiles vs as raw scores). Covariates of interest may also differ between the outcomes (e.g. a series of indicators for different math teachers, versus indicators for English teachers).

   Suppose, for n students you have measured two such outcomes Y_{i1} and Y_{i2}, and three student-specific covariates U_i, X_i, and Z_i. You are interested in implementing linear regression of Y_{i1} on X_i adjusting for Z_i, and linear regression of Y_{i2} on U_i adjusting for Z_i; both linear regressions use an intercept.

   Carefully describe the mean model you would use to implement both of these regressions in a single GEE analysis, and state a choice of working correlation matrix that could take advantage of any correlation between the outcomes to enhance efficiency of the points estimates. Your description should include any manipulation of covariates and/or outcomes that is performed prior to fitting the mean model.

3. **[Efficiency, 3 points]** This question illustrates the extent to which issues of efficiency may matter in practice. Suppose you are using GEE in a situation where testing results (i.e. whether $p < \alpha$ or $p \geq \alpha$) are important, for a null hypothesis that some parameter $\beta \leq 0$ against the alternative that $\beta > 0$. Assuming large sample sizes (i.e. $n \approx \infty$) then the power of the test is given by

$$\text{Power} = 1 - \Phi\left(Z_{1-\alpha} - \frac{\beta}{StdErr(\hat{\beta})} \right),$$

where $Z_{1-\alpha}$ is the $(1-\alpha) \times 100$ percentile of the standard Normal distribution, and Φ is the cumulative distribution function of the standard Normal.

   (a) Suppose you know the Power of the test based on the Wald statistic when the true (efficient) covariance structure is used in the GEE analysis. However, you have used a different (inefficient) working covariance structure in your analysis. If you know the ARE for $\hat{\beta}$ from the two GEEs, level α, and the power of the efficient approach at level α, show how to calculate the power of the inefficient analysis.

   (b) Using your result in a), plot power of the less efficient versus power of the more efficient analysis, for $ARE = 0.1, 0.3, 0.5, 0.7, 0.9$ and $\alpha = 0.05$

   (c) Repeat b) but for $\alpha = 10^{-6}$ instead. Briefly compare the results with those in b)

4. **[GEE efficiency, 9 points]** This questions considers the relative efficiency of GEE, using exchangeable and independence working correlation assumptions. Suppose data are generated in clusters of size $n_i = 6$ as follows;

$$\begin{aligned} Y_{ij} &\sim Bern(p_{ij}) \\ \text{logit}(p_{ij}) &= \beta_0 + b_i + \beta_1 x_{ij} + \beta_2 z_i, \end{aligned}$$

where $x_{ij} = (j-3)/3$ for $j = 1, 2, ...6$, z_i is 0 for $n/2$ clusters and 1 for other $n/2$ clusters, and equal numbers of clusters have $b_i = \sigma\Phi^{-1}((1-0.5)/20), \sigma\Phi^{-1}((2-0.5)/20), ...\sigma\Phi^{-1}((20-0.5)/20)$, i.e. the values returned by `qnorm(ppoints(20))` scaled by σ.

   (a) For estimates obtained from GEE logistic regression of Y on X and Z, calculate the asymptotic relative efficiency (ARE) of using working independence assumptions versus working exchangeable assumptions. Do this fixing $\boldsymbol{\beta} = \{-2.5, 1, 1\}$ and varying σ over the values $0.5, 1, 1.5, 2, 2.5, 3$. As well as giving the numeric values of ARE, briefly say how you did the calculations. Note: numeric evaluation of the AREs is acceptable, and AREs need not be reported accurately to more than two decimal places

   (b) Repeat a), but now assume that in half the clusters (regardless of the values of z_i or Y_{ij}) only the first half of the outcomes (i.e. $j = 1, 2, 3$) are observed. Contrast your answers with those from a). Note: you are not expected to take this missingness (which is technically *ignorable*) into account in your GEE analyses

5. **[Writing about GEE: 8 points]** Read the paper by Liang and Zeger (Biometrika, 1986) that is available on the course website. Using full sentences and paragraphs throughout, write a brief summary of all the main ideas (1–2 double spaced typed pages, with 12 point font and 1 inch margins). While it may be helpful for your summary's structure to be similar to the original paper, do not plagiarize the paper or other sources; write using your own words and do not copy complete sentences or phrases from elsewhere.