Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2006 : Methodological Issues in Measuring Population Health

2006

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2006
Methodological Issues in Measuring Population Health

Flexible Models for Analyzing Longitudinal Data in Population Health

Joel A. Dubin[1]

Abstract

The study of longitudinal data is vital in terms of accurately observing changes in responses of interest for individuals, communities, and larger populations over time. Linear mixed effects models (for continuous responses observed over time) and generalized linear mixed effects models and generalized estimating equations (for more general responses such as binary or count data observed over time) are the most popular techniques used for analyzing longitudinal data from health studies, though, as with all modeling techniques, these approaches have limitations, partly due to their underlying assumptions. In this review paper, we will discuss some advances, including curve-based techniques, which make modeling longitudinal data more flexible. Three examples will be presented from the health literature utilizing these more flexible procedures, with the goal of demonstrating that some otherwise difficult questions can be reasonably answered when analyzing complex longitudinal data in population health studies.

KEY WORDS: Longitudinal; Curves; Non-linear models; Penalized splines; Smoothing.

1. Introduction

The modeling of longitudinal data is described in various textbooks (e.g., Verbeke and Molenberghs, 2000; Diggle et al. 2002, Fitzmaurice et al., 2004). Most (but not all) of the discussion in these texts and most of the longitudinal analyses seen in application journals (such as in biomedicine or population health) focus upon linear models and generalized linear models for longitudinal data. Correlation between the repeated measures over time within a subject is appropriately handled with these models. For linear models, the typical approach used is the linear mixed effects model (Laird and Ware, 1982), and for generalized linear models, the majority of researchers use either marginal models with generalized estimating equations (Liang and Zeger, 1986) or subject-specific approach using generalized linear mixed effect models (e.g., Breslow and Clayton, 1993; McCulloch and Searle, 2000). However, many datasets require more flexible approaches to modeling either by a different parameterization of the above popular models or instead new models altogether. We will provide three examples of longitudinal datasets that required these types of flexible approaches.

In this review paper, all of these examples have been published in various forms (and journals) in the biostatistical and biomedical literature. In Section 2 below, we will describe the examples, one at a time, with the problem of interest defined and subsequent model specified. Each approach has the characteristic that the trajectory of interest over time is not linear at a constant rate during the entire course of follow-up, with trajectories from some of the examples being quite non-linear. In Section 3, we provide some summary remarks.

2. Examples

2.1 Example 1: pharmacokinetics of Theophylline

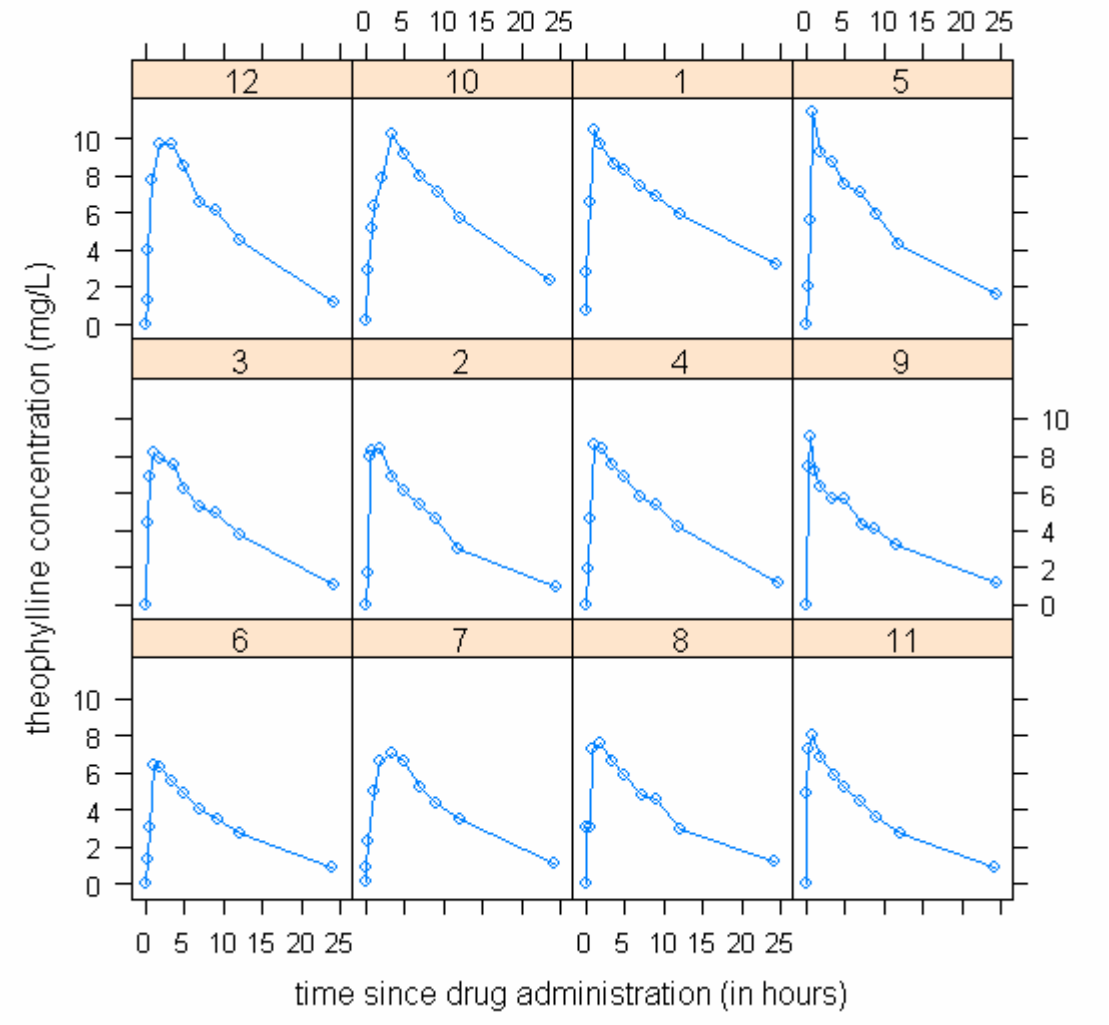
Pharmacokinetics, which studies how our bodies process the intake of a drug, has long focused on models that are non-linear. Parametric non-linear models have been the norm and the parameters in the model represent functions such as absorption of the drug, as well as elimination and clearance. Some researchers have focused on longitudinal

[1]Joel A. Dubin, University of Waterloo, 200 University Ave W., Waterloo, ON, Canada, N2L 3G1, jdubin@uwaterloo.ca

modeling of these processes, for example, from random assignment of different dosing levels to a group of subjects in a clinical trial, including the use of non-linear mixed effects models (e.g., Pinheiro and Bates, 2000).

Pinheiro and Bates (2000) present one such example, a designed pharmacokinetic study of 12 subjects, with between-subject differences on doses received. The longitudinal data for the 12 subjects are presented in Figure 1. The non-linear trajectories for each subject are apparent, as well as between-subject differences in the absorption rate and peak, as well as the elimination and clearance. Random effects may be needed in any modeling of this data, in case dosage amount by itself cannot explain these between-subject differences.

Figure 1: pharmacokinetics of Theophylline concentration



The following non-linear longitudinal (mixed effects) parametric model, a first-order compartmental model, is described in Pinheiro and Bates (2000) for the Theophylline data to reflect the concentration Y at time t after initial dose d:

$$Y_{ij} = \frac{d_i E_i A_i}{C_i (A_i - E_i)} [\exp(-E_i t_{ij}) - \exp(-A_i t_{ij})] + \varepsilon_{ij} \qquad (1)_{ij}$$

In (1),

Y_{ij} is the Theophylline concentration for subject i at time t_{ij}

d_i is the initial dose for subject i

E_i is the elimination rate constant for subject i, where $E_i = \beta_1 + b_{1i}$, with β_1 a population-level fixed elimination effect and b_{1i} a subject-specific random elimination effect, with mean 0 and variance σ^2_{b1}

A_i is the absorption rate constant for subject i, where $A_i = \beta_2 + b_{2i}$, with β_2 a population-level fixed absorption effect and b_{2i} a subject-specific random absorption effect, with mean 0 and variance σ^2_{b2}

C_i is the clearance for subject i, where $C_i = \beta_3 + b_{3i}$, with β_3 a population-level fixed clearance effect and b_{1i} a subject-specific random clearance effect, with mean 0 and variance σ^2_{b3} .

ε_{ij} is the within-subject error term for subject i at time j, assumed conditionally independent $N(0, \sigma^2)$

Though (1) appears as a very complex model, each of the parameters has a biologically interpretable meaning, specifically absorption, clearance and elimination. It is an extension of a general non-linear parametric model: $Y_{ij} = f(X_{ij}, \beta, b_i) + \varepsilon_{ij}$, where the β terms represent fixed population-level effects. In addition, enough flexibility has been built into (1) to allow for between-subject heterogeneity that cannot be explained by observed covariates such as dosage level. This is a complex and flexible model, yet with interpretative value. As far as the fitting of such non-linear parametric longitudinal models, there is an excellent function available in both R and S-Plus, called *nlme*, written by Pinheiro and Bates (2000), which comes from the nlme package/library. In SAS, *Proc NLMIXED* can be used.

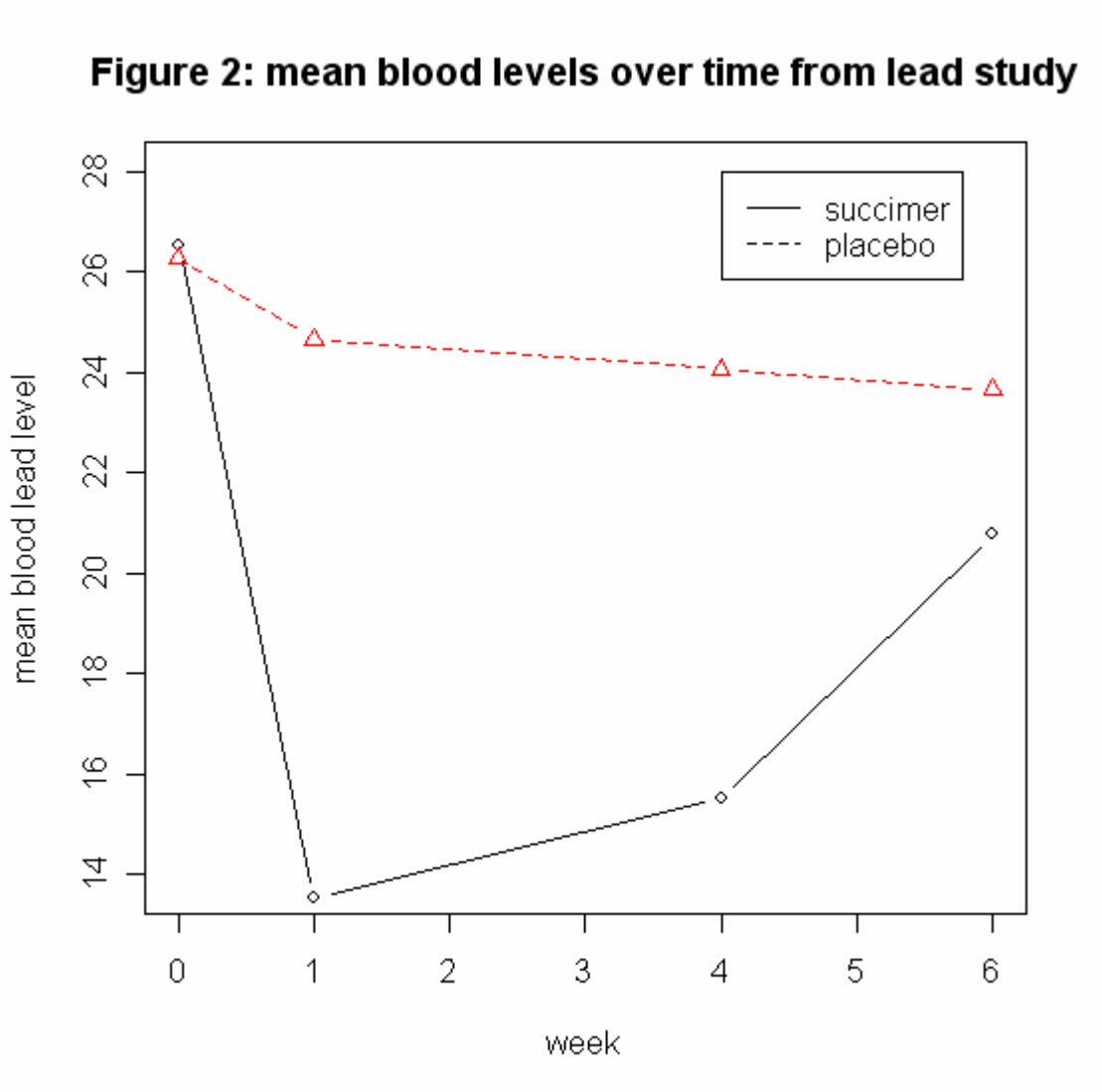
2.2 Example 2: children lead-exposure clinical trial

This second example comes from a placebo-controlled randomized trial of a chelating agent, succimer, in children with elevated blood level between 20 to 44 μg/dL. Enrolled were 100 children from poor inner city housing; ages between 12 and 33 months. This example is well-described in Fitzmaurice et al. (2004). The goal was to see if succimer would do better than the placebo in lowering blood lead levels for children with elevated lead levels.

Figure 2 presents four consecutive mean longitudinal measurements (week 0, week 1, week 4, and week 6), one mean trajectory for each treatment group. A couple of key features can be detected from this plot. First, there is a noticeable dive in blood levels after the first week among those in the succimer group, as compared to a small decrease (possibly a placebo effect and/or simple regression to the mean) in the placebo group. Secondly, after the initial dive in the succimer group, the blood levels rise again, though maintaining a mean lower than the placebo group throughout the six-week follow-up period. This rise is due to the body of the children in the succimer group re-balancing the blood lead level after the initial loss, with lead being released into the bloodstream from bone and muscles.

This is an example that requires a much simpler parametric model than that used in Example 1. Here, the initial focus is simply on a single change in an otherwise straight mean trajectory of the response over time. Hence, a linear mixed effects model can be used, with the implementation of a single changepoint. This is the simplest form

of what otherwise is referred to as a longitudinal *piecewise linear spline* model, with a subject-specific random effect and a single knot at the changepoint.



Fitzmaurice et al. (2004) suggested the following longitudinal piecewise linear spline model:

$$Y_{ij} = (\beta_0 + b_{0i}) + \beta_1(week_{ij}) + \beta_2(week_{ij} - 1)_+ + \beta_3(trt_i * week_{ij}) + \beta_4(trt_i * (week_{ij} - 1)_+) + \varepsilon_{ij} \quad (2)$$

In (2),

Y_{ij} is the longitudinal lead level response for subject i at time $week_{ij}$, where subject i = 1 to 100, and measurement time j = 1 to 4

β_0 is population-level intercept

b_{0i} is the subject-specific random intercept for subject i, assumed to be distributed independent $N(0, \sigma^2_{b0})$; this random effect induces the correlation between the repeated measures over time for each child

β_1 represents population-level slope (initial slope)

$(week_{ij} - 1)_+ = (week_{ij} - 1)$ if $(week_{ij} - 1) > 0$, and 0 otherwise; here, the changepoint (knot) is specified at week 1

β_2 represents population-level adjustment to slope from week 1 forward

β_3 represents interaction between treatment group (1 = succimer, 0 = placebo) and population-level initial slope

β_4 represents interaction between treatment group (1 = succimer, 0 = placebo) and population-level adjustment to slope from week 1 forward

ε_{ij} is the within-subject error term for subject i at time j, assumed to be distributed $N(0, \sigma^2)$, and conditionally independent.

Note the following **expected responses** for subject i at time j, which can be inferred from (2):

$E(Y_{ij}) = \beta_0 + \beta_1(week_{ij}) + \beta_2(week_{ij} - 1)_+$ [for placebo group]

$E(Y_{ij}) = \beta_0 + (\beta_1 + \beta_3)(week_{ij}) + (\beta_2 + \beta_4)(week_{ij} - 1)_+$ [for succimer group]

So, for the placebo group, this translates into a slope of β_1 prior to the changepoint (week 1) and $(\beta_1 + \beta_2)$ after the changepoint.

For the succimer group, this translates into a slope of $(\beta_1 + \beta_3)$ prior to the changepoint (week 1) and $(\beta_1 + \beta_3 + \beta_2 + \beta_4)$ after the changepoint.

In summary, equation (2) represents a relatively simple longitudinal model but one that has been adapted to allow for a change in trajectory over time, including this trajectory change to be a function of treatment group. It can be argued that a second changepoint, at week 4, would provide a better fit to the blood level data. It is not hard to imagine extending models such as (2) to include several changepoints (knots); that is, models such as (2) can easily be extended to accommodate greater non-linearity in the trajectory over time. One such complex model will be described on a different dataset, in Example 3 below. On a programming note, these types of fixed changepoint models, as described in (2) above, can easily be fitted in a linear mixed effects modeling program such as *lme* (in R or S-Plus) or *lmer* (in R) or in *Proc Mixed* in SAS.

2.3 Example 3: respiratory health study

This example, presented only briefly here, comes from an investigation of the association between respiratory health and respirable particulate pollution in a sample of 41 schoolchildren, over a period of 109 consecutive days in 1991. Daily measures of peak expiratory flow (PEF) were collected, as well as amount of particulate matter (PM) and several weather variables such as lowest temperature of the day (LT). Unfortunately, a Figure from this example was unavailable to be included in this paper.

The following initial linear mixed effects model is an extension of the linear model used by the initial investigators (Pope et al., 1991):

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1(PM)_j + \beta_2(LT)_j + \beta_3(Time)_j + \varepsilon_{ij} \quad (3)$$

In (3), the response Y_{ij} represents PEF measurement for child i (i = 1 to 41) at time j (j = 1 to 109), and b_{0i} is a single between-subject random effect representing unexplained variability (heterogeneity) between children on baseline PEF levels. Also, notice that the *i* subscript is suppressed from PM, LT, and Time, as these are not specific to an individual in this study.

Coull et al. (2001) realized the above was an inadequate model, as more flexible relationships were required between PEF and LT, as well as between PEF and Time. In addition, unexplained variability needed to be accounted for regarding the relationship between PEF and PM, and accounting for serial correlation was also necessary. Hence, Coull et al. (2001) suggested the following *additive mixed model* for this data:

$$Y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})\,(PM)_j + f(LT)_j + g(Time)_j + \varepsilon_{ij} \qquad (4)$$

The model (4) reflects the following changes from (3):

b_{1i} has been added to the model to account for unexplained between-child variability on the relationship between PEF and PM

ε_{ij} now accounts for serial correlation, such that $\varepsilon_{ij} = \rho\varepsilon_{i,j-1} + \alpha_{ij}$

f and g represent non-linear relationships between PEF and LT, as well as PEF and Time, respectively. Coull et al. (2001) reflected these non-linear relationships using smoothed functions, via *penalized splines* (e.g., Eilers and Marx, 1996), a more complex implementation of splines than the ones used in the children lead example model in (2).

As complex as the above model (4) may appear, additive mixed models, including models such as that presented in (4) for the respiratory data, can use standard linear mixed effects model software, such as *lme* in R and S-Plus, and *Proc MIXED* in SAS. For more complex additive models, specialized software is required. The SemiPar package in R may be one option, and complex additive mixed models are described in Ruppert et al. (2003). For generalized additive mixed models (e.g., Lin and Zhang, 1999; Wood, 2006), one can use the mgcv package in R.

3. Discussion

Several informative examples were presented in this review paper, with the focus on longitudinal data where complex modeling was required. These were only a few approaches among many proposed in the statistical and biostatistical literature, though several of these flexible approaches have not yet become mainstream in the population health and biomedical literature. Some of these other approaches can fall into classes of *varying coefficient models*, where the β parameters themselves are a function of time (e.g., Chiang et al., 2001; Lin and Ying, 2001; Fan and Li, 2004), fully non-parametric curve-based models (e.g., Rice and Silverman, 1991; Dubin and Müller, 2005), and functional data analytic models (e.g., Ramsay and Silverman, 2005; Yao et al., 2005), the latter two approaches using curves (representing the trajectory of responses over time) as the basis for analysis. Fortunately, there has recently been a concerted effort by many of the researchers developing these new methods to create software such that these flexible longitudinal methods can be utilized more often in the applied health literature. The studies contained in this paper are examples of that effort.

References

Breslow, N.E., and Clayton, D.G. (1993), "Approximate inference in generalized linear mixed models", *Journal of the American Statistical Association*, 88, 125-134.

Chiang, C.T., Rice, J.A., Wu, C.O. (2001), "Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables", *Journal of the American Statistical Association*, 96, 605-619.

Coull, B.A., Schwartz, J., and Wand, M.P. (2001), "Respiratory health and air pollution: Additive mixed model analyses", *Biostatistics*, 2: 337-349.

Diggle, P.J., Heagerty, P., Liang, K.-Y., and Zeger, S.L. (2002), *Analysis of Longitudinal Data*, 2nd ed., Oxford: Oxford.

Dubin, J.A., and Müller, H.G. (2005), "Dynamical correlation of multivariate longitudinal data", *Journal of the American Statistical Association*, 100, 872-881.

Eilers, P.H.C. and Marx, B.D., (1996), "Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, 11: 89-121.

Fan, J.Q., Li, R. (2004), "New estimation and model selection procedures for semiparametric modeling in longitudinal data", *Journal of the American Statistical Association*, 99, 710-723.

Fitzmaurice, G.M., Laird, N.M., Ware, J.H. (2004), *Applied Longitudinal Analysis*, Hoboken: Wiley.

Laird, N.M., and Ware, J.H. (1982), "Random-effects models for longitudinal data", *Biometrics*, 38, 963-974.

Liang, K.-Y., and Zeger, S.L. (1986), "Longitudinal data analysis using generalized linear models", *Biometrika*, 73, 13-22.

Lin, D.Y., Ying, Z. (2001), "Semiparametric and nonparametric regression analysis of longitudinal data", *Journal of the American Statistical Association*, 96, 103-113.

Lin, X., and Zhang, D. (1999), "Inference in generalized additive mixed models by using smoothing splines", *Journal of the Royal Statistical Society, Series B*, 61: 381-400.

McCulloch, C.E. and Searle, S.R. (2000), *Generalized, Linear, and Mixed Models*, New York: Wiley.

Pinheiro, J.C., and Bates, D.M. (2000), *Mixed-Effects Models in S and S-PLUS*, New York: Springer-Verlag.

Pope, C.A., Dockery, D.W., Spengler, J.D., and Raizenne, M.E. (1991), "Respiratory health and PM_{10} pollution: A daily time series analysis", *American Review of Respiratory Disease*, 144, 668-674.

Ramsay, J.O. and Silverman, B.W. (2005), *Functional Data Analysis, 2nd ed.*, New York: Springer-Verlag.

Rice, J.A., and Silverman, B.W. (1991), "Estimating the mean and covariance structure nonparametrically when the data are curves", *Journal of the Royal Statistical Society, Series B*, 53: 233-243.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003), *Semiparametric Regression*, Cambridge: Cambridge.

Verbeke, G. and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer-Verlag.

Wood, S. (2006), *Generalized Additive Models: An Introduction with R*, London: Chapman & Hall.

Yao, F., Müller, H.G., Wang, J.-L. (2005), "Functional data analysis for sparse longitudinal data", *Journal of the American Statistical Association*, 100, 577-590.