Multilevel analysis of a large, observational, longitudinal data set

Douglas Bates
University of Wisconsin, Madison
20 June, 2006

Abstract

Statisticians can be confident that whenever they develop techniques for the analysis of a certain type of data, such as longitudinal data, users of the techniques will encounter data and models that stress the technique to its limits. The general class of statistical models called multilevel models have been successfully applied to many types of longitudinal data and in particular have been widely applied to modeling the test scores of students over time. It is exactly this application, which has become very important in the United States because of the No Child Left Behind act that mandates annual testing of students, that has stressed the capabilities of the statistical models and software.

Some of the data sets to be modeled are very large, sometimes tens of millions of observations on millions of students, and all of them are from observational studies, not designed experiments. Such observational studies typically result in highly unbalanced designs with multiple sources of variation (students, teachers, schools, ...) that are neither nested nor fully crossed. Analysis of large data sets with partially crossed sources of variation is a formidable task. I will discuss techniques for specifying and fitting multilevel models to this type of data, including methods for model building and hypothesis testing.