ST 790, Homework 5
Spring 2018

1. *Unbiasedness of the weighted generalized estimating equation.* In Section 8.7 of the notes, we saw that, with missing data, the usual linear estimating equation (8.70) for estimation of β in a population-averaged model of the form (8.3) with exogenous among-individual covariates is not unbiased unless the missingness mechanism is missing completely at random (MCAR) and thus need not lead to a consistent estimator for the true value of β under the assumption of a missing at random (MAR) mechanism.

   One approach to modifying the equation (8.70) to render it unbiased under a MAR mechanism is to modify it by weighting the contribution for each individual i to the equation by the reciprocal of the probability of the individual dropping out at the time s/he did given her/his past history, as in (8.80), that is

$$\sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} w_{ij} \boldsymbol{X}_i^{(j)T}(\beta) \{ \boldsymbol{V}_i^{(j)}(\beta, \xi, \boldsymbol{x}_i) \}^{-1} \begin{pmatrix} Y_{i1} - f_1(\boldsymbol{a}_i, \beta) \\ \vdots \\ Y_{ij} - f_j(\boldsymbol{a}_i, \beta) \end{pmatrix} \right\} = \boldsymbol{0}. \tag{1}$$

   where

$$w_{ij} = \frac{I(D_i = j+1)}{\overline{\pi}_j(\mathcal{H}_{i,j-1}) \lambda_{j+1}(\mathcal{H}_{ij})},$$

   and all quantities are as defined in the class notes.

   Assuming that the cause-specific hazard functions $\lambda_j(\mathcal{H}_{i,j-1})$ are correctly specified and the missing mechanism is MAR, show that (1) is an unbiased estimating equation.

2. *Laplace approximation.* Consider the nonlinear mixed effects model with first stage individual model with second stage model substituted, of the form

$$E(\boldsymbol{Y}_i | \boldsymbol{x}_i, \boldsymbol{b}_i) = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i), \quad \text{var}(\boldsymbol{Y}_i | \boldsymbol{x}_i, \boldsymbol{b}_i) = \boldsymbol{R}_i(\gamma, \boldsymbol{x}_i), \tag{2}$$

   so that the within-individual covariance matrix does not depend on β_i and thus does not depend on β or $\boldsymbol{b}_i$. Assume further that $\boldsymbol{b}_i$ are independent of $\boldsymbol{x}_i$ and satisfy $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$ and that that the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ and $\boldsymbol{b}_i$ is normal with moments given in (2).

   Consider the derivation of the approximate expression for $p(\boldsymbol{Y}_i | \boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D})$ in (9.85) using Laplace's approximation. Using (9.86) and your matrix algebra prowess, show that (9.85) can be expressed equivalently as (9.87). Thus, you will have shown that the contribution of individual i to the likelihood (9.46) can be approximated by a normal density with mean and covariance matrix given in the first bullet on page 322.

   Hints: It will be convenient to define

$$\boldsymbol{u}_i = \boldsymbol{Y}_i + \boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i) \widehat{\boldsymbol{b}}_i$$

   and to use shorthand notation such as $\boldsymbol{Z}_i = \boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i)$, $\boldsymbol{f}_i = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i)$, etc.

3. *Clinical trial in rheumatoid arthritis, continued.* Recall the data from the clinical trial in patients with rheumatoid arthritis in Problem 4 of Homework 4. The data in the file `arthritis.dat` are on $m = 290$ subjects who were randomized to placebo or auranofin therapy. Baseline age

was recorded, and the outcome of interest is Arthritis Self-Assessment, the dichotomized version of the ordinal Arthrtitis Categorical Scale, obtained at baseline and at months 2, 4, and 6 thereafter, as shown below.

1 subject ID
2 treatment (0 = placebo, 1 = auranofin
3 age at baseline (years)
4 month
5 Arthritis Categorical Scale (ordinal)
6 Arthritis Self-Assessment (0 = tolerable, 1 = severe)

Let Y_{ij} be Arthritis Self-Assessment for subject i at month t_{ij}, where $t_{ij} = 0, 2, 4, 6$ months for $j = 1, \ldots, 4$ for all subjects and $i = 1, \ldots, 290$. Let $\delta_i = 0$ if i was assigned to placebo and 1 if assigned to auranofin therapy. Here, there are no within-individual covariates, so that $\boldsymbol{z}_{ij} = t_{ij}$, and the among-individual covariates $\boldsymbol{a}_i = (\delta_i, a_i)$, where a_i is the age of subject i at baseline, so that $\boldsymbol{x}_{ij} = (t_{ij}, \delta_i)^T$ and $\boldsymbol{x}_i = (t_{i1}, \ldots, t_{in_i}, \delta_i, a_i)^T$.

Consider the following two models for these data:

(i) Assume that for each i and $j = 1, \ldots, n_i$

$$E(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = \frac{\exp(\beta_{0i} + \beta_{1i} t_{ij})}{1 + \exp(\beta_{0i} + \beta_{1i} t_{ij})}, \tag{3}$$

where

$$\begin{aligned} \beta_{0i} &= \beta_0 + b_{1i} \\ \beta_{1i} &= \beta_1 + \beta_2 \delta_i; \end{aligned} \tag{4}$$

$b_{1i} \sim \mathcal{N}(0, D)$; and Y_{ij}, $j = 1, \ldots, n_i$, are mutually independent conditional on $\boldsymbol{z}_i$ and b_{i1}.

(ii) Assume that for each i and $j = 1, \ldots, n_i$

$$E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|\boldsymbol{x}_{ij}) = \frac{\exp(\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}\delta_i)}{1 + \exp(\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}\delta_i)}, \tag{5}$$

$$\text{var}(Y_{ij}|\boldsymbol{x}_i) = \text{var}(Y_{ij}|\boldsymbol{x}_{ij}) = E(Y_{ij}|\boldsymbol{x}_{ij})\{1 - E(Y_{ij}|\boldsymbol{x}_{ij})\}, \qquad \text{corr}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i), \tag{6}$$

where $\boldsymbol{\Gamma}(\alpha, \boldsymbol{x}_i)$ is some correlation matrix.

(a) Give interpretations of the parameters β_0, β_1, and β_2 in model (i) given by (3) and (4).

(b) Give interpretations of the parameters β_0, β_1, and β_2 in model (ii) given in (5) and (6).

(c) Consider model (i). An alternative to (4) is

$$\begin{aligned} \beta_{0i} &= \beta_0 + b_{1i} \\ \beta_{1i} &= \beta_1 + \beta_2 \delta_i + b_{2i}, \end{aligned} \tag{7}$$

where now $\boldsymbol{b}_i = (b_{1i}, b_{2i})^T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$, with Y_{ij}, $j = 1, \ldots, n_i$, are mutually independent conditional on $\boldsymbol{z}_i$ and $\boldsymbol{b}_i$.

In this part of this problem, you will get to observe how well various methods for fitting models of the form (3)-(4) and (3)-(7) based on different analytical and numerical ways of approximating the likelihood agree.

Fit each of the models defined by (3) and (4) and (3) and (7), respectively, using the following methods:

- An analytical approximation to the likelihood based on a linear approximation of the model about the random effects equal to zero.
- An analytical approximation to the likelihood based on a linear approximation of the model about the random effects equal to current empirical Bayes estimates.
- A numerical approximation to the likelihood in which the integrals are approximated using a full Laplace approximation.
- A numerical approximation to the likelihood in which the integrals are approximated using adaptive Gaussian quadrature.

Compare the resulting estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ and $D/\boldsymbol{D}$ in each case and comment the results. Which of second stage models (4) or (7) would you prefer to adopt, and which method would you feel most comfortable using?

Hint: It may well be that you will find it difficult to implement all of the fits in (a), especially for second stage model (7).

(d) Fit model (ii) in (5) and (6) taking $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ to be a compound symmetric matrix. Comment on how the results compare to those obtained in (c).

4. *Pharmacokinetics of Hexapinerdone.* In the data set `pk.dat` on the course website you wil find data from 53 subjects suffering from mild to moderate schizophrenia who volunteered to participate in a single dose pharmacokinetic study of the experimental antipsychotic drug hexapinerdone. Schizophrenia is a severe mental disorder that is ordinarily treated with antipsychotic drugs. At time 0, each participant received a single 1000 mg oral dose of the drug, and blood samples were drawn and assayed for hexapinerdone concentrations at several time points during the next 24 hours. The age and weight of each subject were recorded, along with the subject's CYPD26 phenotype; the CYPD26 gene is expressed primarily in the liver and encodes an enzyme, cytochrome P450 2D6, which is implicated in the way an individual metabolizes a drug. Elimination of a drug encompasses both metabolism of the drug and its excretion from the body. If an individual has the poor CYP2D6 phenotype, s/he may exhibit elimination of the drug at a slower rate than an individual who has the extensive (normal) phenotype. Intermediate metabolizers might show elimination rate somewhere between poor and extensive metabolizers. Because it is believed that hexapinerdone's effect is not through the drug itself but through another compound that is the byproduct of metabolism, a so-called metabolite, there is reason to believe that the drug's elimination characteristics might be associated with CYP2D6 phenotype. Poor metabolizers may require higher doses than extensive metabolizers, and extensive metabolizers run the risk of adverse events at too high a dose, so it is important to investigate this possible relationship as well as relationships between any of age, weight, and CYP2D6 phenotype and pharmacokinetic processes more generally.

   The columns of the data set are as follows:

   1 subject ID
   2 time (hours)
   3 hexapinerdone concentration (mg/dL)
   4 weight (kg)
   5 age (years)
   6 CYP2D6 phenotype (1=poor, 2=intermediate, 3=extensive)

As we have discussed, it is well established that drug concentrations at the individual patient level exhibit nonconstant variance that is often represented the "power of the mean" variance model. It is also well established that the distributions of pharmacokinetic parameters in the population tend to be right skewed.

The investigators had the following goals:

(i) Characterizing the typical or mean pharmacokinetic properties of hexapinerdone absorption, distribution, and elimination in this patient population.

(ii) Determining whether or not there is evidence that these pharmacokinetic characteristics are systematically associated with subject characteristics (weight, age, and CYP2D6 phenotype here) and, if so, with which characteristics.

Identify an appropriate pharmacokinetic model and, using it, carry out analyses to address these questions and write a brief report summarizing what you did and the results, following the basic outline for writing a data analysis report in Appendix F of the course notes. As in the guidelines there, be sure to describe how you formalized the questions of interest within the framework of these models and interpret the results in the context of the subject matter. Comment on any limitations or concerns you might have and on how confident you feel about the reliability of the inferences and conclusions.

Please turn in your code and output along with your report (you can edit the output to include only the portions that pertain directly to your report and embed it in your report if you like).

Hint: As is often the case with fitting nonlinear mixed effects models, challenges can arise, and often these challenges are software-specific. It is a good idea to use more than one software implementation to fit these models (e.g., those in SAS and that in R) and to try several different sets of starting values in the fit of a specific model using each software implementation to feel confident in the results. Accordingly, it would be sensible to carry out analyses using implementations in both SAS and R and to compare and try to reconcile the results.