# ST 790, Homework 3
# Spring 2018

1. *Connection with individual estimation.* Consider the linear mixed effects model expressed as a two-stage hierarchy as in (6.42) and (6.43) of the notes. We will focus in this problem on the special case where, for individual $i$, $i = 1, \ldots, m$, the model is

$$\boldsymbol{Y}_i = \boldsymbol{C}_i\beta_i + \boldsymbol{e}_i \ \ (n_i \times 1), \qquad E(\boldsymbol{e}_i|\boldsymbol{x}_i) = \boldsymbol{0}, \ \ \text{var}(\boldsymbol{e}_i|\boldsymbol{x}_i) = \boldsymbol{R}_i \ (n_i \times n_i), \tag{1}$$

for full column rank design matrix $\boldsymbol{C}_i$ $(n_i \times k)$ depending on $t_{i1}, \ldots, t_{in_i}$; and

$$\beta_i = \boldsymbol{A}_i\beta + \boldsymbol{b}_i \ \ (k \times 1), \qquad E(\boldsymbol{b}_i|\boldsymbol{x}_i) = \boldsymbol{0}, \ \ \text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i) = \boldsymbol{D} \ (k \times k), \tag{2}$$

where $\beta$ $(p\times 1)$ is the vector of fixed effects, $\boldsymbol{A}_i$ $(k\times p)$ is a design matrix depending on among-individual covariates $\boldsymbol{a}_i$, and $\boldsymbol{b}_i$ and $\boldsymbol{e}_i$ are independent of each other for each $i = 1, \ldots, m$ and $\boldsymbol{b}_i$, $\boldsymbol{e}_i$, $\boldsymbol{x}_i$ and thus $\boldsymbol{Y}_i$ are all independent across $i$. Thus, substituting (2) in (1) yields the linear mixed effects model

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

where $\boldsymbol{X}_i = \boldsymbol{C}_i\boldsymbol{A}_i$ $(n_i \times p)$ and $\boldsymbol{Z}_i = \boldsymbol{C}_i$ $(n_i \times k)$. Suppose that the matrices $\boldsymbol{R}_i$, $i = 1, \ldots, m$, and $\boldsymbol{D}$ are known and nonsingular.

If we focus on individual $i$ only, under (1), assuming that $n_i > k$, an obvious estimator for $\beta_i$, the individual-specific parameter characterizing $i$'s inherent trajectory, is the weighted least squares estimator based on $i$'s data only, given by

$$\widehat{\beta}_i = (\boldsymbol{Z}_i^T\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i^T\boldsymbol{R}_i^{-1}\boldsymbol{Y}_i.$$

(a) Show that the usual estimator for $\beta$, given here as

$$\widehat{\beta} = \left(\sum_{i=1}^m \boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^m \boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Y}_i, \quad \boldsymbol{V}_i = \boldsymbol{R}_i + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T, \tag{3}$$

can be expressed equivalently as

$$\widehat{\beta} = \left(\sum_{i=1}^m \boldsymbol{A}_i^T\boldsymbol{W}_i^{-1}\boldsymbol{A}_i\right)^{-1}\sum_{i=1}^m \boldsymbol{A}_i^T\boldsymbol{W}_i^{-1}\widehat{\beta}_i \tag{4}$$

for some matrix $\boldsymbol{W}_i$, and give the form of $\boldsymbol{W}_i$.

Thus, the estimator (3) for $\beta$ can be viewed as a weighted average of individual-specific estimators.

(b) Find $E(\widehat{\beta}_i|\boldsymbol{x}_i)$ and $\text{var}(\widehat{\beta}_i|\boldsymbol{x}_i)$, expressing the latter in terms of $\boldsymbol{W}_i$.

(c) Using your results in (b), show that $\widehat{\beta}$ as expressed in (4) is conditionally unbiased (given all covariates $\tilde{\boldsymbol{x}}$) and find its conditional covariance matrix in terms of $\boldsymbol{W}_i$.

(d) Take $\boldsymbol{R}_i = \sigma^2\boldsymbol{I}_{n_i}$ for all $i$, and suppose that the matrix $\boldsymbol{D} = \boldsymbol{0}$. Give the form of $\widehat{\beta}$ under these conditions, and give an intuitive explanation for why this form is what it is.

(e) Take $\boldsymbol{R}_i = \sigma^2\boldsymbol{I}_{n_i}$ for all $i$, and suppose now that the matrix $\boldsymbol{D} = d\boldsymbol{I}_k$ for scalar $d > 0$. Give the form of

$$\lim_{d\to\infty}\widehat{\beta},$$

and give an intuitive explanation for why this form is what it is.

2. *Loglikelihood tricks.* On page 186 of the notes, we indicate that estimation of the parameters $\beta$ and $\boldsymbol{\xi} = \{\gamma^T, \text{vech}(\boldsymbol{D})^T\}^T$ in a linear mixed effects model can be carried out by maximizing the normal loglikelihood (5.31) following from the implied population averaged model (6.45). In principle, this can be accomplished by maximizing in all of $\beta$ and $\boldsymbol{\xi}$ directly using standard optimization techniques. However, in reality, as discussed on pages 134-135 of the notes, to improve computational efficiency and convergence, maximization of this loglikelihood in linear mixed model software, e.g., SAS `proc mixed`, R `lme()`, is implemented by recasting this optimization problem as one of lower dimension. In this problem, we will consider some of the ways this is typically done.

   Consider the particular case of the linear mixed effects model (6.44),

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \quad \boldsymbol{b}_i|\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D}), \quad \boldsymbol{e}_i|\boldsymbol{x}_i \sim \mathcal{N}\mathbf{0}, \sigma^2\boldsymbol{I}_{n_i}),$$

   where $\boldsymbol{Y}_i$ is $(n_i \times 1)$, $\boldsymbol{b}_i$ is $(q \times 1)$, $i = 1, \ldots, m$, and $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are of full column rank. Ordinarily, $q$ is relatively small ($q = 2$ or 3 at the most), while $n_i$ is larger.

   (a) The first standard tactic is to reparameterize the problem by a rescaling so that

$$\boldsymbol{D} = \sigma^2\boldsymbol{D}_*. \tag{5}$$

   Under (5), the unknown parameters are now $\beta$, $\sigma^2$, and vech($\boldsymbol{D}_*$). (Of course, if the loglikelihood is maximized in these parameters, the estimate of $\boldsymbol{D}$ can be obtained from those for $\sigma^2$ and $\boldsymbol{D}_*$.) Under this reparameterization, write down the resulting expression for $\boldsymbol{V}_i = \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$, and obtain an expression for $\sigma^2$ solving (5.38) (and thus maximizing the loglikelihood) in terms of $\beta$ and $\boldsymbol{D}_*$.

   (b) Substitute the expression for $\sigma^2$ you obtained in (a) into the loglikelihood (5.31), ignoring any constants that arise, to obtain an expression depending only on $\beta$ and $\boldsymbol{D}_*$. This expression is often referred to as the *profile* or *variance-profile* loglikelihood.

   *Note:* Because the value of $\sigma^2$ maximizing the full loglikelihood (5.31) has been substituted, maximizing the profile loglikelihood in $\beta$ and $\boldsymbol{D}_*$ will yield the values of $\beta$ and $\boldsymbol{D}_*$ maximizing (5.31). Thus, this trick reduces the dimension of the optimization by 1.

   (c) Assume that $\boldsymbol{D}_*$ is nonsingular (as we'd hope it is!), with inverse $\boldsymbol{D}_*^{-1}$; the inverse of a covariance matrix is often referred to as a precision matrix. Reexpress the profile loglikelihood you found in (b) in terms of $\beta$ and $\boldsymbol{D}_*^{-1}$ to obtain the *precision matrix parameterization* of the profile loglikelihood.

   *Note:* The need to invert a matrix of dimension $(n_i \times n_i)$ in the original profile loglikelihood in (b) has been replaced by the need to invert a matrix of dimension $(q \times q)$. Thus, if $q \ll n_i$ for all $i$, maximizing the profile loglikelihood in this parameterization reduces the computational burden.

   (d) Neither of the tricks in (b) and (c) take advantage of the fact that the value of $\beta$ solving (5.27) (and thus maximizing the loglikelihood) satisfies (5.39). Incorporate this fact into your result in (c) to obtain an expression to be maximized that is a function solely of $\boldsymbol{D}_*^{-1}$. This is referred to as the *full profile* loglikelihood.

   *Note:* Here, the optimization problem has been further reduced to solving for the distinct elements of $\boldsymbol{D}_*^{-1}$ $(q \times q)$, where $q$ is small. Estimates of $\beta$ and $\sigma^2$ can then be reconstructed from that for $\boldsymbol{D}_*$.

3. *HIV clinical trial, continued.* Recall AIDS Clinical Trials Group (ACTG) Study 193A, which was introduced in Section 5.2. Data from $m = 1309$ participants are in the file `cd4.dat` on the class webpage. These subjects were each randomly assigned to one of four treatment groups: 1, 600 mg zidovudine (ZDV) alternating monthly with 400 mg of didanosine (ddI); 2 600 mg ZDV plus 2.25 mg zalcitabine (ZAL); 3, 600 mg ZDV plus 400 mg ddI concurrently; and 4, 600 mg ZDV plus 400 mg ddI lus 400 mf nevirapine (NVP). The three treatment regimens coded as 1–3 are classified as *dual therapy*, as they each involve two antiretroviral agents, while that coded as 4 is classified as *triple therapy*, involving three agents.

   These patients were severely immunocompromised, with CD4 counts of less than 50 cells/mm$^3$. Accordingly, treatment regimens that can maintain CD4 counts at level at baseline, without them falling further, are desirable; treatment regimens that in fact cause CD4 counts to rise are even better.

   CD4 T-cell counts were planned to be measured at baseline (week 0) and at 8 week intervals thereafter to 40 weeks; that is, at weeks 8, 16, 24, 32, and 40. As discussed in class and as evident from the data, most participants did not return to the clinic at exactly these times. The exact times at which CD4 measures were taken are recorded in the data set.

   The data set has the following columns:

| | |
|---|---|
| 1 | patient ID |
| 3 | treatment group (1 – 4 as indicated above) |
| 4 | age (years) |
| 5 | gender (0=female, 1=male) |
| 2 | week |
| 6 | *log*(CD4 + 1), log-transformed CD4 count (log cells/mm$^3$) |

   Some subjects clearly skipped intended visits altogether, and others apparently dropped out of the study without completing all visits. Thus, *be careful* to account for this and to state any assumptions necessary to justify your analyses.

   The investigators were interested in comparing the typical features of patient log CD4 trajectories between dual therapy and triple therapy. They conjectured that the addition of NVP to other the other agents involved in regimens 1–3, as in triple therapy regimen 4, would lead to a different typical pattern of change of these trajectories than for dual therapy. In particular, they hypothesized that the log CD4 trajectories for patients receiving triple therapy would rise during the first 4 months (16 weeks) and then level off and remain relatively constant in the remaining study period. In contrast, they hoped that those for patients receiving dual therapy would at least stay constant (and not drop) during the first 4 months (16 weeks) and that this would continue for the rest of the study period, with no decline in log CD4.

   In particular, they had the following questions:

   (i) Is the typical or mean rate of change of log CD4 during the first 16 weeks different for dual and triple therapy? What is the typical or mean rate of change of log CD4 in the first 16 weeks for each therapy? Is there evidence that the typical or mean rate of change of log CD4 is positive for either or both therapies during the first 16 weeks?

   (ii) Is the typical or mean rate of change of log CD4 from week 16 through the end of the study different for dual and triple therapy? What is the typical or mean rate of change of log CD4 for each therapy during the second part (post-week 16) part of the study? Is there evidence that the typical or mean rate of change is different from zero for either group?

(iii) What is the mean log CD4 count for each group at the end of the study (week 40)? Are there differences in mean log CD4 at the end of the study among the treatments?

(iv) Is there evidence that typical or mean log CD4 at baseline is associated with age or gender? Is the typical or mean rate of change in the first 16 weeks associated with age or gender for either therapy? Is the typical or mean rate of change in the second part of the study (post-16 weeks) associated with age or gender?

These questions are subject-specific in nature. Using methods in Chapter 6 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).

*Note:* Given the nature of the questions, it is reasonable to combine groups 1–3 into a single dual therapy group.