ST 762, HOMEWORK 5 EXTRA PROBLEMS, FALL 2009

These problems are from previous years and are for you to work on or not as you choose; they are not to be turned in. Solutions will be posted when the homework problems to be turned in are due.

1. *Subject-specific versus population-averaged modeling.* An investigator has collected data from m randomly-chosen subjects in the form of $(n_i \times 1)$ vectors $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{in_i}]^T$ taken at time points x_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$. The subjects each started an anti-fungal treatment at time 0, and the response Y_{ij} is binary, indicating whether the subject's toenail fungus has flared up or not at time x_{ij} after the start of treatment.

   The investigator approaches 2 statisticians for help in fitting a model to the data. She tells each statistician she would like to know whether the probability of flare-up changes over the period of administration. She notes that she would like to model the probability of flare-up at a particular time using a logistic regression model, as this is standard in the literature in her field.

   Both statisticians recognize that the elements of the vectors $\boldsymbol{Y}_i$ may be correlated.

   Statistician A decides to assume that $E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|x_{ij})$ and $\text{var}(Y_{ij}|\boldsymbol{x}_i) = \text{var}(Y_{ij}|x_{ij})$. She postulates the model

$$E(Y_{ij}|x_{ij}) = H(\beta_0 + \beta_1 x_{ij}), \quad \text{var}(Y_{ij}|x_{ij}) = H(\beta_0 + \beta_1 x_{ij})\{1 - H(\beta_0 + \beta_1 x_{ij})\}, \qquad (1)$$

   and

$$H(u) = \frac{\exp(u)}{1 + \exp(u)}.$$

   Thus, A has specified logistic regression model for the marginal mean. To account for possible correlation among the components of $\boldsymbol{Y}_i$, she assumes that the conditional (on $\boldsymbol{x}_i$) correlation between any two elements of $\boldsymbol{Y}_i$ is equal to some constant α.

   Statistician B, on the other hand, decides to adopt the following model. Writing $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in_i})^T$, he assumes that $E(Y_{ij}|\boldsymbol{x}_i, b_i) = E(Y_{ij}|x_{ij}, b_i)$ and $\text{var}(Y_{ij}|\boldsymbol{x}_i, b_i) = \text{var}(Y_{ij}|x_{ij}, b_i)$, where he takes

$$E(Y_{ij}|x_{ij}, b_i) = H(\beta_0 + \beta_1 x_{ij} + b_i), \quad \text{var}(Y_{ij}|x_{ij}, b_i) = H(\beta_0 + \beta_1 x_{ij} + b_i)\{1 - H(\beta_0 + \beta_1 x_{ij} + b_i)\}. \qquad (2)$$

   Here, b_i is assumed to be a $\mathcal{N}(0, D)$ random effect independent of $\boldsymbol{x}_i$. Thus, this statistician specified an ordinary logistic regression model for subject i. He figures that incorporating the random effect b_i will "automatically" induce a correlation structure, so will account for correlation among the elements of $\boldsymbol{Y}_i$.

   (a) Give interpretations of the parameters β_0 and β_1 in Statistician A's model.

   (b) Give interpretations of the parameters β_0 and β_1 in Statistician B's model.

   (c) For the modeling strategy of Statistician A, the marginal conditional expectation $E(Y_{ij}|x_{ij})$ is given exactly. However, to facilitate comparison with that implied by B's model in (d) below, it is useful to write an approximation to $E(Y_{ij}|x_{ij})$ using the fact that the standard logistic distribution has CDF $F(u) = 1/(1 + e^{-u})$, $-\infty < u < \infty$, that may be approximated by the CDF of a normal random variable with mean 0 and standard deviation $c = 15\pi/(16\sqrt{3}) \approx 1.7$. Using this result, give an approximation to the model for $E(Y_{ij}|x_{ij})$ given by A in terms of the standard normal CDF.

(d) For the modeling strategy of Statistician B, find an approximation to the marginal conditional expectation $E(Y_{ij}|x_{ij})$ using the standard logistic distribution result in (c) and any other manipulations you find necessary. *Do not quote results from books; you must derive the result yourself.*

(e) Compare the models for marginal conditional expectation $E(Y_{ij}|x_{ij})$ under the two strategies. What, if any, common characteristics do they share, and how are they different? In general, is it possible for β_0 and β_1 in each model to have the same interpretation? Under what conditions would the interpretation of these parameters be the same?

(f) Suppose Statistician B's model is correct. Statistician A fits her model by the methods in Chapter 14 of the course. Will the estimators of the parameters β_0 and β_1 obtained by Statistician B consistently estimate the parameter she thinks she is estimating? *Explain your answer.*

2. Consider the population-averaged marginal model

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{f}_i(\boldsymbol{x}_i\boldsymbol{\beta}), \quad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{x}_i), \tag{3}$$

where $\boldsymbol{Y}_i$ is $(n_i \times 1)$ and $v_{ijk}(\boldsymbol{\beta}, \boldsymbol{\xi}) = v_{ijk}$ is the (j,k) element of $\boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{x}_i)$. Define the $\{n_i(n_i+1)\}/2\times 1\}$ vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ as on page 380 of the class notes, and define the "gradient" matrix $\boldsymbol{E}(\boldsymbol{\beta}, \boldsymbol{\xi})$ as on page 381. Assume that the $\boldsymbol{Y}_i$, $i = 1, \ldots, m$, are independent across i. Consider for fixed $\boldsymbol{\beta}$ the estimating equation for $\boldsymbol{\xi}$ given by

$$\sum_{i=1}^{m} \boldsymbol{E}_i^T(\boldsymbol{\beta}, \boldsymbol{\xi}) \boldsymbol{Z}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\xi}) \{\boldsymbol{u}_i - \boldsymbol{v}_i(\boldsymbol{\beta}, \boldsymbol{\xi})\} = \boldsymbol{0}, \tag{4}$$

where $\boldsymbol{Z}_i(\boldsymbol{\beta}, \boldsymbol{\xi})$ is the "working covariance matrix" for $\boldsymbol{u}_i|\boldsymbol{x}_i$.

All of the following calculations should be carried out under the supposition that model (3) is correctly specified.

(a) Suppose that the "working covariance matrix" is specified using the "independence working assumption" on page 384 with the additional assumption that the Y_{ij}, $j = 1, \ldots, n_i$, are normally distributed. That is, pretend that the elements Y_{ij}, $j = 1, \ldots, n_i$, are independent and normal for all i. Find the form of $\boldsymbol{Z}_i(\boldsymbol{\beta}, \boldsymbol{\xi})$ under these conditions.

(b) Show that, under the conditions of (a), the kth row of the estimating equation (4) is the same as the pseudo-likelihood equation (14.10) on page 377.

(c) Now suppose that the data are *bivariate*, so that $n_i = 2$ for all i. Thus, each $\boldsymbol{Y}_i$ is a bivariate response $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})^T$ (not necessarily independent). Suppose further that the "working covariance matrix" is specified using the "Gaussian working assumption." Under these conditions, find $\boldsymbol{Z}_i(\boldsymbol{\beta}, \boldsymbol{\xi})$.

(d) Show explicitly that, under the conditions in (c), the kth row of (4) is identical to the pseudo-likelihood equation (14.10) on page 377.

Note: in (b) and (d), you must carry out the analytical steps to verify the equality explicitly; do not quote general results in the notes.