Software Design for Longitudinal Data Analysis

Douglas M. Bates
U. of Wisconsin

José C. Pinheiro
Bell Labs

Outline

- Why revise current `lme/nlme` code?
- Examples of trellis graphics for longitudinal data
- A class for longitudinal data
- Special characteristics of the class
- Self-starting nonlinear regression models
- Some simple modelling

Linear and nonlinear mixed effects in S

- Initial definition and algorithms with Mary Lindstrom (1988, 1990)
- `lme/nlme` **S** code (and further theory) by Pinheiro and Bates
- Version 2.1 of the `nlme` library is most recent
- This version is shipped with **S-PLUS** 3.4 (Sept. ’96)
- It is also available via anonymous ftp or from `Statlib`
- Shipped as source code (GPL’d)

Capabilities of nlme 2.1 library

- Linear mixed effects using **S** formula language
- Nonlinear mixed effects using **S** nonlinear formulas
- Only one level of random effects (no nesting)
- Maximum likelihood or REML (Gaussian)
- General or patterned covariance of random effects
- optional variance function for intra-group noise
- optional correlation pattern in intra-group noise

Why change?

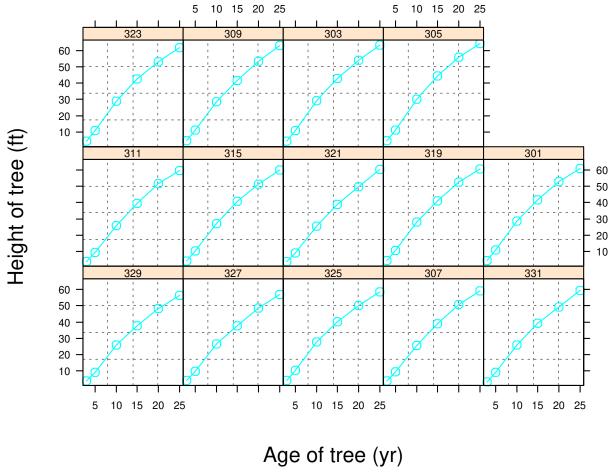
New ideas in software development

- Primary influences are *trellis* and *S version 4*
- Use the power of trellis graphics
- Borrow the ideas of formula-data representation and conditional expressions.
- Object-orientation (classes and methods) more important in S4.
- Improve maintainability.
- Add new capabilities

Today we show the relationship to trellis and two important classes for our version 3.0 code

Trellis graphics for longitudinal data

Growth of Loblolly pine trees by seed source as quoted in Kung (1986).




```
> Loblolly <- groupedData(height ~ age | Seed,
+    data = read.table("Loblolly", header = T),
+    labels = list(x = "Age of tree", y = "Height of tree"),
+    units = list(x = "(yr)", y = "(ft)"))
> plot(Loblolly)
```


Characteristics of these data

- Balanced design
- Very similar patterns
- Nonlinearity in growth pattern. Might use the *asymptotic regression model*

$$\theta_1 + (\theta_2 - \theta_1)\exp(-\theta_3 t), \quad \theta_3 > 0$$

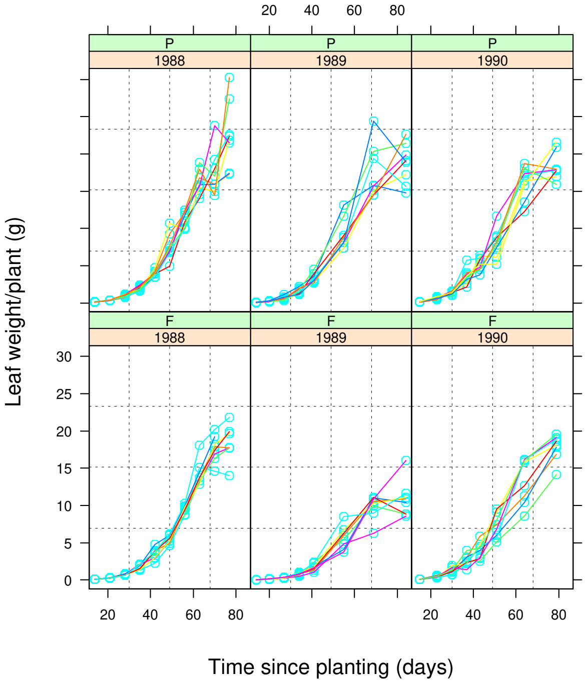
for the height versus time for each seed source.

Characteristics of the plot

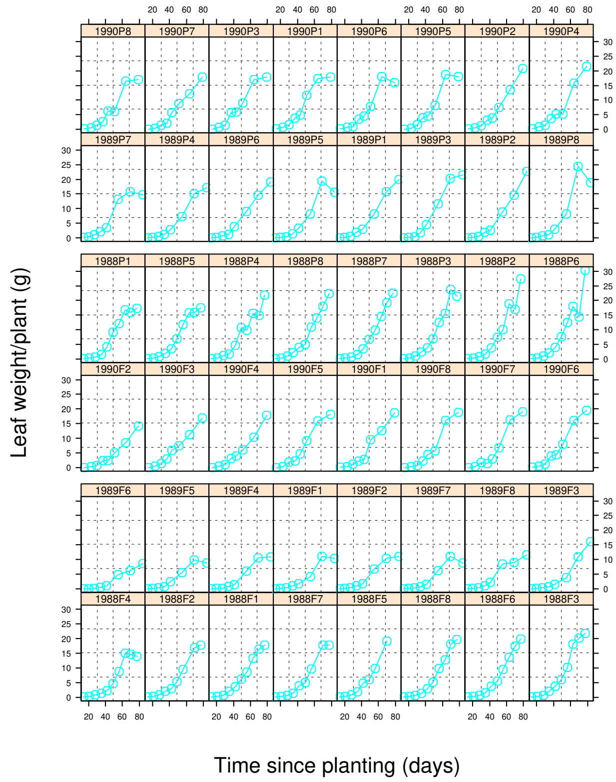
- Array of plots like a latticework or trellis. Individual plots are called panels.
- Identical axes on each panel to facilitate comparison
- Panels are ordered — by maximum response in this case
- Aspect ratio tuned to enhance comparison of slopes (*banking to 45^o*)

Soybean growth

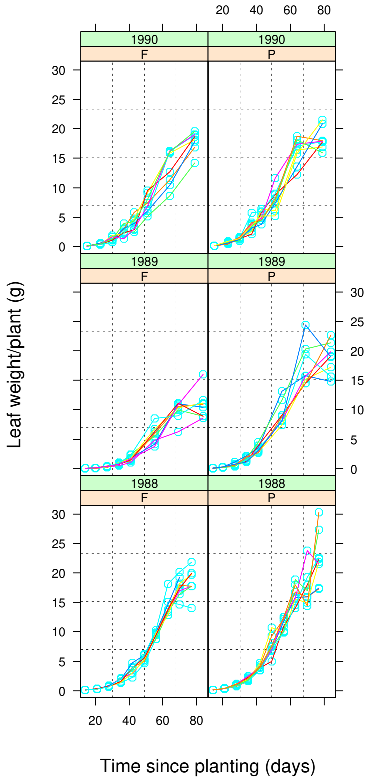
Eight plots each for 2 varieties of soybean in 3 years (Davidian and Giltinan 1995)



By plot



Alternate orientation



Code for Soybean plots

```
> Soybean <- groupedData(
+  formula = weight ~ time | Plot,
+  data = read.table("Soybean", h= T),
+  outer = ~ Variety * Year,
+  labels = list(
+           x = "Time since Planting",
+           y = "Leaf weight/plant"),
+  units = list(x = "(days)",
+               y = "(g)"))
> plot(Soybean, outer = ~Year*Variety)
> # two trellis parameters enhance the
> # layout for the second plot
> plot(Soybean, layout = c(8, 6),
+   between = list(y = c(0, 0.5)))
> plot(Soybean, outer = T)
```

Characteristics of Soybean Data

- Sigmoidal shape wrt time. D&G suggest a *logistic function* for the intra-group model.

$$\theta_1 / \left[1 + \exp\left(\frac{\theta_2 - t}{\theta_3} \right) \right]$$

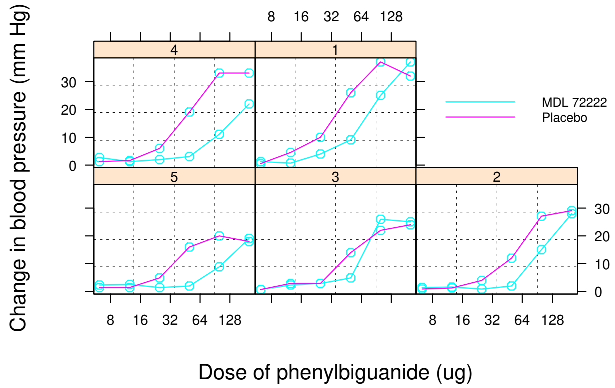
- Two *outer* factors - `Year` and `Variety`.
- Probably an interaction in these factors when modelling the asymptote (θ_1) for the weight.
- Greater level of “noise” when response is greater. May want to model intra-group variance.
- Design could lead to serial correlation for the intra-group noise.

Use of `outer` formulas

- The *grouping factor* (`Plot`) represents sampling from the population to which random effects will apply.
- *Outer* factors (`Variety` and `Year`) apply to all observations within a group.
- Re-ordering of grouping factor levels in default plot preserves outer factors.
- When `outer` is set in a plot, the outer formula is used for dividing panels, the grouping factor determines lines.

Phenylbiguanide and rabbits’ blood pressure

Change in blood pressure for 5 rabbits with ascending doses of PBG after treatment by placebo and by a 5-HT$_3$ antagonist (Ludbrook 1994, Venables and Ripley 1996)




```
> PBG <- groupedData(deltaBP ~ dose | Rabbit,
+  data = read.table("PBG", header = T),
+  inner = ~ Treatment,
+  labels = list(x = "Dose of phenylbiguanide",
+                y = "Change in blood pressure"),
+  units = list(x = "(ug)", y = "(mm Hg)"))
> plot(PBG, scales = list(x = list(log = 10)))
```


Use of `inner` factors

- Levels of an `inner` factor can change within groups
- On plots an `inner` factor does not affect the panel — it determines which observations are joined in each panel.

`groupedData` objects in general

- Extension of the idea of a data frame
- Incorporates a (continuous) response, a primary covariate and a grouping factor in a display formula

```
response ~ primary | grouping
```

- By default the grouping factor is converted to an ordered factor according to a summary function applied to the response
- Outer or inner factors can be specified.
- Labels and units can be specified.

From data display to model building

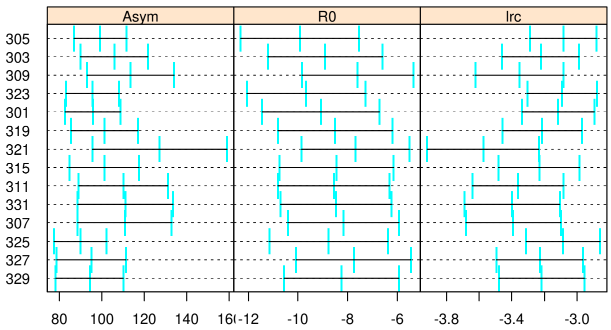
- Assuming an iterative model building process
- Generally would begin by fitting the intra-group model to each group of observations
- From these fits determine how the parameters are varying between groups, how they depend on other covariates, etc.
- Start fitting more complicated models as indicated
- If intra-group model is nonlinear, need starting estimates

`selfStart` objects

- A class of nonlinear regression models with an auxillary “initial” function
- Contains the model function itself and possibly code for derivatives
- Before the iterations start, the “initial” function is called to create suitable starting estimates
- Described in Venables and Ripley (1996). Interface to “initial” function will change in nlme 3.0.

Separate fits to Loblolly pine trees

```
> Loblolly.lis <- nlsList(asymptotic, Loblolly)
> coef(Loblolly.lis)
         Asym        R0       lrc
329  94.12877 -8.250685 -3.217588
327  94.94058 -7.757495 -3.229325
325  89.88361 -8.759223 -3.086195
307 110.69818 -8.169516 -3.390332
331 111.00287 -8.462608 -3.397574
311 109.98436 -8.558667 -3.362500
315 101.05602 -8.443651 -3.232822
321 127.13533 -7.679281 -3.575346
319 101.08717 -8.502381 -3.214016
301  95.66688 -9.078241 -3.116381
323  95.55627 -9.665035 -3.092266
309 113.51390 -7.595622 -3.352815
303 105.71443 -8.906502 -3.222905
305  99.17167 -9.916694 -3.084837
> plot(intervals(Loblolly.lis))
```



From individual fits to mixed effects

```
> Loblolly.nlme <- nlme(Loblolly.lis)
> Loblolly.nlme1 <- nlme(Loblolly.lis,
+    random = list(Asym ~ ., lrc ~ .))
> Loblolly.nlme2 <- nlme(Loblolly.lis,
+    random = list(Asym ~ .))
> anova(Loblolly.nlme, Loblolly.nlme1,
+    Loblolly.nlme2)
               Model df    AIC    BIC  loglik
 Loblolly.nlme     1 10 242.22 266.53 -111.11
Loblolly.nlme1     2  7 238.25 255.26 -112.12
Loblolly.nlme2     3  5 239.49 251.64 -114.74

                  Test Lik.Ratio P-value
 Loblolly.nlme
Loblolly.nlme1 1 vs. 2    2.0235 0.56754
Loblolly.nlme2 2 vs. 3    5.2389 0.07284
> summary(Loblolly.nlme2)
...
Random Effects Variance-Covariance
 Standard Deviation:
[1] 3.650654

 Within-Group Variance: 0.5167627

Fixed Effects Estimate(s):
           Value  Std.Error   z ratio
Asym  101.447153 2.41718972  41.96905
  R0   -8.627632 0.31222361 -27.63286
 lrc   -3.233713 0.03365146 -96.09429

 Correlation of Fixed Effects Estimates
          Asym        R0
 R0  0.7039401
lrc -0.9077674 -0.8271033
```


General conclusions

- Trellis graphics provides compact, informative graphical presentation of small sample longitudinal data
- A class for grouped data encapsulates the data and key aspects of the structure of the data
- The groupedData class facilitates both graphical display and initial model fits
- Creating a self-starting nonlinear regression model takes a lot of the drudgery out of the model building.
- An object-oriented approach to longitudinal software presents challenging design issues but enhances usability and maintainability.