.

ST 732, HOMEWORK 4, SPRING 2007

1. Consider a straight line model for individual behavior as in Equation (9.1) of the notes, which for unit i is of the form

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}, \tag{1}$$

where Y_{ij} is the random variable representing the observation that might be seen for unit i at time t_{ij}; $j = 1, \ldots, n_i$ indexes the time points for unit i; β_{0i} and β_{1i} are the unit-specific intercept and slope, respectively, dictating the "inherent trajectory" for unit i; and e_{ij} is a mean-zero random deviation representing how Y_{ij} deviates from the inherent trajectory. Let

$$\boldsymbol{\beta}_i = \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix}$$

be the vector of unit-specific parameters for individual i in model (1), and let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ denote the random vector of observations on i, with $\boldsymbol{e}_i$ defined similarly.

(a) If we write (1) in the form $\boldsymbol{Y}_i = \boldsymbol{Z}_i \boldsymbol{\beta}_i + \boldsymbol{e}_i$, give the form of $\boldsymbol{Z}_i$ if $n_i = 5$. e (b) Now suppose that units arise from 4 populations, labeled A, B, C, and D. Write down a second-stage population model that allows each population to have its own mean intercept and slope $\beta_{0,k}$ and $\beta_{1,k}$, respectively, where $k = A, B$, C, or D about which unit-specific intercepts and slopes vary in each population. Express your model in the form in Equation (9.5) in the notes; i.e.,

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i \boldsymbol{\beta} + \boldsymbol{b}_i,$$

where $\boldsymbol{\beta} = (\beta_{0,A}, \ldots, \beta_{0,D}, \beta_{1,A}, \ldots, \beta_{1,D})'$. Define $\boldsymbol{b}_i$ and give the form of $\boldsymbol{A}_i$ when unit i is from each of populations A, B, C, and D, respectively.

(c) Define

$$\begin{aligned}
\delta_{Ai} &= 1 \text{ if unit } i \text{ is from population A} \\
&= 0 \text{ otherwise} \\
\delta_{Bi} &= 1 \text{ if unit } i \text{ is from population B} \\
&= 0 \text{ otherwise} \\
\delta_{Ci} &= 1 \text{ if unit } i \text{ is from population C} \\
&= 0 \text{ otherwise} \\
\delta_{Di} &= 1 \text{ if unit } i \text{ is from population D} \\
&= 0 \text{ otherwise}
\end{aligned}$$

Express the $\boldsymbol{A}_i$ matrices found in (b) compactly by giving form of $\boldsymbol{A}_i$ for any unit i in terms of $\delta_{Ai}, \delta_{Bi}, \delta_{Ci}, \delta_{Di}$.

(d) As shown on pages 321–322 of the notes, the model under the conditions in (a)–(c) can be expressed as

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

Give the form of $\boldsymbol{X}_i$ for a unit i with $n_i = 5$ if the unit is from populations A and D, respectively.

(e) Give the form of $\boldsymbol{X}_i$ for any unit i in terms of $\delta_{Ai}, \delta_{Bi}, \delta_{Ci}, \delta_{Di}$ defined in (c). Note that writing $\boldsymbol{X}_i$ this way corresponds to how we think about how `model` statements in `proc mixed` and `proc glm` are constructed under the "explicit parameterization" (see Section 8.9 of the class notes).

2. Consider the random coefficient model with individual first stage model

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij},$$

where individual i is observed at times $t_{i1}, \ldots, t_{in_i}$, $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})'$, and

$$\text{var}(\boldsymbol{e}_i) = \sigma^2 \boldsymbol{I}_{n_i};$$

and population second stage model

$$\beta_{0i} = \beta_0 + b_{0i}, \quad \beta_{1i} = \beta_1 + b_{1i}, \quad \boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix},$$

where

$$\text{var}(\boldsymbol{b}_i) = \boldsymbol{D} = \begin{pmatrix} D_{11} & D_{12} \\ D_{12} & D_{22} \end{pmatrix},$$

and $\boldsymbol{b}_i$ is statistically independent of $\boldsymbol{e}_i$ as on p. 320 of the note.

(a) Use results on variances and covariances covered earlier in the course to demonstrate that

$$\text{var}(Y_{ij}) = D_{11} + D_{22} t_{ij}^2 + 2 D_{12} t_{ij} + \sigma^2, \quad \text{cov}(Y_{ij}, Y_{ik}) = D_{11} + D_{22} t_{ij} t_{ik} + D_{12}(t_{ij} + t_{ik}),$$

thus verifying a generalization of the result at the top of p. 329 of the notes.

(b) Suppose that $D_{12} = 0$, so that b_{0i} and b_{1i} are uncorrelated. Are Yij and Y_{ik} correlated under this condition? Explain.

(c) Suppose instead that $\text{var}(\boldsymbol{e}_i) = \sigma_1^2 \boldsymbol{\Gamma}_i + \sigma_2^2 \boldsymbol{I}_{n_i}$, where $\boldsymbol{\Gamma}_i$ is the $(n_i \times n_i)$ Markov correlation model with parameter $\rho > 0$. Find $\text{var}(Y_{ij}$ and $\text{cov}(Y_{ij}, Y_{ik})$ in this case, where all the other conditions given above still hold.

3. Recall the lead level study from Homework 3, Problem 3. Suppose that a new group of investigators studying treatment of lead exposure asked the original investigators for their data. This new group is took a different approach to modeling these data. In particular, as an initial model, they ignored the age and gender variables and considered the random coefficient model with straight-line first stage

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}$$

for child i, where we may define $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i})'$ for child i. They assumed that, for treatment $k = 1, 2, 3$, where $k = 1$ is placebo, $k = 2$ is low-dose succimer, and $k = 3$ is high-dose succimer, $\beta_{0,k}$ is the "typical" mean value of intercepts β_{0i} and $\beta_{1,k}$ is the "typical" mean value of slopes β_{1i} for children receiving treatment k.

Define $\boldsymbol{\beta} = (\beta_{0,1}, \beta_{0,2}, \beta_{0,3}, \beta_{1,1}, \beta_{1,2}, \beta_{1,3})'$. Then the investigators assumed the second stage population model is

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i \boldsymbol{\beta} + \boldsymbol{b}_i, \quad \boldsymbol{b}_i = (b_{0i}, b_{1i})',$$

and $\boldsymbol{A}_i$ is the appropriate design matrix for child i that "picks off" the correct mean intercept and slope from $\boldsymbol{\beta}$ corresponding to the treatment i took.

The investigators were ultimately interested in learning whether the patterns of blood lead levels over the study period were different depending on treatment. In particular, they were interested in whether there is evidence that the "typical" mean slopes were not all the same.

(a) From the spaghetti plots shown in Homework 3, do you think that the assumption that blood lead levels for children in each treatment group follow "inherent trajectories" that may be represented by child-specific straight lines seems reasonable?

(b) The investigators were willing to assume the following:

(i) The assay used to ascertain blood lead levels from blood samples collected from the children committed errors whose magnitude is unrelated to the lead level in the sample being measured; and

(ii) Lead level samples were taken sufficiently far apart in times that correlation due to local within-child fluctuations in lead levels was negligible, and the magnitude of such fluctuations was constant over time for all treatments. The magnitudes of such fluctuations are independent of the magnitude of the true lead levels.

In developing their model further, the investigators wanted to investigate the following:

(iii) whether the magnitudes of within-child fluctuations in lead levels are the same for all treatments (they constant for all treatments, but are they the same?)

(iv) whether the way in which child-specific intercepts and slopes vary and co-vary are the same under the three treatments.

Using `proc mixed`, fit using REML three different versions of the random coefficient model, all of which incorporate assumptions (i) and (ii) above but allow different assumptions about (iii) and (iv), namely:

- Magnitude of within-child fluctuations in lead level and the way child-specific intercepts and slopes vary/co-vary are both *the same* under all three treatments
- Magnitude of within-child fluctuations in lead levels are possibly *different* under different treatments, but the way child-specific intercepts and slopes vary and covary is *the same*
- Magnitude of within-child fluctuations in lead levels is *the same* under all treatments but the way in which child-specific intercepts and slopes vary/co-vary are possibly *different*.
- *Both* the magnitude of within-child fluctuations in lead levels and the way in which child-specific intercepts and slopes vary/co-vary are possibly *different* across treatments.

(c) From inspection of AIC and BIC for each model fit, which set of assumptions on within-child fluctuations and among-child variation/covariation in intercepts/slopes do you prefer?

(d) Under the model that embodies the assumptions you chose in (c), is there evidence to suggest that the "typical" mean slopes of blood lead level patterns for the three treatments are not the same? To address this, include an appropriate `contrast` statement in the fit of your preferred model and obtain the Wald test statistic. State the value of the statistic, the associated p-value, and your conclusion regarding the strength of the evidence supporting the contention that the mean slopes differ.