

Software Design for Longitudinal Data Analysis

Douglas M. Bates*

*University of Wisconsin, Madison
United States*

José C. Pinheiro

*Bell Laboratories
United States*

ABSTRACT Software for exploring and modelling longitudinal data can be made much easier to use by incorporating an object-oriented design. Current versions of S-Plus provide some object-oriented capability but experimental versions of S emphasize an even stronger commitment to object orientation. These new capabilities, combined with the development of Trellis graphics by Cleveland and Becker, caused us to reexamine the basic design of our mixed-effects modelling functions - lme and nlme. We chose to implement a groupedData class with associated constructors, modeling, and display methods. This provides powerful visualization capabilities almost automatically. In addition it dramatically simplifies the interface to the modelling code.

We describe and illustrate the use of this approach and its impact on the modeling software with special emphasis on nonlinear mixed-effects models.

Key words and phrases: Mixed-effects models, trellis graphics, nonlinear regression, self-starting models.

1 Introduction

Just as the statistical methodology for longitudinal data can be considerably more complicated than for cross-sectional data, the design of software for the exploration and modelling of longitudinal data is more complicated. As always with statistical software tradeoffs must be made between convenience for the user, sophistication of the analysis methods, and efficiency of the code. With longitudinal data additional issues of delineating experimental units or longitudinal characteristics must be considered. When so-

*This research was supported by the National Science Foundation through grant DMS-930901.

phisticated graphical techniques for data exploration or model diagnostics are also included the design problems become even greater.

For several years we have been developing classes and methods in the S language (Chambers and Hastie, 1992) for modelling longitudinal data with linear or nonlinear mixed-effects models. The `nlme` library is now included with versions 3.4 and higher of S-Plus from the Data Analysis Products Division of MathSoft, Inc. Its use is described in MathSoft (1996, chapter 2) and Venables and Ripley (1996, sections 6 and 9).

Although the library is useful in its current state, we have decided to change the design of the code and put a much greater emphasis on object-orientation. The primary reasons for the change are to enhance the use of trellis graphics displays; to provide a cleaner, more intuitive user interface; and to facilitate migration of the code to version 4 of the S language (Chambers, 1993). The redesign has affected several areas of the code. Here we will focus on two important areas: the use of `groupedData` objects to encapsulate the data and key aspects of the structure of the data, and the use of self-starting nonlinear regression models to derive starting estimates for nonlinear regression model parameters.

In §2 we describe trellis graphics displays for longitudinal data and the motivation they provide for `groupedData` objects. These objects are described in §3. The `selfStart` class of nonlinear regression models are described in §4 and our conclusions are given in §5.

2 Trellis graphics and longitudinal data

One of the most exciting recent developments in graphical display of data is the `trellis` approach (Cleveland, 1994; Becker, Cleveland and Shyu, 1996) to multi-panel displays of conditional plots. The approach is particularly valuable for longitudinal data that typically represent measurements of a response over time for different experimental units. We want to examine the behaviour within these units but also allow ourselves to compare behaviour between units. Trellis displays are well suited to this.

The approach is best shown with an example. Kung (1986) presents data, shown in Figure 1, on the growth of Loblolly pine trees. A common method of representing such data would be a plot with fourteen curves on it, one for each tree. Usually this type of plot becomes cluttered and difficult to read. In a trellis display like Figure 1 the measurements from different trees are plotted in separate panels but with consistent axes to facilitate comparison between panels. The panels are labelled with a strip giving an identifier, the seed source in this case, for the tree.

There are other, more subtle aspects to this plot. The order in which the panels are plotted is determined by a characteristic of the data. In this case, the panels have been ordered according to the maximum height measured