Homework II (2022)

1. Write the following models in matrix notation and in each case determine the marginal mean and variance of y. If a factor is not specified, assume it is fixed.

   (a) $y_{ij}|a_i \sim^{indep.} N(\mu + a_i + \beta_j, \sigma^2)$; $a_i \sim^{iid} N(0, \sigma_a^2)$ for $i = 1, 2, \ldots, m$; $j = 1, 2 \ldots, n$. (Two-way mixed model).

   (b) $y_{ij}|a_i, b_j \sim^{indep.} N(\mu + a_i + b_j, \sigma^2)$; $a_i \sim^{iid} N(0, \sigma_a^2)$; $b_j \sim^{iid} N(0, \sigma_b^2)$; a_i and b_j are independent for $i = 1, 2, \ldots, m$; $j = 1, 2 \ldots, n$. (Two-way random model).

   (c) $y_{ijk}|a_i, g_{ij} \sim^{indep.} N(\mu + a_i + \beta_j + g_{ij}, \sigma^2)$; $a_i \sim^{iid} N(0, \sigma_a^2)$; $g_{ij} \sim^{iid} N(0, \sigma_g^2)$; a_i and g_{ij} are independent for $i = 1, 2, \ldots, m$; $j = 1, 2 \ldots, n$; $k = 1, 2, \ldots, r$. (Two-way mixed model with interaction).

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

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

where subject i is observed at times $t_{i1}, \ldots, t_{in_i}$, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T$, and

$$var(\epsilon_i) = \sigma^2 I_{n_i}$$

and population second stage model:

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

where

$$var(b_i) = D = \begin{pmatrix} D_{11} & D_{12} \\ D_{12} & D_{22} \end{pmatrix},$$

and b_i is statistically independent of ϵ_i.

   (a) Use results on variances and covariances to demonstrate that

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

   thus verifying a generalization of the result.

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

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

3. Generate correlated data with different forms for the correlation (covariance) structure. Let the number of observations per subject being $n_i = 10$, evaluate at times $t_{ij} = j$ for $j = 1, 2, \cdots, 10$. Use the mean model:

$$E(Y_{ij}|X_i) = \beta_0 + \beta_1 t_{ij}.$$

For each of the scenarios below generate vectors, $Y_i = (Y_{i1}, \cdots, Y_{i10})$ with the specified covariance structure. Generate data for $m = 25$ subjects for each scenario, using parameter value $\beta = (10.0, 1.0)$.

(a) **Random Intercepts**: To introduce correlation we assume that each subject has its own intercept. The complete model is given by:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_{0,i} + \epsilon_{ij}, \\ b_{0,i} &\sim N(0, \tau^2), \\ \epsilon_{ij} &\sim N(0, \sigma^2), \end{aligned}$$

where $b_{0,i}$ and ϵ_{ij} are mutually independent.
- Give the general form for the covariance matrix $\Sigma = Cov(Y_i)$.
- Give Y_i and plot (lines) versus t_i for $m = 25$ using $\sigma = 1.0$ and $\tau = 1.0$.
- Give Y_i and plot (lines) versus t_i for $m = 25$ using $\sigma = 1.0$ and $\tau = 2.0$.
- Give Y_i and plot (lines) versus t_i for $m = 25$ using $\sigma = 1.0$ and $\tau = 5.0$.

(b) **Random Intercepts and Slopes**: To introduce correlation we assume that each subject has its own intercept and slope. The complete model is given by:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_{0,i} + b_{1,i} t_{ij} + \epsilon_{ij}, \\ b_j &\sim N(0, D), \\ \epsilon_{ij} &\sim N(0, \sigma^2), \end{aligned}$$

where $b = (b_{0,i}, b_{1,i})$ and ϵ_{ij} are mutually independent.
- Give the general form for the covariance matrix $\Sigma = Cov(Y_i)$.
- Generate Y_i and plot (lines) versus t_i for $m = 25$ using $\sigma = 1.0$,

$$D = \begin{pmatrix} 2.0 & 0 \\ 0 & 2.0 \end{pmatrix}.$$

- Generate Y_i and plot (lines) versus t_i for $m = 25$ using $\sigma = 1.0$,

$$D = \begin{pmatrix} 2.0 & -0.2 \\ -0.2 & 2.0 \end{pmatrix}.$$

- Generate Y_i and plot (lines) versus t_i for $m = 25$ using $\sigma = 1.0$,

$$D = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.4 \end{pmatrix}.$$

(c) **Serial Correlation**: To introduce correlation we assume that each subject has his own "process" that is serially correlated. The complete model is given by:

$$\begin{aligned}
&Y_{ij} = \beta_0 + \beta_1 t_{ij} + W_i(t_{ij}) + \epsilon_{ij}, \\
&W_i \sim N(0, D), \\
&Var[W_i(t_{ij})] = \tau^2, \\
&Cov[W_i(t_{ij}), W_i(t_{ik})] = \tau^2 \rho^{|t_{ij} - t_{ik}|}, \\
&\epsilon_{ij} \sim N(0, \sigma^2),
\end{aligned}$$

where W_i and ϵ_{ij} are mutually independent.

- Give the general form for the covariance matrix $\Sigma = Cov(Y_i)$.
- Generate Y_i and plot (lines) versus t_i for $m = 25$ using $\sigma = 1.0$, $\tau = 2.0$, and $\rho = 0.7$.
- Generate Y_i and plot (lines) versus t_i for $m = 25$ using $\sigma = 1.0$, $\tau = 2.0$, and $\rho = 0.9$.
- Generate Y_i and plot (lines) versus t_i for $m = 25$ using $\sigma = 2.0$, $\tau = 2.0$, and $\rho = 0.9$.

4. Exposure to lead can produce a variety of adverse health effects in infants and children, including hyperactivity, hearing or memory loss, learning disabilities, and damage to the nervous system. Although the use of lead as a gasoline additive has been discontinued in the US, so that airborne lead levels have been reduced dramatically, a small percentage of children continue to be exposed to lead at levels that can produce such health problems. Much of this exposure is due to deteriorating lead-based paint that may be chipping and peeling in older homes. Lead-based paint in housing was banned in the US in 1978; however, many older homes (built pre-1978) do contain lead-based paint, and chips and dust can be ingested by young children living in these homes during normal teething and hand-to-mouth behavior. This is especially a problem among children in deteriorating, inner-city housing. The US Centers of Disease Control and Prevention (CDC) has determined that children with blood levels above 10 micrograms/deciliter ($\mu g/dL$) of whole blood are at risk of adverse health effects.

   Luckily, there are so-called chelation treatments that can help a child to excrete the lead that has been ingested. The researchers were interested in evaluating the effectiveness of one such chelating treatment, succimer, in children who had been exposed to what the CDC views as dangerous levels of lead. They conducted the following study. 120 children aged 12-36 months with confirmed blood levels of > 15 $\mu g/dL$ and 40 $\mu g/dL$ in a large, inner-city housing project were identified; these lead levels are above the at-risk threshold determined by the CDC. A clinic was set up in the housing project staffed by personnel from the city's Department of Public Health. The personnel randomized the children into three groups: 40 children were assigned at random to receive a placebo (an inactive agent with no lead-lowering properties), 40 children were assigned at random to receive a low dose of succimer, and 40 children were assigned at random to receive a higher dose of succimer. Blood lead levels were measured at the clinic for each child at baseline (time 0), prior to initiation of the assigned treatments. Then assigned treatment was started, and ideally, each child was to return to the clinic at weeks 1, 2, 4 and 8. At each visit, blood lead level was measured for each child.

The data are available in the file **lead.dat** on icampus. The data are presented in the form of one data record per observation; the columns of the data set are as follows:

1 Child id; 2 Indicator of age (=0 if $\leq$ 24 months; =1 if $>$ 24 months); 3 Gender indicator (=0 if female, =1 if male); 4 Week; 5 Blood lead level ($\mu g/dL$); 6 Treatment indicator (=0 if placebo, =1 if low dose, =2 if higher dose)

Let Y_{ij} denote the jth lead level measurement on the ith child at time t_{ij} for that child, $j = 1, \ldots, n_i$. Note that t_{ij} for each child and n_i may be different. We consider the random coefficient model with straight-line first stage

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

for child i, where we may define $\beta_i = (\beta_{0i}, \beta_{1i})^T$ 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, β_{0k} is the "typical" mean value of intercepts β_{0i} and β_{1k} is the "typical" mean value of slopes β_{1i} for children receiving treatment k.

Define $\beta = (\beta_{01}, \beta_{02}, \beta_{03}, \beta_{11}, \beta_{12}, \beta_{13})^T$. Then the investigators assumed the second stage population model is

$$\beta_i = A_i \beta + b_i, \quad b_i = (b_{0i}, b_{1i})^T,$$

and A_i is the appropriate design matrix for child i that "picks off" the correct mean intercept and slope from β 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) Using spaghetti plots, do you think that the assumption that blood lead levels for children in each treatment group follow "'inherent trajectories' that maybe 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 in 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 leas 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.

Fit 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 co-vary 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 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.