Mickey Warner

Stat 637 – Mini Project # 1

Introduction

In this report, we review Zeger *et al.*'s 1988 paper titled *Models for Longitudinal Data: A Generalized Estimating Equation Approach*. The authors consider two classes of models: subject-specific (SS) and population average (PA). These models are employed when several measurements are collected on individuals across time. These measurements are typically correlated which adds an additional challenge in the analysis.

Subject-specific models are used when the response for an individual is the focus. These models taken on the form of a generalized linear mixed model. For subject i at time t, let y_{it} be the response, $\mathbf{x}_{it}$ a $p \times 1$ vector of fixed covariates associated with $p \times 1$ fixed effects $\boldsymbol{\beta}$, and $\mathbf{z}_{it}$ a $q \times 1$ vector of covariates associated with $q \times 1$ random effects $\mathbf{b}_i$. Let $u_{it} = E(y_{it}|\mathbf{b}_i)$. We assume the responses satisfy

$$h(u_{it}) = \mathbf{x}_{it}^\top \boldsymbol{\beta} + \mathbf{z}_{it}^\top \mathbf{b}_i \quad \text{and} \quad \text{var}(y_{it}|\mathbf{b}_i) = g(u_{it}) \cdot \phi$$

where $\mathbf{b}_i$ is an independent observation from some distribution, F, and $i = 1, \ldots, K$ and $t = 1, \ldots, n_i$. Typically, $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$. The functions h and g are referred to as the "link" and "variance" functions, respectively. Choices for h include the log, logit, or probit links. The variance function g may be defined by the choice of likehood (e.g. with a Poisson likelihood, we have $\text{var}(y_{it}|\mathbf{b}_{it}) = u_{it}$). The scale parameter ϕ can be used to define quasi-likelihoods.

When inference on a population is of more interest than inference on a subject, an alternative model is the population average model. Let $\mu_{it} = E(y_{it})$ be the marginal expectation. The responses then satisfy

$$h^*(\mu_{it}) = \mathbf{x}_{it}^\top \boldsymbol{\beta}^* \quad \text{and} \quad \text{var}(y_{it}) = g^*(\mu_{it}) \cdot \phi,$$

for link and variance functions h^* and g^*. The population parameters $\boldsymbol{\beta}^*$ describe the relationship between the covariates and the average response across all subjects.

Parameter estimation

The authors' method of estimating the parameters $(\boldsymbol{\beta}, \mathbf{D}, \phi)$ is comparable to the iteratively reweighted least squares approach as discussed in class. However, the covariance between responses and the random effects must be taken into account. This is done through generalized estimating equations.

Estimates may be calculated for the random effects $\mathbf{b}_i$, but these merely provide offsets for each subject and so specific values are not of particular interest. When random effects are present in the model, interpretation of the fixed effects $\boldsymbol{\beta}$ has a subtle difference than that of $\boldsymbol{\beta}^*$. We discuss interpretation in the next section.

We omit the actual procedure as described in the article since it is rather involved. Software for parameter estimation is provided in the R package `geepack` with the `geeglm()` function.

Example: respiratory illness

We use the `respiratory` data set from the R package `geepack`. The data are from a clinical trial of patients with respiratory illness. A placebo or active treatment was randomly given to 111 patients from two clinical centers. Patients were examined at baseline then again examined at four return visits. The outcome is respiratory status, where good = 1 and poor = 0. Sex and age are also given covariates.

We demonstrate both the PA and SS models on the respiratory illness data set. Zeger *et al.* (1988) show that the parameters estimates and t-values are fairly robust to misspecification in the correlation $\mathbf{R}$, so we will choose one structure in our analysis. In Table 1 we show the sample correlation among the binary outcomes. The correlations appear to be very similar, hence we will use the exchangeable correlation structure: $R_{jk} = \alpha, j \neq k$ and $R_{jj} = 1$.

Visit	1	2	3	4
1	1.000	0.509	0.443	0.514
2	0.509	1.000	0.582	0.530
3	0.443	0.582	1.000	0.587
4	0.514	0.530	0.587	1.000

Table 1: Correlation among of the binary response for the respiratory data