Problem Set 5: Intro to Longitudinal Data

S-043/Stat-151 (Fall 2023)

Overview

This short problem set kicks off our work on longitudinal data.

Skills to be developed

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

- Fitting and interpreting linear growth curve models.
- Plotting and visualizing longitudinal data.
- Thinking about the connection of statistics to real-world substance.

Before you start

Please do the following

1. Review the reading for Units 1-3, and the first parts of Unit 4.
2. Read through the problem before starting so you have a sense of where you are going.

1. Adolescent alcohol use[1]

[1] This problem is taken from R-H&S 7.7 (p. 380).

Singer and Willett (2003) analyzed a dataset from Curran, Stice, and Chassin (1997). As part of a larger study of substance abuse, 82 adolescents were interviewed yearly from ages 14-16 and asked about their alcohol consumption. Specifically, they were asked to report the frequency in the past 12 months of each of the following behaviors on an 8-point scale from 0 (not at all) to 7 (every day):

1. drinking wine or beer
2. drinking hard liquor
3. drinking five or more drinks in a row
4. getting drunk

Following Singer and Willett, we will use the square root of the mean of these four items as the response variable.[2]

Square Root → emphasize lower end of a scale

[2] The square root is a common transform to put more focus on change in the lower end of a scale. With square root, the distance from "1" to "2" is $\sqrt{2} - \sqrt{1} = 0.41$, and from 6 to 7 is $\sqrt{7} - \sqrt{6} = 0.20$. This transform also adjusts for data that is right skewed, likely for this kind of risk assessment data.

At age 14, the adolescents were also asked how many of their peers drank alcohol 1) occasionally and 2) regularly over the past 12 months, with each answer scored on a 6–point rating scale from 0 (none) to 5 (all). The square root of the mean of these two items was used as a covariate.

The original sample comprised 246 children who had a parent with a diagnosis of alcohol abuse or dependence (recruited through court records, wellness questionnaires from health maintenance organizations, and community telephone surveys) and 208 demographically-matched control children. Each child in the former group had at least one biological and custodial parent with a lifetime Diagnostic Interview Schedule III (DSM-III) diagnosis of alcohol abuse or dependence. The matched control children did not have this.

The dataset `alcuse.dta` has the following variables:

- id: identifier for the adolescents
- alcuse: frequency of alcohol use (square root of mean on four alcohol items, Y_{tj})
- age_14: age - 14, number of years since first interview ($time_{tj}$)
- coa: dummy variable for being a child with a parent diagnosed with alcohol abuse or dependence (w_{1j})
- peer: alcohol use among peers at age 14 (square root of mean of two items) w_{2j}

We can model these data with linear growth curves. Consider the level 1 model

$$Y_{tj} = \pi_{0j} + \pi_{1j} time_{tj} + \epsilon_{tj}$$

and the level 2 models

coa peer

$$\pi_{0j} = \gamma_{00} + \gamma_{01} w_{1j} + \gamma_{02} w_{2j} + u_{0j}$$
$$\pi_{1j} = \gamma_{10} + \gamma_{11} w_{1j} + \gamma_{12} w_{2j} + u_{1j}$$

with (u_{0j}, u_{1j}) bivariate normal as usual.

Your tasks

a. Substitute the level–2 models into the level–1 model to obtain the reduced form model.

b. Determine the correct `lmer()` call and fit the model.

c. *Succinctly* identify and interpret all the parameter estimates. No more than one sentence per parameter. Make a list, with each item being, e.g., "γ_{00} = #, which means blah, blah blah."

d. Is there evidence that children of parents diagnosed with alcohol dependency are different than the other children? Formally test and then explain any differences.

e. Make an 80% confidence interval for the effect of peer use on the growth rate of a kid. Using average growth as a benchmark, interpret both the top and bottom end of your confidence interval in terms of being a large or small effect.

f. Plot your data. You can either plot the individual growth curves of all your children, use `ggeffects` to make a plot of growth for children of parents diagnosed with alcohol dependency vs. the other children, or some other plot of your choosing that shows your model results.[3]

g. Using the empirical Bayes estimates predictions from your model, estimate the fraction of the study participants who decreased their use of alcohol over the course of the study.

h. We can estimate the *reliability* of our empirical Bayes estimates, which is a measure of how well we think we are measuring the estimate. The reliability is calculated as

$$reliability = 1 - \frac{\text{uncertainty in estimate}}{\text{uncertainty in est} + \text{variance of effects}} = 1 - \frac{SE_j^2}{SE_j^2 + \tau}$$

where τ is the variance of the random effects being evaluated (e.g., it would be τ_{00} if we wanted the reliability of the intercepts and τ_{11} if we wanted the reliability of the slopes).

What is the reliability of the estimated growth rates in your model? Does this make you trust or not trust your estimates in part (d), above?[4]

[3] See the online textbook for sample code to make these plots.

[4] Usually a reliability of 0.80 or so is considered "reliable" and less than that we have to account for the fact that our estimates are in large part noise.