Biost/Stat 571: Homework # 3

Due 5pm, Fri Feb 17 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 instructor) and unclear work will not receive credit. Clarity in derivations and exposition count as these are essential in any professional setting.

Problem 1. Consider the six city data on the class website. The data file is sixcity.dat and the variable names are given in sixcity.docx. This data set contains 537 children from Steubenville, Ohio, each of whom was examined annually from age 7 to age 10 for the presence of wheezing (which suggests a diagnosis of asthma). Mother's smoking status was reported. Consider fitting marginal logistic models with covariates age, smoking and interaction between age and smoking.

(1) Fit a regular logistic regression ignoring within-subject correlation and calculate the naive SEs.

(2) Analyze this data set using GEE1 assuming working independence and exchangeable. Compare the results with the naive logistic regression. Interpret each regression coefficient, except for the intercept.

(3) (BONUS) Analyze this data set using alternating logistic regression. Is there a strong within-subject correlation?

(4) Propose a GEE1 model that can separate the baseline age effect from the longitudinal time effect.

Problem 2. Conduct a literature search and find your favorite algorithm to simulate correlated binary outcomes which have exchangeable correlation $\rho = 0.25$ and marginal means follow

$$logit(\mu_{ij}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2ij},$$

where i indicates subject $i = 1, \cdots, m = 300$ and j indicates time point $j = 1, 2, 3$, $\beta_0 = -1.5, \beta_1 = 0.5, \beta_2 = 0.5$, $X_{1i} = 0/1$ is gender with half being males, and $X_{2ij} = j - 1$ indicates time.

(1) Simulate 200 data sets and run GEE1 with exchangeable correlation to estimate the regression coefficients and their sandwich SEs, and the correlation parameter ρ.

(2) Calculate the average regression coefficient estimates and the average correlation parameters across the 200 runs, and compare them with the true values, i.e., evaluate the empirical biases of the GEE parameter estimates.

(3) Calculate the average sandwich SEs, i.e, average "estimated SEs", and compare them with the "empirical SEs", which are the standard deviations of the estimate regression coefficients across

the 200 simulations. This allows you to evaluate whether the estimated SEs using the sandwich estimators work well in estimating the true variation of the GEE estimates $\hat{\beta}$'s.

(4) Does your program allow you to simulate correlated binary data with correlation 0.75?

Problem 3. GEEs vs. LMMs.

(1) Re-analyze the Framingham data from homework 2 using a GEE. Compare the results between using an LMM and a GEE.

(2) (Open Ended Problem) An investigator is interested in conducting a study in which one of two treatments will be administered to an even number of m individuals ($m/2$ in each treatment group). Subsequently, the investigator plans to measure a quantitative (continuous) biomarker on each individual n times longitudinally.

The investigator considers different strategies for analyzing the data: (1) by using a linear mixed model (2) using a GEE (3) using OLS with a fixed effect (indicator) for each participant.

What are the relative merits and/or limitations of these approaches? Conduct a simulation study to assess these approaches. If you were given more time, what additional simulation scenarios would you consider?

Problem 4. Multiple discrete and continuous endpoints are common in many biomedical studies. For example, in toxicology, fetal weight (continuous) and fetal death (binary) are observed. One is interested in studying the effects of dose levels on fetal weight and fetal death and jointly modeling these two outcomes, which may be correlated. Denote by Y_{1i} a continuous response (e.g., fetal weight) and by Y_{2i} a binary response (e.g., fetal death). Let $\mathbf{X}_i$ be a vector covariates, e.g., dosage. Note that Y_{1i} and Y_{2i} might be correlated.

Propose a **marginal/population-average model** to jointly model (Y_{1i}, Y_{2i}) by modeling their means as functions of $\mathbf{X}_i$. Write down the model. Propose estimating equations for estimating regression coefficients. Propose covariance estimators of the estimated regression coefficients. You need not implement/run this.

Problem 5. (BONUS) Consider longitudinal/clustered data $(Y_{ij}, \boldsymbol{X}_{ij})$, where i indicates subject i ($i = 1, \cdots, m$) and j indicates the jth observation of the ith subject ($j = 1, \cdots n_i$). For simplicity, we assume $n_i = n$. Denote by $\mu_{ij} = E(Y_{ij}|\boldsymbol{X}_i)$ the marginal mean of Y_{ij} conditional on the covariates $\boldsymbol{X}_i$ and $var(Y_{ij}|\mathbf{X}_i) = v(\mu_{ij})$ the marginal variance of Y_{ij}. Suppose the marginal mean μ_{ij} satisfies the generalized linear model

$$g(\mu_{ij}) = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}, \tag{1}$$

where $g(\cdot)$ is a monotonic differential link function, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients. Let $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{in})^T$, $\boldsymbol{\mu}_i = (\mu_{i1}, \cdots, \mu_{in})^T$ and $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \cdots, \boldsymbol{X}_{in})^T$. Consider the Generalized Estimating Equations (GEEs)

$$\boldsymbol{U}_\beta(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{m} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}(\alpha)\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} \tag{2}$$

where $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}^T = \boldsymbol{\Delta}_i^{-1} \boldsymbol{X}_i$ and $\boldsymbol{\Delta}_i = diag\{g'(\mu_{i1}), \cdots, g'(\mu_{in})\}$, $\boldsymbol{V}_i(\alpha) = \boldsymbol{V}_{Mi}^{1/2} \boldsymbol{R}_i(\alpha) \boldsymbol{V}_{Mi}^{1/2}$ is a working covariance matrix, $\boldsymbol{R}_i(\alpha)$ is a working correlation matrix assuming an exchangeable correlation, α is the corresponding working exchangeable correlation parameter, $\boldsymbol{V}_{Mi} = diag\{v(\mu_{ij})\}$ contains the marginal variances of the Y_{ij}.

Suppose the mean model is correctly specified by the GLM (1). The working correlation $\boldsymbol{R}_i(\alpha)$ might be misspecified. Show that under appropriate regularity conditions, the GEE estimator $\widehat{\boldsymbol{\beta}}$ is asymptotically normal

$$\sqrt{m}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \boldsymbol{A}^{-1} \boldsymbol{B} \boldsymbol{A}^{-1}$, and

$$\boldsymbol{A} = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \qquad (3)$$

$$\boldsymbol{B} = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i, \qquad (4)$$

$\boldsymbol{A}^{-1}$ is called the model-based covariance of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{B}^{-1}$ is called the empirical covariance of $\widehat{\boldsymbol{\beta}}$. When the working covariance $\boldsymbol{V}_i$ is correctly specified, $\boldsymbol{A} = \boldsymbol{B}$ and $\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}$.