BIO 245: Analysis of Longitudinal Data
Homework #2
Due Oct. 18, 2023

Question 1:

Continuing our investigations with the MACS data, the `MACS-VL.RData` dataset on the course website has longitudinal information on CD4+ cell counts for K=225 MACS participants with baseline viral load data. In this question we are going to consider the relationship between baseline viral load and the rate of decline of CD4 count.

(a) Summarize the key variables using simple numerical and/or graphical summaries as relevant to the scientific question of interest.

(b) Use appropriate exploratory methods to characterize the covariance structure of the data. What structured covariance model(s) appear plausible/reasonable?

(c) Use the `gls()` command in the `nlme` library to fit the model:

$$\mathrm{E}[Y_{ki}] = \beta_0 + \beta_1 t_{ki} + \beta_2 x_k + \beta_3 t_{ki} x_k$$

where x_k is the (possibly transformed) baseline viral load and t_{ki} is time since seroconversion in months. Use compound symmetric correlation but consider both maximum likelihood and restricted maximum likelihood for estimation. Present your results in a concise manner that would be suitable for a journal and provide a precise interpretation of the estimates for the mean model. Ccomment on whether there is a significant association between baseline viral load and the rate of decline in CD4+ based on the estimates from this model.

(d) The model in part (c) restricts the analysis in that it is estimating a linear relationship between (possibly-transformed) baseline viral load and CD4 count over time. As a way of relaxing this restriction, consider categorizing baseline viral load. Given a categorization with J levels, one alternative to the model in part (c) is

$$\mathrm{E}[Y_{ki}] = \beta_0 + \beta_1 t_{ki} + \sum_{j=2}^{J} \beta_{2j} x_k(j) + \sum_{j=2}^{J} \beta_{3j} t_{ki} x_k(j)$$

where $x_k(j)$ is a dummy variable that indicating membership in the j^{th} category. Again using a compound symmetric correlation structure but only using restricted maximum likelihood,

fit the model and present your results in a concise manner that would be suitable for a journal. Provide a precise interpretation of the estimates in this regression model, and comment on whether there is a significant association between baseline viral load and the rate of decline in CD4+ based on the estimates from this model.

(e) (optional) The models in parts (c) and (d) can be viewed as special cases of a 'varying coefficient' model:

$$\mathrm{E}[Y_{ki}] = \gamma_0(x_k) + \gamma_1(x_k)t_{ki},$$

in which the slope for time depends on the value of the covariate. Specifically, while the model in (c) assumes that $\gamma_1(x_k) = \beta_1 + \beta_3 x_k$, the model in (d) assumes a discrete function where $\gamma_1(x_k) = \beta_1 + \sum_j \beta_{3k} x_k(j)$. Hence, the model in (c) utilizes viral load in its continuous form, but is restrictive in the nature of the relationship (i.e. linearity), the model in (d) utilizes a categorial version of viral load but makes no assumptions regarding the functional form of how the rate of decline differs across the viral load categories.

Beyond these two special cases, allowing $\gamma_0(x_k)$ and $\gamma_1(x_k)$ to consider richer functional forms than the linear form used in the model in (c) provides a more flexible description of how the rate of decline differs for different values of baseline viral load. With this in mind, use a varying coefficient model for the rate of decline in CD4+ that characterizes how the rate of decline depends on baseline viral load. I recommend that you use natural or restricted cubic splines for the coefficient functions and simply choose two knots. Plot the estimated coefficient function $\hat{\gamma}_1(x_k)$ with pointwise 95% confidence bands, and interpret specific values. What does this plot suggest about the adequacy of the model in (c)?

Question 2 (Optional):

Consider the one-way analysis of variance model:

$$Y_{ki} = \mu + \gamma_k + \epsilon_{ki},$$

with $i = 1, \ldots, n$ replicates on $k = 1, \ldots, K$ units and

$$\begin{aligned} \gamma_k &\sim \text{Normal}(0,\, \tau^2), \\ \epsilon_{ki} &\sim \text{Normal}(0,\, \sigma^2), \\ \gamma_k &\perp \epsilon_{ki}. \end{aligned}$$

The following may be useful: Let $\boldsymbol{I}_m$ denote the $m \times m$ identity matrix and $\mathbf{1}_m$ denote the $m \times 1$ vector of 1's. Then:

$$(a\boldsymbol{I}_m + b\mathbf{1}_m)^{-1} = \frac{1}{a}\left(\boldsymbol{I}_m - \frac{b}{a+mb}\mathbf{1}_m\right)$$

for $a \neq 0$ and $a \neq -mb$ and:

$$|a\boldsymbol{I}_m + b\mathbf{1}_m| = a^{m-1}(a+mb)$$

(a) Derive the likelihood and log-likelihood as a function of $(\mu,\, \sigma^2,\, \tau^2)$.

(b) Show that the MLEs for μ, σ^2, and τ^2 are given by:

$$\begin{aligned} \hat{\mu} &= \bar{Y}_{\cdot\cdot} \\ \hat{\sigma}^2 &= \text{MSE} \\ \hat{\tau}^2 &= \frac{(1-1/n)\text{MSA} - \text{MSE}}{m} \end{aligned}$$

where $\text{MSA} = n\sum_i(\bar{Y}_{k\cdot} - \bar{Y}_{\cdot\cdot})^2/(K-1)$ and $\text{MSE} = \sum_k\sum_i(Y_{ki} - \bar{Y}_{k\cdot})^2/[K(n-1)]$. Hint: It may be helpful to write $\lambda = \sigma^2 + n\tau^2$.

(c) Obtain the form for $\text{Var}[\hat{\mu}]$ and hence an estimate of this quantity.

(d) Find the REML estimators for σ^2 and τ^2 by integrating μ out of the likelihood in part (a).

(e) In the one-way random effects model with balanced data, it can be shown that:

$$\frac{\text{MSA}/(\sigma^2 + m\tau^2)}{\text{MSE}/\sigma^2} \sim F_{K-1,\, K(n-1)}$$

where $F_{K-1,\, K(n-1)}$ denotes the F distribution with $K-1$ and $K(n-1)$ degrees of freedom. Hence explain why $F^\star = \text{MSA}/\text{MSE}$ may be compared to an $F_{K-1,\, K(n-1)}$ to test the hypothesis H: $\tau^2 = 0$.