The R Statistical Computing Environment
Basics and Beyond
Mixed-Effects Models and Repeated-Measures MANOVA and ANOVA in R: Exercises

John Fox
(McMaster University)
ICPSR/Berkeley

2016

The file `Goldstein.txt` on the workshop website contains data on 728 11-year-old students in 48 inner-London primary schools. The data are analyzed by Harvey Goldstein in *Multilevel Statistical Models, Third Edition* (Arnold, 2003). The data set includes the following variables:

- `math.8`: a math-test score when the student was eight years old.
- `math.11`: a current math-test score.
- `female`: a dummy variable coded 1 for girls and 0 for boys.
- `manual`: a dummy variable coded 1 if the student's parent (presumably the main wage earner) is in a manual occupation and 0 otherwise.
- `school`: a number (ranging from 1 to 50) indicating which school the student attends. (Yes, there are only 48 schools!)

Add the following two variables to the data set:

1. the mean age-8 math score in the student's school;
2. the deviation between the student's own age-8 math score and the mean score in his or her school (i.e., compute the school-centered age-8 math score).

If you have difficulty creating these variables and adding them to the data set, you will find the necessary R code in the file `Goldstein.R` on the workshop web site.

1. Using Trellis graphics (i.e., the R `lattice` package), examine scatterplots of age-11 math score by centered age-8 math score for each school. Do these relationships seems reasonable linear? Note that some schools have very small numbers of observations and none has very many; it therefore isn't useful to plot nonparametric-regression smooths on the scatterplots. Then examine the relationship between age-11 math score and gender, and between age-11 math score and "social class." (If you have trouble formulating these graphs, the requisite R code is in `Goldstein.R`.)
2. Using the `lmList` function in the `nlme` package, regress age-11 math scores on centered age-8 scores and the dummy variables for gender and class. Look at the within-schools coefficients. Why are some missing? Then plot each set of coefficients (i.e., starting with the intercepts) against the schools' mean age-8 math scores. Do the coefficients appear to vary systematically by the school's mean age-8 scores? (Once again, you'll find R commands for these computations and graphs in `Goldstein.R`.)

3. Fit linear mixed-effects models to the Goldstein data (using `lmer` in the `lme4` package), proceeding as follows:

   - Begin with a one-way random-effects ANOVA of age-11 math scores by schools. How much of the variation in age-11 scores is between schools (i.e., what is the intra-class correlation)?
   - Fit a random-coefficients regression of age-11 math scores on the students' centered grade-8 scores, gender, and class. Initially include random effects for the intercept and all three explanatory variables. Test whether each of these random effects is needed and eliminate from the model those that are not. How, if at all, are grade-11 math scores related to the three explanatory variables? *Note*: Some of these mixed models take awhile to converge.
   - Introduce the mean school age-8 math score as a level-2 explanatory variable, but only for the level-1 coefficients that were found to vary significantly among schools in part (b). Test whether the random effects that are in the model are still required now that there is a level-2 predictor in the model.
   - Briefly summarize your findings.

Winer's venerable 1971 text *Statistical Principles in Experimental Design, Second Edition* contains data from a "modified version" of an experiment attirbuted to Meyer and Noble (1958): Six subjects high in anxiety and six low in anxiety were randomly assigned to two conditions of muscular tension (no tension and high tension), yielding three subjects in each combination of conditions of anxiety and tension. The response variable is the number of errors on a learning task made by the subjects during four trial blocks of the experiment. The data are in the file `Winer.txt`, where the variables are named `anxiety`, `tension`, `errors.1`, `errors.2`, `errors.3`, and `errors.4`.

1. Graph the mean number of errors as a function of anxiety, tension, and trial blocks. How do errors appears to be related to these factors?
2. Perform a repeated-measures analysis of variance or MANOVA of the data. What conclusions would you draw?