Data Analysis Exercises for Chapter 23: *Applied Regression Analysis, Generalized Linear Models, and Related Methods, Third Edition* (Sage, 2016)

John Fox

Last modified: 2015-03-18

Exercise D23.1 The High School and Beyond study data set in the file `HSB.txt` includes variables in addition to those employed in the example developed in the text. At the individual level, there are each student's gender and minority status, and at the school level, the school's size and the percentage of students in the school in the academic track. It is also possible to form compositional variables for gender composition (or to differentiate single-gender and co-ed schools) and for percentage of minority students. After exploring how these variables may influence students' math-achievement test scores, formulate a richer mixed-effects model incorporating additional explanatory variables. What do you discover?

Exercise D23.2 The file `Snijders.txt` contains data on 4106 grade-8 students (who are approximately 11 years old) in 216 primary schools in the Netherlands. The data are used for several examples, somewhat different from the analysis that we will pursue below, by Snijders and Boskers in *Multilevel Analysis, 2nd Edition* (Sage, 2012).

The data set includes the following variables:

- `school`: a (non-consecutive) ID number indicating which school the student attends.
- `iq`: the student's verbal IQ score, ranging from 4 to 18.5 (i.e., *not* traditionally scaled to a population mean of 100 and standard deviation of 15).
- `test`: the student's score on an end-of-year language test, with scores ranging from 8 to 58.
- `ses`: the socioeconomic status of the student's family, with scores ranging from 10 to 50.
- `class.size`: the number of students in the student's class, ranging from 10 to 42; this variable is constant within schools, apparently reflecting the fact that all of the students in each school were in the same class.
- `meanses`: the mean SES in the student's school, calculated from the data; the original data set included the school-mean SES, but this differed from the values that I computed directly from the data, possibly it was based on all of the students in the school.
- `meaniq`: the mean IQ in the student's school, calculated (for the same reason) from the data.

There are some missing data, and I suggest that you begin by removing cases with missing data. How many students are lost when missing data are removed in this manner? Then add the following two variables to the data set:

1. school-centred SES, computed as the difference between each student's SES and the mean of his or her school; and

2. school-centred IQ.

(a) Examine scatterplots of students' test scores by centered SES and centred IQ for each of 20 randomly sampled schools. Do the relationships in the scatterplots seem reasonable linear? *Hint*: In interpreting these scatterplots, take into account the small number of students in each school, ranging from 4 to 34 in the full data set.

(b) Regress the students' test scores on centred SES and centred IQ within schools for the full data set — that is, compute a separate regression for each school. Then plot each set of coefficients (starting with the intercepts) against the schools' mean SES, mean IQ, and class size. Do the coefficients appear to vary systematically by the schools' characteristics (i.e., by the Level 2 explanatory variables centred SES, centred IQ, and class size)?

(c) Fit linear mixed-effects models to the Snijders and Boskers data, proceeding as follows:

- Begin with a one-way random-effects ANOVA of test scores by schools. What proportion of the total variation in test scores among students is between schools (i.e., what is the intra-class correlation)?
- Fit a random-coefficients regression of test scores on the students' centred SES and centred IQ. Initially include random effects for the intercept and both explanatory variables. Test whether each of these random effects is needed, and eliminate from the model those that are not (if any are not). How, if at all, are test scores related to the explanatory variables? *Note*: Depending on the software you use, you may obtain a convergence warning in fitting one or more of the null models that remove variance and covariance components; this warning should not prevent you from performing the likelihood-ratio test for the corresponding random effects.
- Introduce mean school SES, mean school IQ, and class size as Level 2 explanatory variable, but only for the Level 1 coefficients that were found to vary significantly among schools in the random-coefficients model. *Hint:* Recall that modeling variation in Level 1 coefficients by Level 2 explanatory variables implies the inclusion of cross-level interactions in the model; and don't forget that the intercepts are Level 1 coefficients that may depend on Level 2 explanatory variables. It may well help to write down the mixed-effects model first in hierarchical form and then in Laird-Ware form. Test whether the random effects that you retained in the random-coefficients model are still required now that there are Level 2 predictors in the model. *Note*: Again, you may obtain a convergence warning.
- Compute tests of the various main effects and interactions in the coefficients-as-outcomes model. Then simplify the model by removing any fixed-effects terms that are non-significant. Finally, interpret the results obtained for the simplified model. If your final model includes interactions, you may wish to construct effect displays to visualize the interactions.

Exercise D.23.3 Laird and Fitzmaurice ("Longitudinal Data Modeling," in Scott, Simonoff, and Marx, eds., *The SAGE Handbook of Multilevel Modeling*, Sage, 2013) analyze longitudinal data from the MIT Growth and Development Study on the change over time of percent body fat in 162 girls before and after menarch (age at first menstruation).[1] The data are in the file `Phillips.txt` and include the following variables:

[1] The data are originally from Phillips et al., "A longitudinal comparison of body composition by total body water and bioelectrical impedance in adolescent girls," *Journal of Nutrition*, 2003, 133: 1419–1425.

- `subject`: subject ID number, 1–162.
- `age`: age (in years) at the time of measurement; the girls are measured at different ages, and although the measurements are approximately taken annually, the ages are not generally whole numbers.
- `menarche`: age at menarch (constant within subjects).
- `age.adjusted`: age − age at menarch.
- `body.fat`: percentage body fat at the time of measurement.

Laird and Fitzmaurice fit a linear mixed-effects model to the data,

$$Y_{ij} = \beta_1 + \beta_2 t_{ij-} + \beta_3 t_{ij+} + \delta_{1i} + \delta_{2i} t_{ij-} + \delta_{3i} t_{ij+} + \varepsilon_{ij}$$

where

- Y_{ij} is the body-fat measurment for girl i on occasion j;
- t_{ij-} is adjusted age prior to menarche and 0 thereafter;
- t_{ij+} is adjusted age after menarch and 0 before;
- $\beta_1, \beta_2, \beta_3$ are fixed effects; and
- $\delta_{1i}, \delta_{2i}, \delta_{3i}$ are subject-specific random effects.

(a) Examine the data by plotting body fat versus adjusted age for all of the girls simultaneously; following Laird and Fitzmaurice, add a lowess smooth to the scatterplot. Now randomly select a subset (say, 30) of the girls and plot body fat versus adjusted age *separately* for each of the selected girls. What can you say about the apparent relationship between body fat and age before and after menarche? Is Laird and Fitzmaurice's model reasonable given your exploration of the data? Explain what each fixed-effect and random-effect coefficient in the model represents.

(b) Fit the mixed-effects model as specified by Laird and Fitzmaurice. What do you conclude? Consider the possibility of dropping each of the random effects from the model.

(c) An alternative but equivalent mixed model for these data is

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + \beta_3' t_{ij+} + \delta_{1i} + \delta_{2i} t_{ij} + \delta_{3i}' t_{ij+} + \varepsilon_{ij}$$

where t_{ij} is adjusted age, used in the model instead of adjusted age prior to menarch. The fixed-effect coefficients β_1 and β_2 have the same meaning as before, as do the random effects δ_{1i} and δ_{2i}, but β_3' represents the fixed-effect change in trend post-menarche (and δ_{3i}' the post-menarch random effect change in trend for subject i). Verify that this form of the model has the same fit to the data as the original model in part (b). Given that the study focuses on change at menarche, why might you prefer the alternative form of the model?

(d) The equivalent mixed models fit in parts (b) and (c) assume independently sampled errors ε_{ij}. There are likely too few observations per subjects (about 6.5 on average) to fit a model with autocorrelated errors *and* complex random effects (determine whether this is the case), but we can simplify the random effects to a random intercept δ_{1i} and fit a continuous first-order autoregressive process to the errors. Do this if you encounter convergence problem when you add autocorrelated errors to the original mixed-effects model, and then compare the fit of this model to that of the corresponding model estimated in part (b) or (c). Which model appears preferable? *Hint:* In comparing the model fit in part (d) to that fit in part (b) or (c), recognize that the models are not nested.