

Biostatistics 140.653
Third Term, 2022
Problem Set 3 REVISED on 3/4/2022

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 annual growth rates of Nepali children and explore whether this differs by mother's parity

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 ($i = 1, \dots, 200$) 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 motivation!

Linear mixed models allow us to specify subject-specific regression models and then subsequently describe how the parameters from subject-specific regression models vary in the population of interest.

- We imagine that the subjects in our sample are representative of persons from the population of interest, e.g. Nepali children from birth to 5 years of age
- The data we get to observe for each subject is generated from a subject-specific regression line, for example:

$$wt_{ij} = \beta_{0i} + \beta_{1i}age_{ij} + \varepsilon_{ij},$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$, $Cov(\varepsilon_{ij}, \varepsilon_{ik}) = 0$ for all $j \neq k$

This model says that each child has their own linear growth in weight for ages 0 to 60 months.

Further the model above says that the weights we observe at a given age are measured with error (ε_{ij}), i.e. random fluctuations in an individual child's weights. These random fluctuations are independent of each other over time and have constant variance.

- Now we can think of the population of all children, where each child has their own intercept (β_{0i}) and linear growth rate (β_{1i}). In the population, there would be a population average intercept (β_0) and population average growth rate (β_1) PLUS measures of how variable the intercepts and linear growth rates are across children!

At the population, we have:

$$\begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right),$$

where $E(\beta_{0i}) = \beta_0$, $E(\beta_{1i}) = \beta_1$, $Var(\beta_{0i}) = \sigma_0^2$, $Var(\beta_{1i}) = \sigma_1^2$,
 $Cov(\beta_{0i}, \beta_{1i}) = \sigma_{01} = \sigma_{10}$

- From the model, we can describe the population average growth rate (β_1) but also how growth rates for individual children vary with respect to the average growth rate (i.e. σ_1^2).

Based on the assumptions above, we could say that 95% of children will have linear growth during the first 5 years of life that range from $\beta_1 \pm 1.96 \sigma_1$

Absent the knowledge of how to fit the model described above (which you will have after Tuesday's lecture), we can conduct some simple intuitive analyses that would allow us to explore variation in growth rates.

In what follows, we will assume that growth is linear. This is a strong assumption that is likely violated but will keep the analyses simple ☺

1. Fit a simple linear regression of weight on age (linear) for each child in the sample and save the estimated slope.
2. Scale the estimated slopes to represent the expected change in weight per year.
3. Plot the expected change in weight per year as a function of the child's baseline age (i.e. age when fuvisit = 0). Describe any patterns you observe in
 - a. The population average change in weight per year as a function of the child's baseline age
 - b. Variation in the expected change in weight per year across children as a function of the child's baseline age (be quantitative, i.e. estimate the variance)
4. Repeat 3) but stratify by mother's parity.

Part V: Summarize your findings

Write a **brief** report with sections: objective, data, methods, results, summary as if for a health services journal. You may include up to 2 figures (which may have multiple panels). Remember to be enumerate when possible!