Exercise #8
Due: March 10, 2010

Biostat/Stat 571
Winter 2010
P. Heagerty

Reading: ◦ Diggle, Heagerty, Liang & Zeger, Chapters 8 & 9

GEE Efficiency

1. One motivation for the adoption of a "working correlation" structure with GEE is to obtain more efficient estimates, $\widehat{\boldsymbol{\beta}}$.

Consider the logistic regression model:

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

where X_1 is a covariate that varies within a cluster and X_2 is a covariate that varies between clusters.

Assume that $X_{1,ij} = (j-3)/3$ for all clusters (ie. a "time" variable). Assume that $X_{2,i} = 0$ for half the clusters and $X_{2,i} = 1$ for half the clusters (ie. a "treatment" variable).

(a) Assume that the data are balanced with $n_i = 6$ observations per cluster. Calculate the asymptotic relative efficiency of the exchangeable GEE estimator relative to to the independence GEE estimator when the data truly have an exchangeable correlation structure. Numerically calculate the ARE for $\widehat{\boldsymbol{\beta}}$ (each element) for a range of correlations (but keep them legal!) if $\boldsymbol{\beta} = (-2.5, 1.0, 1.0)$.

(b) Now assume that the data are comprised of subjects that complete only visits 1, 2, and 3 (50%) and subjects that complete all six visits (50%). Calculate the ARE $\widehat{\boldsymbol{\beta}}$ (each element) as in part (a) for a range of correlations if $\boldsymbol{\beta} = (-2.5, 1.0, 1.0)$.

(c) Assume that the data are balanced with $n_i = 6$ observations per cluster. Calculate the asymptotic relative efficiency of the AR-1 GEE estimator relative to to the independence GEE estimator when the data truly have an AR-1 correlation structure. Numerically calculate the ARE for $\widehat{\boldsymbol{\beta}}$ (each element) for a range of positive correlations if $\boldsymbol{\beta} = (-2.5, 1.0, 1.0)$.

(d) Now assume that the data are comprised of subjects that complete only visits 1, 2, and 3 (50%) and subjects that complete all six visits (50%). Calculate the ARE $\widehat{\boldsymbol{\beta}}$ (each element) as in part (c) for a range of positive correlations if $\boldsymbol{\beta} = (-2.5, 1.0, 1.0)$.

Data Analysis

2. The Madras Longitudinal Study collected monthly symptom data on schizophrenia patients after their initial hospitalization. One scientific question is whether subjects with a younger age-at-onset tend to recover more/less quickly, and/or whether female subjects recover more/less quickly. Recovery is measured by a reduction in the presentation of symptoms. For this analysis we will consider the

outcome *thought disorders* which is a binary indicator of whether the patient was observed to present this "positive symptom" during the month.

(a) Summarize the prevalence of symptoms over time for groups of patients defined by age-at-onset and gender.

(b) Summarize the serial dependence in the response series using a matrix of pairwise correlations, and a matrix of pairwise odds ratios. For this summary you need not adjust for age-at-onset or gender. Interpret the patterns in these dependence summaries.

(c) Use appropriate regression methods to determine if these data suggest different rates of recovery (ie. change in prevalence over time) for younger age-at-onset or female patients.

Conditional and Marginal Means

3. Consider the GLMM for count data given by:

$$\begin{aligned} Y_{ij} \mid \boldsymbol{X}_i, b_{i,0} &\sim \text{Poisson} \\ \log E[Y_{ij} \mid \boldsymbol{X}_i, b_{i,0}] &= \beta_0 + \beta_1 X_{ij} + b_{i,0} \\ b_{i,0} &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

(a) Using information about the normal moment generating function derive the expression for $\mu_{ij} = E[Y_{ij} \mid \boldsymbol{X}_i]$ as a function of $\boldsymbol{\beta}$ and σ^2.

(b) Give a precise interpretation of the parameter β_1.

(c) If the marginal model, $\log E[Y_{ij} \mid \boldsymbol{X}_i] = \beta_0^* + \beta_1^* X_{ij}$ were fit to data generated via this hierarchical model would the estimators of β_0^* and/or β_1^* be consistent for β_0 and/or β_1, their analogue in the conditional model? Justify.

(d) Now consider the model that contains random intercepts and random slopes:

$$\begin{aligned} Y_{ij} \mid \boldsymbol{X}_i, b_{i,0}, b_{i,1} &\sim \text{Poisson} \\ \log E[Y_{ij} \mid \boldsymbol{X}_i, b_{0,i}, b_{i,1}] &= \beta_0 + \beta_1 X_{ij} + b_{i,0} + b_{i,1} X_{ij} \\ b_{i,0} &\sim \mathcal{N}(0, \sigma_0^2) \\ b_{i,1} &\sim \mathcal{N}(0, \sigma_1^2) \end{aligned}$$

where $b_{i,0}$ and $b_{i,1}$ are assumed independent.

Give a precise interpretation of the conditional parameter β_1 and give an interpretation of σ_0 and σ_1.

(e) Derive the marginal mean $\mu_{ij} = E[Y_{ij} \mid \boldsymbol{X}_i]$.

4. (**optional**) One numerical integration method used to obtain likelihood estimates for GLMMs is known as *Gauss-Hermite* integration. Gauss-Hermite integration uses "quadrature" to numerically evaluate an integral. The evaluation is based on choosing a number of points to be used (K) and then approximating the integral with a weighted sum:

$$\int_x g(x) \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)dx \quad \approx \quad \sum_{i=1}^{K} g(x_i) \cdot w_i$$

On the class web page for Exercise #8 there is **R/S+** code to obtain the quadrature points (x_i) and the weights (w_i) for either $K = 5$, $K = 10$, or $K = 20$ nodes. Use this routine to calculate the marginal logistic regression parameter that is induced by

$$\begin{aligned} \text{logit}E[Y_{ij} \mid \boldsymbol{X}_i, b_{i,0}] &= \beta_0 + \beta_1 X_{ij} + b_{i,0} \\ b_{i,0} &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

for a model with a single binary covariate $X_{ij} = 0/1$ and $\boldsymbol{\beta} = c(-2.0, 1.0)$. Use $K = 20$, and $\sigma = 0.5, 1.0, 1.5$.