Biostatistics 140.653
Third Term, 2022
Problem Set 3

Instructions: Feel free to discuss and complete the analysis with other students. However, each student must write-up their own solutions. Write as if for a scientific journal. Be brief and accurate. Submit your text answers along with your code in an html or pdf file generated via RMarkdown.

Due in CoursePlus drop box: Friday, March 11 by 12:00pm (noon) EST

For this problem set, use the complete Nepal Anthropometry Study (NAS) Dataset with up to 5 measurements on each child over time.

The goals of the analysis are to:

1) Determine if the average growth rates of children differ by mother's parity (number of previous live births)
2) Estimate the population variation in weights of 6-month old Nepali children and estimate the population variation in annual growth rates of Nepali children

Part I: Get familiar with the data

1. Make a table of mother's parity (*alive* variable). Ideally, we would compare children of nulliparous women to categories of women of parity > 0. However, in this dataset, there are only 19 children from nulliparous women. So, we will create two categories of women: parity ≤ 3 (i.e. 1 to 4 live births) vs. parity > 3 (5 or more live births).

2. Make a spaghetti plot of children's weight as a function of age; connecting the measured weights within a child over time. Color code the data by parity group. Add smoothing splines for each parity group. Note any similarities or differences in the growth rates across the two parity groups.

Part II: Model checking and recommendations

Fit the following model to the data:

$Y_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 I(parity_i > 3) + \beta_3 I(parity_i > 3) age_{ij} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2)$, $Cov(\varepsilon_{ij}, \varepsilon_{ik}) = 0$,

where *i* indicates the child (*i* =1, …, 200) and *j* denotes the follow-up (*j* = 1, 2, 3, 4, 5).

1. Conduct appropriate checking of this model; i.e. check for appropriateness of the mean model, and the independence and constant variance assumptions for the residuals.

2. Based on your model checking, propose an alternative model for the data that can address the first goal of the analysis (i.e. determine if the growth rates of children differ by mother's parity (number of previous live births) while satisfying the observed patterns in data with respect to the mean model and distribution of residuals. NOTE: If you modify the mean model, you may want to iterate between model checking for the mean.

Part III: Marginal model for longitudinal data

1. Use the *gls* function in R to fit the model you proposed in Part I. From the fit of the model, compute the estimated $Corr(\varepsilon_{i0}, \varepsilon_{ij})$ for *j = 1, 2, 3, 4* where the follow-up visits (fuvisit) have values 0 (baseline) and 1, 2, 3, 4 (representing the 4 follow-up visits each 4 months apart).

2. Conduct a likelihood ratio test to address the first goal of the analysis; i.e. to determine if the average growth rates of children differ by mother's parity (number of previous live births).

3. Fit the mean model you proposed in Part I using the *gee* function but where you allow the correlation structure to be "independence". The *gee* function will produce standard error estimates assuming the independence assumption (labeled as "naïve" or "model-based" standard error estimates) and "robust" standard error estimates (using the Huber-White sandwich estimator). Compare the estimated coefficients and standard errors from the *gls* and *gee* model fits.

   HINT:
   *fit = gee(wt~ns(age,2) * parity, data=data,id = id, corstr="independence")*
   summary(fit)$coefficients
   sqrt(diag(fit$naive.variance))
   sqrt(diag(fit$robust.variance))

4. The bootstrap procedure can also be applied to longitudinal or clustered data to estimate standard errors of estimated coefficients (or functions of). To preserve the within-subject dependency, the bootstrap procedure samples children (with replacement) as opposed to assessments. See the ProblemSet3.rmd file for code to implement a clustered bootstrap. Compute the bootstrap standard error estimates and compare these to the standard errors from the *gls* and *gee* model fits. Comment on similarities and differences.

Part IV: Linear mixed model

1. Fit the following two models using the *lme* function in R.

Random Intercept Model:

$$Y_{ij} = (\beta_0 + u_{0i}) + \beta_1(age_{ij} - 6) + \beta_2\left(age_{ij} - 24\right)^+ + \beta_3\left(age_{ij} - 48\right)^+ + \beta_4 I(parity_i > 3) + \beta_5 I(parity_i > 3)(age_{ij} - 6) + \beta_6\left(age_{ij} - 24\right)^+ I(parity_i > 3) + \beta_7\left(age_{ij} - 48\right)^+ I(parity_i > 3) + \varepsilon_{ij},$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$, $Cov\left(\varepsilon_{ij}, \varepsilon_{ik}\right) = 0$ for all $j \neq k$, $u_{0i} \sim N(0, \tau_0^2)$, and $Cov\left(\varepsilon_{ij}, u_{0i}\right) = 0$ for all *j*.

Random Intercept and Random Slope for Age Model:

$$Y_{ij} = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})(age_{ij} - 6) + \beta_2\left(age_{ij} - 24\right)^+ + \beta_3\left(age_{ij} - 48\right)^+ + \beta_4 I(parity_i > 3) + \beta_5 I(parity_i > 3)(age_{ij} - 6) + \beta_6\left(age_{ij} - 24\right)^+ I(parity_i > 3) + \beta_7\left(age_{ij} - 48\right)^+ I(parity_i > 3) + \varepsilon_{ij},$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$, $Cov\left(\varepsilon_{ij}, \varepsilon_{ik}\right) = 0$ for all $j \neq k$, $u_{0i} \sim N(0, \tau_0^2)$, $u_{1i} \sim N(0, \tau_1^2)$, $Cov(u_{0i}, u_{1i}) = \tau_{01}$, and $Cov\left(\varepsilon_{ij}, u_{0i}\right) = 0$, $Cov\left(\varepsilon_{ij}, u_{1i}\right) = 0$, for all *j*.

2. Interpret the random intercept variance from the Random Intercept Model.

3. Interpret the random intercept and random slope for age variance from the Random Intercept and Random Slope for Age Model.

4. Obtain the fitted values from the models and make a plot with 3 panels comparing the observed data (spaghetti plot) and the fitted values from both the Random Intercept and Random Intercept and Random Slope for Age Models (i.e. spaghetti plots of fitted values). Comment on which of the two linear mixed models you think is most consistent with the data.

5. <u>OPTIONAL</u> Using the fit of the Random Intercept and Random Slope for Age Model, estimate the variance in weights of children at 6-months, 12-months and 36-months, from the parity ≤ 3 group.

Part V: Summarize your findings

Write a brief report with sections: objective, data, methods, results, summary as if for a health services journal.