ST 790, Homework 4
Spring 2018

1. *Bayesian view of the linear mixed effects model.* In Bayesian inference, given a random variable $\boldsymbol{Y}$ with density $p(\boldsymbol{y}|\eta)$ depending on parameter η, inference on η is carried out by deriving its posterior distribution given $\boldsymbol{Y}$ under a prior distribution $p(\eta)$ for η; namely

$$p(\eta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\eta)\, p(\eta)}{\int\, p(\boldsymbol{y}|\eta)\, p(\eta)\, d\eta}.$$

The usual Bayesian estimator for η given data $\boldsymbol{Y}$ is the mode of the posterior distribution, which for symmetric distributions is also the mean.

Here, we consider the linear mixed effects model from a Bayesian perspective. Write the model in "stacked notation" as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{b}$ and $\boldsymbol{e}$ are independent of one another and of $\tilde{\boldsymbol{x}}$, with $\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{D}})$ and $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$. Suppose that $\widetilde{\boldsymbol{D}}$ and $\boldsymbol{R}$ are *known*. From a Bayesian point of view, both $\boldsymbol{b}$ and β are regarded as parameters, where the distribution of $\boldsymbol{b}$ is regarded as a prior distribution. To complete the Bayesian view of the model, specify a prior distribution for β,

$$\beta \sim \mathcal{N}(\beta^*, \boldsymbol{H}), \tag{2}$$

for some covariance matrix $\boldsymbol{H}$. We surpress dependence on $\tilde{\boldsymbol{x}}$ for brevity.

Summarizing, from (1) and (2), we have the Bayesian hierarchy

$$\boldsymbol{Y}|\beta, \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b}, \boldsymbol{R}), \qquad \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{D}}), \qquad \beta \sim \mathcal{N}(\beta^*, \boldsymbol{H}). \tag{3}$$

The posterior density of β is then, in obvious notation,

$$p(\beta|\boldsymbol{y}) = \frac{\int p(\boldsymbol{y}|\beta, \boldsymbol{b})\, p(\beta)\, p(\boldsymbol{b})\, d\boldsymbol{b}}{\int\int\, p(\boldsymbol{y}|\beta, \boldsymbol{b})\, p(\beta)\, p(\boldsymbol{b})\, d\boldsymbol{b}\, d\beta}. \tag{4}$$

Similarly, the posterior density of $\boldsymbol{b}$ is

$$p(\boldsymbol{b}|\boldsymbol{y}) = \frac{\int p(\boldsymbol{y}|\beta, \boldsymbol{b})\, p(\beta)\, p(\boldsymbol{b})\, d\beta}{\int\int\, p(\boldsymbol{y}|\beta, \boldsymbol{b})\, p(\beta)\, p(\boldsymbol{b})\, d\boldsymbol{b}\, d\beta}. \tag{5}$$

The Bayesian "estimators" for β and $\boldsymbol{b}$ are then the modes of the posterior distributions (4) and (5), respectively.

Under (3), it is possible to "do" these integrations directly to obtain these posterior distributions; these calculations are rather involved, so we won't do that here. Instead, we will deduce these posteriors from the joint distribution of β, $\boldsymbol{b}$, and $\boldsymbol{Y}$, $p(\beta, \boldsymbol{b}, \boldsymbol{y}) = p(\boldsymbol{y}|\beta, \boldsymbol{b})\, p(\beta)\, p(\boldsymbol{b})$.

(a) It turns out that, by tedious, direct calculations it can be shown that $p(\beta, \boldsymbol{b}, \boldsymbol{y})$ is normal. Find the mean and covariance matrix of this normal distribution.

(b) Using the result in (a), show that the posterior density $p(\beta|\boldsymbol{y})$ in (4) corresponds to a normal distribution with mean

$$E(\beta|\boldsymbol{Y} = \boldsymbol{y}) = \boldsymbol{C}^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y} + \boldsymbol{H}^{-1}\beta^*)$$

and covariance matrix var($\beta|\boldsymbol{Y}=\boldsymbol{y}$) = $\boldsymbol{C}^{-1}$, where $\boldsymbol{V}=\boldsymbol{R}+\boldsymbol{Z}\widetilde{\boldsymbol{D}}\boldsymbol{Z}^T$ as in the notes, where

$$\boldsymbol{C}=\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}+\boldsymbol{H}^{-1}.$$

(c) Using the results in (a) and (b), show that the posterior density $p(\boldsymbol{b}|\boldsymbol{y})$ in (5) corresponds to a normal distribution with mean $E(\boldsymbol{b}|\boldsymbol{Y}=\boldsymbol{y})$ and covariance matrix var($\boldsymbol{b}|\boldsymbol{Y}=\boldsymbol{y}$), and find an expression for $E(\boldsymbol{b}|\boldsymbol{Y}=\boldsymbol{y})$ in terms of $E(\beta|\boldsymbol{Y}=\boldsymbol{y})$ in (b), and find an expression for var($\boldsymbol{b}|\boldsymbol{Y}=\boldsymbol{y}$) in terms of $\boldsymbol{C}$.

(d) If we take $\boldsymbol{H}^{-1}=\boldsymbol{0}$, then this corresponds to a *noninformative prior* for β. Under this condition, using your result in (b), find the posterior distribution of β. Comment on the form of its mean and covariance matrix and what this suggests about the connection between Bayesian and frequentist inference in this model.

(e) As in (d), taking $\boldsymbol{H}^{-1}=\boldsymbol{0}$, so that the prior for β is noninformative, use your result in (c) to find theemacs posterior distribution of $\boldsymbol{b}$ under this condition. Comment on what its mean and covariance matrix suggest about the interpretation of (6.56) and (6.57).

2. *Hepatitis C dynamics for a single subject.* Characterizing the mechanisms governing the interaction between a virus and the immune system taking place within a patient over time is of great interest in the study of infectious diseases such as human immunodeficiency virus (HIV) and hepatitis C virus (HCV) and the development of treatments for these diseases. A key approach is to represent the interplay between virus and immune responses through a series of hypothetical compartments corresponding to populations of viral particles, or *virions*; populations of immune cells that are targeted by or respond to the virus; and mechanisms of the effects of treatments on viral replication and infectiousness. This compartmental representation gives rise to a system of differential equations that can be solved analytically or numerically to yield expressions for observable measures such as *viral load*, roughly a measure of the concentration of virus in the body. Realistic such models rarely lead to closed form analytical expressions for longitudinal viral load, but in the case of short-term dynamics following initiation of treatment, simplifications can be made that do lead to useful closed form expressions.

   For over two decades, interferon-α-2b (IFN) was a standard treatment for HCV infection. Although it can have deleterious adverse effects and has been largely replaced by antiviral agents, it is still used as part of recommended regimens for some patients. The data set `hcv3.dat` on the course webpage contains data from a single acutely-infected participant in an early pharmaceutical study of IFN therapy who began IFN therapy on day 0 (baseline). Viral load (copies HCV RNA/ml), given in the second column of the data set, was ascertained at $n=11$ time points t_j, $j-1,\ldots,n$: baseline, immediately prior to the start of treatment, and at 10 subsequent time points during the next two days (given in the first column).

   As was argued in an early, world-famous paper in the journal *Science* (Volume 282, p. 103), a reasonable model HCV dynamics over the first two days of IFN therapy that can be expressed in a closed form is as follows. Letting $V(t)$ be viral load at time t following initiation of IFN therapy at time t (days) is given by

$$V(t)=V_0[1-\epsilon+\epsilon\exp\{-c(t-t_0)_+\}],\quad 0\le t\le 2,\quad V_0,c,>0,\quad 0<\epsilon\le 1, \tag{6}$$

   where $x_+=x$ for $x>0$, and 0 otherwise; V_0 is the viral load at day 0 (prior to start of therapy); c is the *virion clearance rate*; t_0 is the so-called *pharmacological delay* such that decay in viral load is not seen until IFN has made sufficient progress in distributing through the body;

and ϵ is an *efficacy* parameter. The efficacy parameter has the interpretation that the effect of IFN therapy is to reduce the production of new virions in the system by the fraction $(1-\epsilon)$; if $\epsilon = 1$ then the therapy is said to be "perfect" (which it never is). For this problem, take the pharmacological delay to be known: $t_0 = 0.20$ days. It is well known that viral load measures at the individual patient level exhibit nonconstant variance.

(a) Letting $\boldsymbol{x}_j = t_j$ and Y_j be viral load at t_j, $j = 1, \ldots, n$, assume that $E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta)$, where $f(\boldsymbol{x}_j, \beta)$ is the right hand side of (6), where $\beta = (\beta_1, \beta_2, \beta_3)^T$ is defined such that $V_0 = \exp(\beta_1)$, $c = \exp(\beta_2)$, and

$$\epsilon = \frac{\exp(\beta_3)}{1 + \exp(\beta_3)},$$

where this parameterization enforces positivity of V_0 and c and restricts ϵ to be between 0 and 1. To accommodate the nonconstant variance, adopt the power of the mean variance model

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 f^{2\delta}(\boldsymbol{x}_j, \beta)$$

for $\delta > 0$, and assume that the sampling times t_j are sufficiently intermittent that serial correlation among the Y_j is negligible.

Under this model, use the generalized least squares algorithm (as implemented in the R function `gnls()` in the `nlme` package or using SAS (e.g., the macro using `proc nlin` on the course website) to fit this model to these data and obtain estimates of V_0, c, and ϵ.

Hints: (i) ***Starting values*** are required for β and δ. Finding starting values for nonlinear models is model-dependent and a bit of an art form. You can derive starting values by considering the form of the model and noting that V_0 is determined by the observations during the "delay" period, and c is determined by the decay in viral loads following the "delay." (ii) Viral load values on the order of millions of copies/ml. This makes the parameter V_0 orders of magnitude larger than the others, which can cause numerical difficulties. A common tactic is to rescale viral load so that the parameter values are of similar magnitude (e.g., by dividing by 10^6, so that the result measures millions of copies/ml).

(b) Obtain approximate standard errors to accompany your estimates of V_0, c, and ϵ.

Hint: The model is parameterized in terms of β, so the standard errors output by the software are standard errors for β, not for V_0, c, and ϵ Thus, you need approximate standard errors for a ***transformation*** of the original parameters.

3. *Pretest-Posttest Clinical Trial.* The classic "pretest-posttest" study is one in which subjects are randomly assigned to the interventions to be compared; and the outcome of interest is recorded at baseline, prior to initiation of the assigned interventions and then again at some pre-determined follow-up time after the interventions have been administered.

   The data in the file `leprosy.dat` on the course web page are from such a clinical trial involving 30 patients with leprosy. Each patient was randomized to one of two antibiotics (coded as 1 and 2) or placebo (coded as 0). The outcome was the total number of leprosy bacilli observed at six sites in the body where bacilli tend to congregate and was recorded at baseline, prior to initiation of treatment, and then again after the course of treatment with antibiotic or placebo was completed.

   The investigators wished to determine if treatment with either one or both of the antibiotics reduces the abundance of leprosy bacilli relative to placebo.

   The data set has the following columns:

1 Patient ID
2 Drug
3 Pre-treatment bacilli count
4 Post-treatment bacilli count

Thus, each patient has $n = 2$ outcome measures: for patient i, Y_{i1} is the baseline outcome and Y_{i2} is the follow-up outcome. Let $t_{i1} = 0$ represent baseline and $t_{i2} = 1$ represent follow-up, and let $\delta_i^{(1)} = 1$ if patient i received antibiotic 1 and $\delta_i^{(2)} = 1$ if patient i received antibiotic 2. Letting $\boldsymbol{x}_i = (t_{i1}, t_{i2}, \delta_i^{(1)}, \delta_i^{(2)})^T$, because the outcome is a count and thus nonnegative, a natural model for mean outcome is the loglinear model

$$E(Y_{ij}|\boldsymbol{x}_i) = \exp(\beta_1 + \beta_2 t_{ij} + \beta_3 t_{ij}\delta_i^{(1)} + \beta_4 t_{ij}\delta_i^{(2)}) = f(\boldsymbol{x}_i, \boldsymbol{\beta}), \quad \boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T. \tag{7}$$

Note that this model assumes a common mean, $\exp(\beta_1)$, at baseline ($j = 1$), which is reasonable under randomization; and then allows the mean to be different from baseline at the follow-up time ($j = 2$) in a way that is different depending on treatment received.

Clearly, the pre- and posttest outcomes are correlated.

(a) A natural distributional model for counts is the Poisson distribution. Calculate the sample mean and variance of the outcomes for each treatment at baseline and follow-up. Does the Poisson mean-variance relationship appear to hold for these data?

(b) Take

$$\text{var}(Y_{ij}|\boldsymbol{x}_i) = \sigma^2 f(\boldsymbol{x}_i, \boldsymbol{\beta}), \tag{8}$$

where σ^2 is an overdispersion parameter. Propose a reasonable correlation model for this situation, depending on a parameter α, and write down the covariance matrix $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ resulting from this choice, where $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})^T$.

(c) Consider solving the quadratic estimating equation (8.14) to estimate σ^2 and the correlation parameter α in your model in (b). Assuming your covariance matrix in (b), give the vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_i(\beta, \xi)$, the gradient matrix $\boldsymbol{E}_i(\beta, \xi)$, and the "covariance matrix" $\boldsymbol{Z}_i(\beta, \xi)$ under the "Gaussian working assumption" such that the quadratic estimating equation can be written in the alternative form discussed in Section 8.3 of the notes.

(d) Fit the model defined by (7) and (8) by whatever method you like and use the results to address the investigators' main question and summarize your conclusions. (No need to write a full-blown report.)

4. *Clinical trial in rheumatoid arthritis.* The data in the file `arthritis.dat` on the course webpage are from $m = 290$ participants in a clinical trial comparing auranofin therapy (3 mg of oral gold, twice daily, coded as 1) and placebo (coded as 0) for the treatment of rheumatoid arthritis. Participants were randomized in equal proportions to the two treatments and, at baseline (month 0), prior to initiation of assigned treatment, the outcome, the patient's global impression of his/her current arthritis symptoms, the Arthritis Categorical Scale, was obtained from each participant. This is an example of a "patient reported outcome;" here, a subject rated his/her current arthritis on a five-point ordinal scale: 1 = very good, 2 = good, 3 = fair, 4 = poor, and 5 = very poor. This outcome was then obtained again at months 2, 4, and 6. Also recorded for each subject was his or her age at baseline. In this problem, we will take the outcome of interest to be a dichotomized version of this 5 point scale, which we refer to as Arthritis Self-Assessment, which = 0 if the Arthritis Categorical Scale = 1, 2, or

3, indicating “tolerable” symptoms; and = 1 if Arthritis Categorical Scale = 4 or 5, indicating “severe” symptoms.

The data set has the following columns:

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)

Miraculously, all 290 subjects were evaluated at baseline, and all returned to the clinic for each follow up visit.

The investigators are interested in the following questions: (i) Is there evidence that the probability of severe symptoms changes over the study period for either placebo or auranofin therapy? (ii) Does auranofin therapy lead to a lower probability of severe symptoms than placebo after 6 months in this patient population? What is the probability of having severe symptoms after 6 months for each treatment? (iii) Is there evidence that the probability of severe symptoms at baseline associated with age? (iv) Is any change in probability of severe symptoms over the study period associated with age?

Using methods in Chapter 8 of the notes, 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).