Problem Set 6: Longitudinal data and three level models

S-043/Stat-151 (Fall 2023)

Overview

This problem set is to practice more complex longitudinal growth models and also work on three level models.

Skills to be developed

This problem set is designed to help you develop the following skills:

- Fitting and interpreting quadratic and piecewise growth models.
- Plotting and visualizing longitudinal data.
- Thinking about how models are approximations, and how to interpret estimates as averages of complex processess over time.

Before you start

Please do the following

1. Review the reading for Units 1-4, and the material on three-level models[1]
2. Read through the problem before starting so you have a sense of where you are going.

[1] You can do the first part of the assignment without three level models, if you want to start early.

1. Linear and quadradic growth models for actual growth data[2]

[2] These data are more fully discussed and analyzed in RH&S, Chapter 7

For this problem we consider a data set on Asian children in a British community who are weighed on up to four occasions.[3] The data set `childweights.dta` is a 12% random sample, stratified by gender from the larger data set available from the center for multilevel modeling.[4]

[3] Measurements are at roughly at ages six weeks, then 8, 12, and 27 months.

[4] The full data were previously analyzed by Prosser, Rasbash, and Goldstein (1991). See RB&S pg 343 for link to full data.

We have the following variables:

- `id`: child identifier
- `weight`: weight in kilograms (the outcome)
- `age`: age in years
- `gender`: gender (1: male, 2: female).

Tasks

a. Plot the data for 16 random children. *Do not turn this part in.*
b. Fit a linear growth model to these data.
c. Try and fit a quadratic growth model to these data. You will probably run into issues, due to missing data. Explain why things are failing, and describe the distribution of the number of observations per child. Here are some useful R snippits, assuming `wts` is your data frame:

```
table( wts$id )  # tally up obs/child
table( table( wts$id ) ) # tally the tallys

# Add number of obs for each group to your data.
wts = wts %>% group_by( id ) %>%
    mutate( n = n() )
```

d. Can you subset your data so you can fit your model?[5] Try and keep as much data as possible while still fitting it.
e. Asertain which model (linear vs. quadratic) is a better fit, and provide evidence for your model choice.[6]
f. Re-create your plot from (a) with your final sample of data and augment it with the growth curves from both your linear and quadratic models (in different colors). Color the curves by gender. Turn *this plot* in.

[5] Use `filter()`.

[6] You should also reflect on whether the more complex model is worth it, in your opinion and even if it is a better fit, given the data loss required for fitting it.

2. Piecewise linear growth to model summer and school reading

This problem uses data from James Kim's Project READS. The study was an evaluation of interventions aimed at improving students' literacy skills, but we're ignoring that aspect of the dataset for this assignment. Instead, we're examining growth in student reading comprehension from their third-grade spring through their fifth-grade fall. Specifically, we're trying to answer the following research questions:

RQ 1: Do students learn more quickly over the summer or the school-year?

RQ 2: Do students' rates of growth depend on student-level poverty?

The data are in `MCC_student_data.Rds`.[7]

The `score` variable records reading comprehension. Time of measurement is recorded in `wave`, with `s3` being spring of third grade, `f4` fall of fourth, `s4` spring of fourth, and `f5` fall of fifth. Other variables are `frl`, an indicator for whether the student was eligible for a free or reduced-price lunch, `gender` coding (binary) gender, and `id` the school id. The `lep` indicator is for learning English.

We have created several time variables for you. For each wave, we have recorded the number of summer months and school months that have passed. We have also recorded time (in months) since the first measure. So The values of `time` are 0, 4, 12, and 16 months.

[7] Save and load R objects with `saveRDS()` and `dat = loadRDS( filename )`.

a) Identify the mathematical model

Use the following R code to fit an initial model:

```
m1 <- lmer(score ~ 1 + school + summer +
             (1 + school + summer|id) +
             (1 + school|sch),
           data=studs)
```

Write out in mathematical notation the model being fit.

b) Answer RQ1

Answer RQ 1, making reference to and extending any relevant work from all of the above as needed.[8]

c) Answer RQ2

Modify your model and use your augmented model to answer RQ 2. Use free and reduced price lunch as an indicator of student poverty (this is a common measure in Education). You do not need to fully write out your new mathematical model, but be sure to clearly show what you did and what coefficients you are using to to investigate this question.

[8] You may wish to shift the model from having separate slopes for summer and school to making the summer slope an increment on a baseline learning rate. See lecture on piecewise growth. If you do, write your level 1 model, but do not bother re-writing the variances; just reference your earlier work above.

d) Concept check #1 (random slopes)

Suppose that we have a sample of schools in each of which either almost *all* or almost *no* students receive free or reduced-price lunches. In a short paragraph explain

- if it would be theoretically possible to include school-level random effects for FRL status, and
- if not, why not, and if so, what problems would you expect to run into in fitting the model.