BU-1371-M

SEMI-PARAMETRIC METHODS FOR LONGITUDINAL DATA ANALYSIS

Naomi Altman Cornell University
423 Warren Hall, Ithaca, NY14853

Keywords and phrases: Mixed model; random effects; nonlinear regression; local polynomial regression; self-modeling regression; SEMOR; shape-invariant regression; functional data analysis; semi-parametric regression; sample of curves;

Abstract

Self-modeling regression is a powerful semi-parametric tool for analysis of longitudinal data with time-invariant covariates. This paper explore the use of self-modelling regression for fitting data from a designed experiment in which the response is a curve. Recent advances in mixed nonlinear models and nonparametric regression with time series errors has made the use of mixed model self-modeling regression a feasible extension of parametric mixed model methods for studies in which the response is a curve.

1 Introduction

In areas such as study of tumor growth, sensory response and material wear, the response of each experimental unit is a time curve. The data may be observational, or may result from a controlled experiment. This paper focuses on the analysis of such data using a semi-parametric mixed model based on the self-modeling regression (SEMOR) approach of Lawton et al (1972). The advantages of this approach are:

1. efficient and interpretable data summary via the parametric part of the model
2. modeling of time-invariant treatment and covariate effects via the parametric part of the model
3. separate modeling of effects of treatments and covariates on the time scaling and magnitude of response
4. flexible determination of the shape of the time curve via the nonparametric part of the model
5. insensitivity of the estimates to within subject error correlation

The paper is organized around the data displayed in Figure 1. These data (from Crowder and Hand, 1990, p. 13-18) are serum glucose measurements on 6 healthy volunteers taken following a high carbohydrate meal. A SEMOR model is developed for these data. Computational and modeling issues are discussed, and extensions are suggested in the context of nonlinear mixed modeling.

Section 2 discusses the data. Section 3 is a brief introduction to the SEMOR model. Section 4 discusses two computational algorithms and regression diagnostics for fitting the SEMOR model. Section 5 discusses results for the glucose data. Section 6 suggests extensions to the model. Section 7 contains concluding remarks.

2 The Glucose Data

Figure 1 displays slightly smoothed serum glucose measurements taken on 6 healthy volunteers on 6 different occasions, following ingestion of a high carbohydrate meal. (Crowder and Hand, 1990, p. 13-18). Subjects were measured 15 minutes before and immediately after the meal, then half-hourly to hourly for up to 7 hours following. Meals were given at different times of day, with several days between each meal. Another view of the data is given in Figure 2, which displays all the data, slightly smoothed, for person 2. Some features of the displays, particularly linear and quadratic trends at the boundaries of the measurement interval, are due to the smoothing.

Serum glucose rises in the bloodstream as the body metabolizes carbohydrates from the meal. In response, the body produces insulin which eventually brings the glucose level back to baseline. This produces the basic shape of the glucose response curve. However, the response is not simple – it depends on activity levels as well as on hormones circulating in the body. In particular, human growth

*Supported by Hatch Grant 151410 NYF and NSF Grant DMS-95-25350. The author wishes to thank N. Peckinpaugh for useful information regarding glucose metabolism, and M. Lindstrom for a discussion of the nonlinear mixed model.

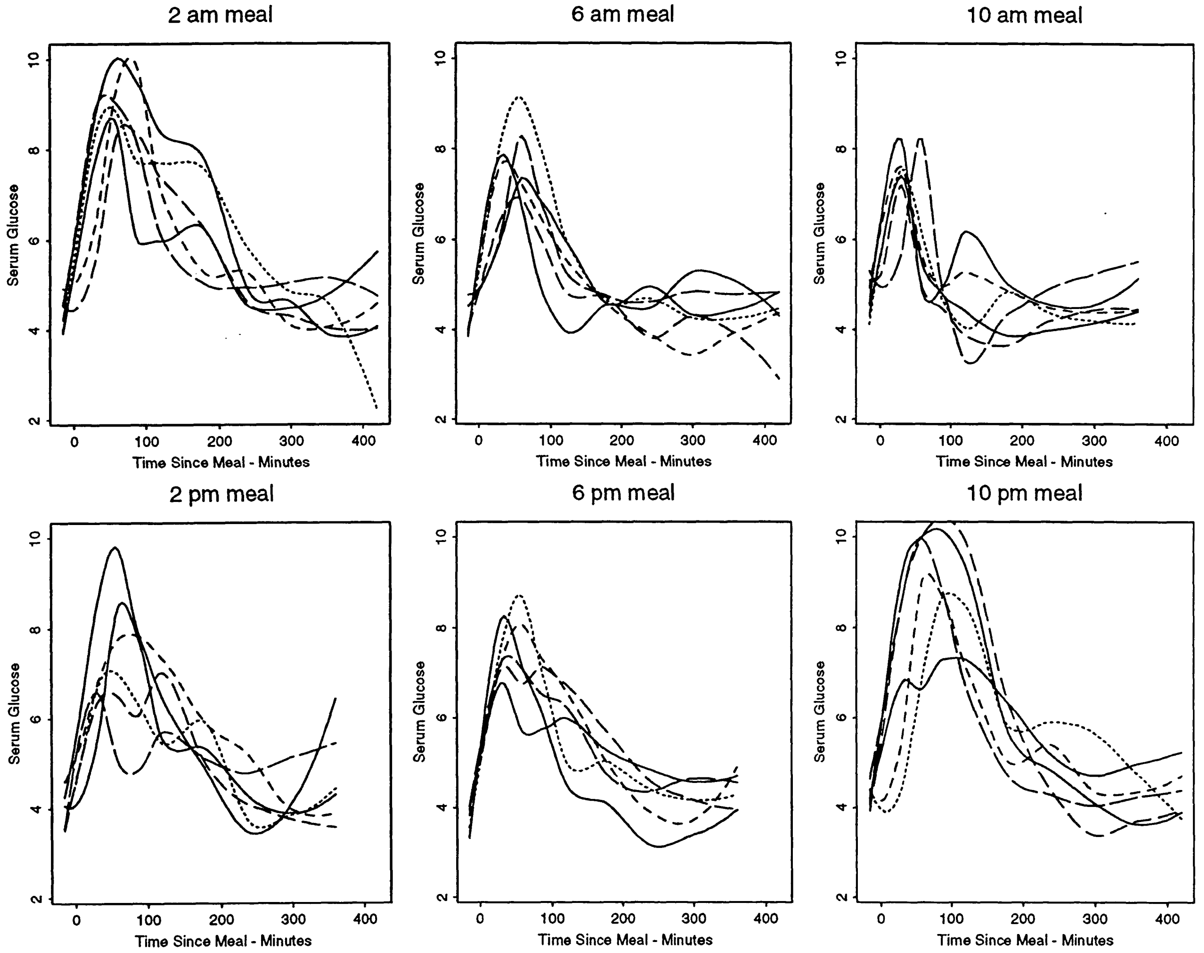

Figure 1: *Slightly smoothed serum glucose measurements taken on 6 volunteers following a high carbohydrate meal. Meals were taken at various times of day, several days apart. Curves for each volunteer have the same line type in each plot.*

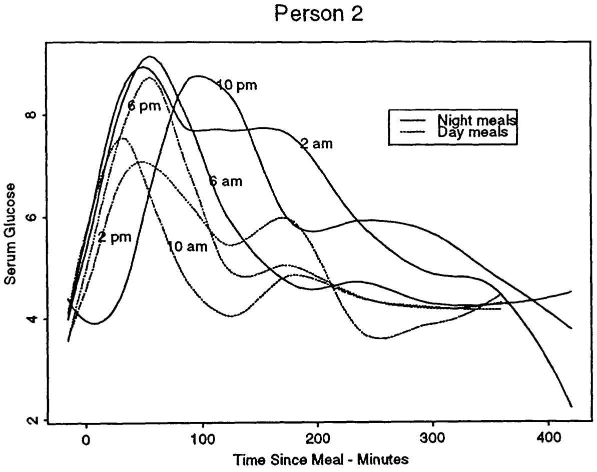

Figure 2: Slightly smoothed serum glucose measurement taken on a volunteer following a high carbohydrate meal. Meals were taken at various times of day, several days apart.

hormone, which is released by most humans in the early morning, is an insulin inhibitor.

A number of questions about the relationship between response and time of day of the meal are suggested by these data. The length of time that glucose remains elevated appears to be longest for the 2:00 a.m. meal. The response time appears to be shortest for the 10:00 a.m. meal, and the maximum response appears to be slightly lower than at other mealtimes.

Some features that are considered to be of clinical interest (N. Peckinpaugh, personal communication) include the maximum response and the time spent above baseline. Another feature of interest is a dip below baseline, which is often followed by another peak before return to baseline. This is indicative of "insulin resistance", which has high prevalence (about 25%) in healthy populations under clinical testing conditions.

One way to think about these data is that we have a two-way ANOVA with fixed effect "mealtime", random effect "person" and response which is a curve (although measured at 10 or 11 time points).

Notice that, although we do not have a parametric form for the data, the response curves at least roughly have the same shape, and that the time of day effect is fairly pronounced and is similar across subjects. We might also expect that there is error autocorrelation within any one curve, but that curves from the same person at different times of day might be independent (conditional on the person).

According to Crowder and Hand (1990, p. 18) the investigators summarized each curve by area under the curve. Crowder and Hand suggest 3 other measures which they consider to be at least as important: peak value, time to peak, and time to return to baseline.

The analysis we will consider in this paper assumes that the "shape" of the response curve is the same for each individual in the study and for each mealtime. Because we do not have a biological model to give a functional form for the shape, we model it nonparametrically. Subject effects and mealtime effects are summarized by parametric functions of time and response level, making modeling straight forward.

3 Self-Modeling and Shape Invariant Regression

One way to think about curve data is to model the response as coming from a basic response curve that is modified by the covariates. When the shapes of the curves are quite similar, it is natural to think of the covariates acting on the response by stretches of the time or response axes. In this case, we may model the response for the i^{th} curve at time t by

$$Y_i(t) = \mu_i(\tau_i) + \epsilon_{it}. \qquad (1)$$

where $\tau_i = \psi_i(t)$ and $\mu_i = \theta_i(\mu_0)$, and ψ_i and θ_i are parametric functions. The simplest example is the shape-invariant model

$$\begin{aligned} \psi_i(t) &= \beta_{i0} + \beta_{i1}t \qquad \text{and} \\ \theta_i(\mu_0) &= \alpha_{i0} + \alpha_{i1}\mu_0 \end{aligned} \qquad (2)$$

with the α's and β's all constants. These models, with i.i.d. errors and μ_0 defined nonparametrically were first proposed by Lawton et al, 1972. While the modeling in this paper uses the shape-invariant model, most of the computations can be extended, suitably modified, to any parametric ψ and θ. Even nonparametric ψ and θ could be considered, but there are important identifiability issues. Identifiability of the model is discussed in detail in Kneip and Gasser, 1988.

Before discussing estimation I want to point out the nice feature of the model - the parameters provide a convenient summary of the data. Take for example, the glucose data. Figure 3 is a picture of a particular estimate of $\mu_0(t)$ and the corresponding estimate of $\mu_i(\tau_i)$. The area under the curve that is of interest is the shaded area from zero to τ_{ib} (point B on Figure 3a) the return to baseline $\mu_i(0)$. Notice that if $\mu_0(t)$ returns to baseline at t_b, and $\mu_i(t) = \alpha_{i0} + \alpha_{i1}\mu_0(\beta_0 + \beta_{i1}t)$

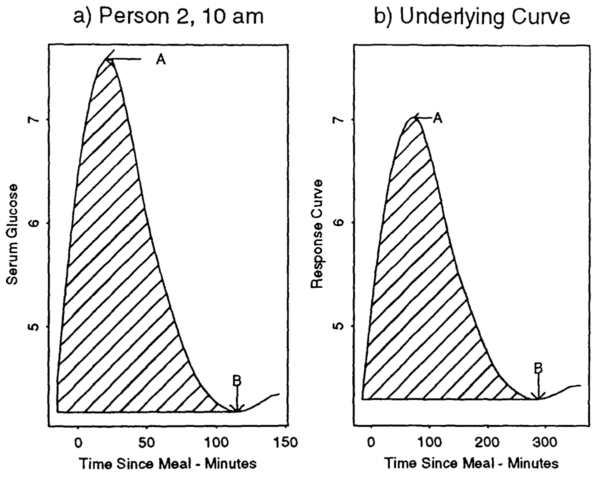

Figure 3: *The fitted curve for person 2 at 10 a.m. (a) and the underlying curve (b), showing the correspondence between summary statistics such as the area under the curve (shaded), height and location of peak (A) and the time of return to baseline. Under the shape invariant model, these summary statistics are functions of the parameters, while the unknown underlying curve is a nuisance parameter.*

then $\tau_{ib} = \beta_{i0} + \beta_{i1} t_b$. If $\int_0^{t_b} \mu_0(t)dt = A_0$ then $\int_{\beta_{i0}}^{\beta_{i0}+\beta_{i1}t_b} \mu_i(t)dt = \alpha_{i1} A_0/\beta_{i1}$. Similarly, if the peak of μ_0 is t_p, the peak of μ_i (point A on Figure 3a) is $\beta_{i0} + \beta_{i1} t_p$, and the value of the peak of μ_i is $\alpha_{i0} + \alpha_{i1}\mu_0(t_p)$. Thus for the shape-invariant model, the unknown curve appears as a nuisance parameter common to all subjects and mealtimes, and would not play any role in analysis these effects.

A number of investigators have discussed estimation of the underlying curve and parameters. Lawton et al (1972) suggested estimating the parameters and μ_0 by minimizing the sum of squares:

$$\sum_{i,t}[Y_{it} - (\alpha_{i0} + \alpha_{i1}\mu_0(\beta_0 + \beta_{i1}t)]^2 \qquad (3)$$

where minimization proceeds by iteratively estimating μ_0 by a nonparametric regression estimator and the parameters by least squares. Kneip and Gasser (1988) talk about model identifiability in detail, and discuss the convergence of the iterative algorithm and consistency of the estimates, as the number of curves goes to infinity and the time points become dense on an interval. They use a Grenander sieve approach (Grenander, 1981) to estimate μ_0 (i.e. a flexible parametric family with the number of parameters increasing with the number of time points)

Härdle and Marron (1990) discuss the special case of 2 curves, and determining if $\mu_1(t) = \theta(\mu_2(\tau))$, using kernel regression to estimate the curves and least squares to estimate the parameters. They show $\sqrt{n}$ convergence of the parameter estimates under a wide choice of bandwidths, and asymptotic normality under a particular choice of bandwidths, when the number of design points for each curve becomes infinite. They also discuss hypothesis testing. Kneip and Engel (1995) use kernel regression to estimate the curves and least squares to estimate the parameters for SIM. They show $\sqrt{n}$ convergence and asymptotic normality under a different bandwidth condition when the number of design points for each curve becomes infinite.

Lindstrom (1995) fits a SIM using regression splines treated selected knots as fixed. This allows her to use approximate likelihood methods which allows extension to mixed models. The computations in this paper are based on fixed effects. Extensions to mixed models are discussed briefly in Section 7.

4 Fitting the Shape Invariant Regression Model

4.1 Computational Algorithms

If μ_0 were known in (1) then the parameters could readily be fitted by non-linear least squares. On the other hand, if the parameters are known, then

$$\mu_0(t) = \theta_i^{-1}(\mu_i(\tau)) \qquad (4)$$

where $t = \psi_i^{-1}(\tau)$.

This suggests the following iterative algorithm for estimating the parameters and μ_0, starting with an initial estimator.

Algorithm 1

1. Smooth all the data to obtain $\hat{\mu}_i$'s.
2. Estimate the θ's and ψ's using least squares, that is, by minimizing $\int (\hat{\mu}_i(t) - \theta_i(\hat{\mu}_0(\tau_i(t)))\, dt$.
3. For the i^{th} curve, substitute all the estimates into (1) to obtain $\hat{\mu}_{0i}$.
4. Average the $\hat{\mu}_{0i}$'s to obtain $\hat{\mu}_0$.
5. Repeat 2-4 until convergence.

The smoothing step may use any nonparametric regression estimator. Härdle and Marron (1990) and Kneip and Engel (1995) use kernel regression. Lindstrom (1995) uses regression splines. In this paper, local polynomial regression is used.

Good performance of smoothing algorithms generally depends on appropriate choice of smoothing parameter. Since the number of observations per curve is small, and the observations within a curve are likely to be correlated, methods based on asymptotics, such as plug-ins and methods based on residual sums of squares, such as leave-one-out cross-validation are not appropriate. Curve-wise cross-validation as in Rice and Silverman (1991) would be a possibility if it were desirable to use the same bandwidth for smoothing all curves. However, as we shall see below, if the treatment or subject effects are large, the bandwidth (or knot placement) needs to be adjusted accordingly.

The smooths are biased estimators of the underlying curves μ_i. Consider, for example, local polynomial smoothing with the shape-invariant model (2). If we use the same bandwidth λ for smoothing each curve, then the asymptotic (in the number of time points per curve) bias is $C\lambda^2\beta_{i1}^2\mu_0^{''}(t)$, which means that we would be averaging curves with different means. This could lead to averaging away shape information. However, if we use bandwidth

$$\lambda/\beta_{i1} \qquad (5)$$

for the i^{th} smooth, we will be averaging curves with almost the same mean. This condition on the relative bandwidth sizes for the smooths is exactly the condition that Härdle and Marron required for asymptotic normality of the estimated parameters.

Another issue of interest is autocorrelation in the errors for each individual. However, if the autocorrelation structure is the same for each person and the autocorrelations decay exponentially over time, the results of Altman (1990) suggest that the convergence rates will be the same as in the i.i.d. case. The results on the relative sizes of the bandwidths will also hold, at least as long as the distribution of measurement times is approximately the same for each individual. The variance of the estimated parameters will undoubtedly be affected by autocorrelation in the errors.

Another algorithm, which has not, to my knowledge, appeared in the literature is:

Algorithm 2

1. Estimate the θ's and ψ's using least squares, that is, by minimizing $\sum(Y_i(t) - \theta_i(\hat{\mu}_0(\tau_i(t)))\,dt$.
2. Obtain an estimate $\hat{\mu}_0$ by simultanously smoothing $\hat{\theta}_i^{-1}Y_i$ against $\hat{\psi}_i(t)$.
3. Repeat 1-2 until convergence.

This algorithm does have two advantages in theory. Firstly, the convergence of nonparametric regression estimators in the papers mentioned in Section 3 requires that the design points become dense on the interval of estimation. However, for many longitudinal studies this type of asymptotic result does not have practical application - it is difficult to take more closely spaced observations on a sampling unit, but possible (at least in theory) to increase the number of sampling units. If the ψ_i's are treated as random with a suitably dense distribution, then even if the number of design points for each sampling is bounded, the design points used in step 2 of Algorithm 2 will become dense on the interval as the number of sampling units increases. (For example, if $\psi_i(t) = \beta_i t$ and the response is recorded at T equally spaced times on the interval, all that is required is that the density of the β's is bounded away from 0 on a suitable interval.) Thus it should be possible to show convergence of Algorithm 2 when the ψ_i's come from a random effects model. This is related to a remark of Lindstrom, 1995. As well, methods for bandwidth selection requiring large sample sizes can be used.

A subtler point concerns error autocorrelation. Using Algorithm 2, smoothing is done against the entire set of values $\{\hat{\tau}_i\}_{i=1\cdots n}$ where n is the number of subjects. As the number of subjects increases (but the number of measurement times per subject remains fixed) measurements which are close in transformed time are unlikely to come from the same individual. For convergence, the smoothing algorithms require that computations are done (effectively) in windows of decreasing width - thus asymptotically data within each window will be uncorrelated.

4.2 Regression Diagnostics

The usual regression diagnostics based on residuals are of course available for the SEMOR model. However, a more sensitive test of the goodness of the model is available. If in Algorithm 1, we do the averaging in Step 4 only within treatment group, we obtain an estimator of the treatment mean curve. If the bandwidths have been selected according to relationship 5 then this provides a good visual assessment of how the regression curve varies with treatment. It is, however, important to note that the selection of bandwidth can have a strong effect on the height and width of features of the estimated curve.

5 Fitting the Glucose Data

The glucose data were fitted using a variant of the shape-invariant model (2) and Algorithm 1. Because time 0 is the time at which the meal was administered, it has a special meaning for the model. Therefore, the time transformation used was

$$\tau_i = \beta_{i1} t.$$

Also, there was some concern that the shape of the response curve varied systematically with mealtime. For this reason, the estimate of μ_0 was developed using the 6 p.m. data only. Local quadratic regression was used, using the the **loess** routine in S-Plus (Becker et al, 1988). The algorithm requires that at least 5 points lie within each smoothing window. Because there were few (10 or 11) time points per curve, and only 6 subjects, no attempt was made to do data-adaptive bandwidth selection. The same adjustment 5 was made for each curve within a mealtime. The bandwidth selected for the mealtime with the maximum value of the mean estimated value of β_1 was 50%, and the bandwidth for each other mealtime was adjusted inversely to the mean estimated value of β_1 for that mealtime.

Algorithm for Glucose Data

```
0. Smooth all the data using a fixed bandwidth.
1. Pick the median person (Jones and Rice, 1992) at 6 p.m. as the starting estimator of $\mu_0$.
2. Iterate over all subjects at 6 p.m. to get an estimator of $\mu_0$.
3. Estimate parameters for each person and mealtime using nonlinear least squares.
4. Resmooth using bandwidths proportional to $1/\bar{\hat{\beta}}_{\bullet j}$ for mealtime $j$.
5. Repeat 1-4.
```

The average curves within each mealtime are displayed in Figure 4. To assess the similarity of the curves, recall that the estimate of μ_0 is the smooth curve for the 6 p.m. meal. We can see that all of the estimated mean curves for the mealtimes have the same basic shape, except for that of the 2 p.m. meal, which has a second mode. Recall that the time and response axes of these plots have been adjusted so that, if the model is correct, the mean curves should be identical - i.e. they should not only look similar in shape, the maxima should have the same location and height and the peaks should have the same width. Except for the extra mode at 2 p.m., the observed differences are likely due to bandwidth selection.

To assess the bimodality of the mean response curve at 2 p.m., ideally we would have confidence bands about the curve. However, methodology for this is not yet available. Instead, an informal assessment is done by comparison with the slightly smoothed data in Figure 1.

We see that at 10 a.m. and 2 a.m. there are some individuals who appear to have a plateau or a second mode following the primary mode. However, these modes and plateaus are much lower than the primary mode. By contrast, at 2 p.m. at least 3 subjects appear to have modes which are comparable in height to the primary mode, and others appear to have a lower plateau. Thus the bimodality of the mean response at 2 p.m. appears to be a plausible feature of the data, which might merit further investigation.

6 More Modeling for the Glucose Data

One of the strengths of the SEMOR model is that it produces a set of regression coefficients which can be used for inference about treatment or covariate effects. To date only asymptotic results for the distribution of parameter estimates are available, which do not seem applicable to a study with 6 subjects and a maximum of 10 time points per curve. Therefore this paper will rely informally on graphical displays and "two-stage" analyses which treat the parameter estimates as input data to standard multivariate techniques.

The data were collected as a randomized complete block design. We might expect some systematic trends in the parameters over time, particularly some type of periodic effect due to diurnal effects.

Figure 5 is a plot of the estimated parameters and their estimated mean over time. Although the mean values of α_{i0} and α_{i1} do have a roughly sinusoidal pattern, the spread in the individual values is very large. By contrast, the mean value of β_{i1} is non-sinusoidal, with a very sharp peak at 10 a.m. (indicating a very short duration of elevated glucose at that mealtime), and little variation otherwise. The individual values are tightly clustered about the mean. A randomized complete block MANOVA, using the parameter estimates as the variables confirmed that the only parameter estimates whose means differ significantly with time are the $\hat{\beta}_{i1}$'s.

As an illustrative approximation to the observed

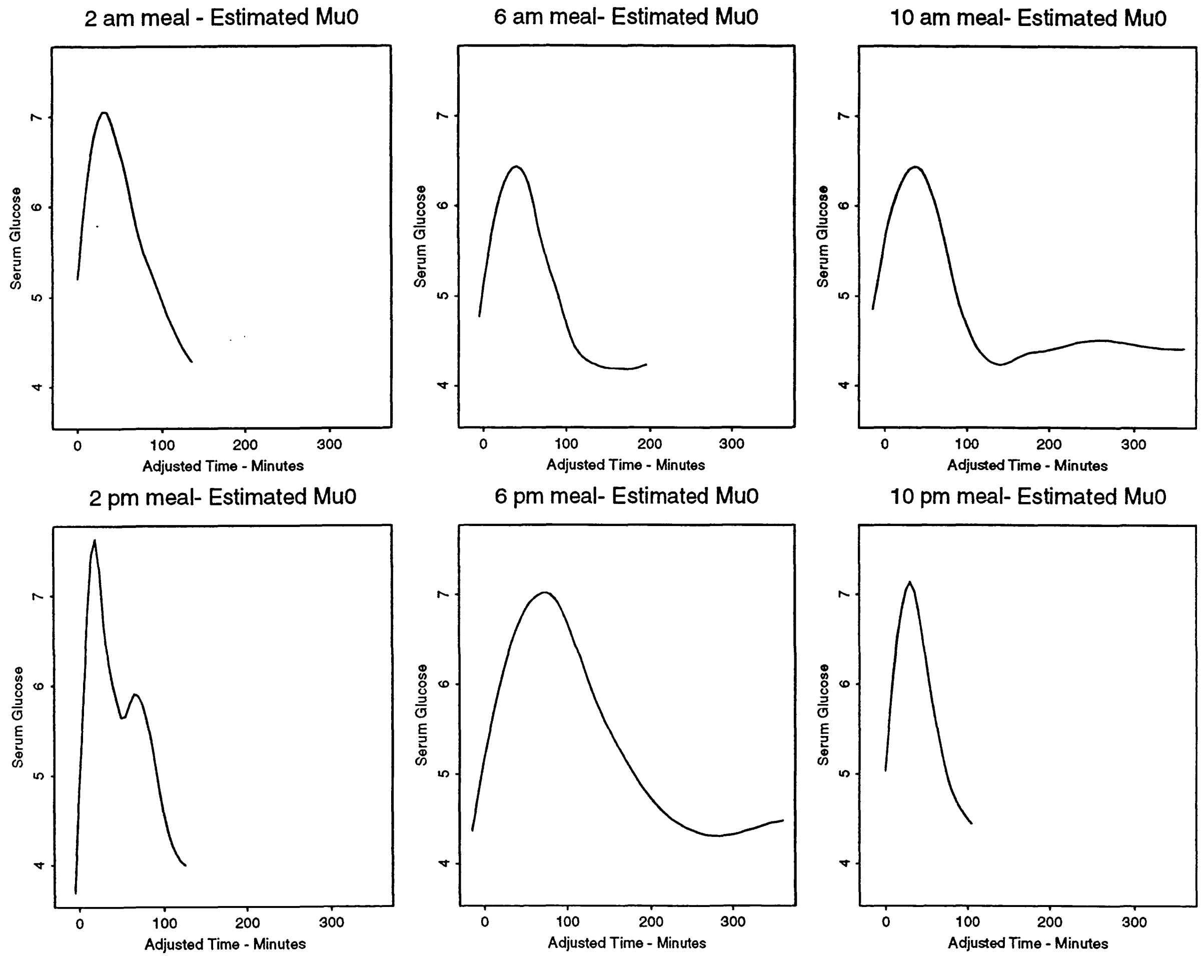

Figure 4: *The estimated mean curve for each mealtime. Notice that the shapes are quite similar except for the double peak for the 2 p.m. meal. The individual data in Figure 1 does show evidence of a double peak for most volunteers at that meal.*

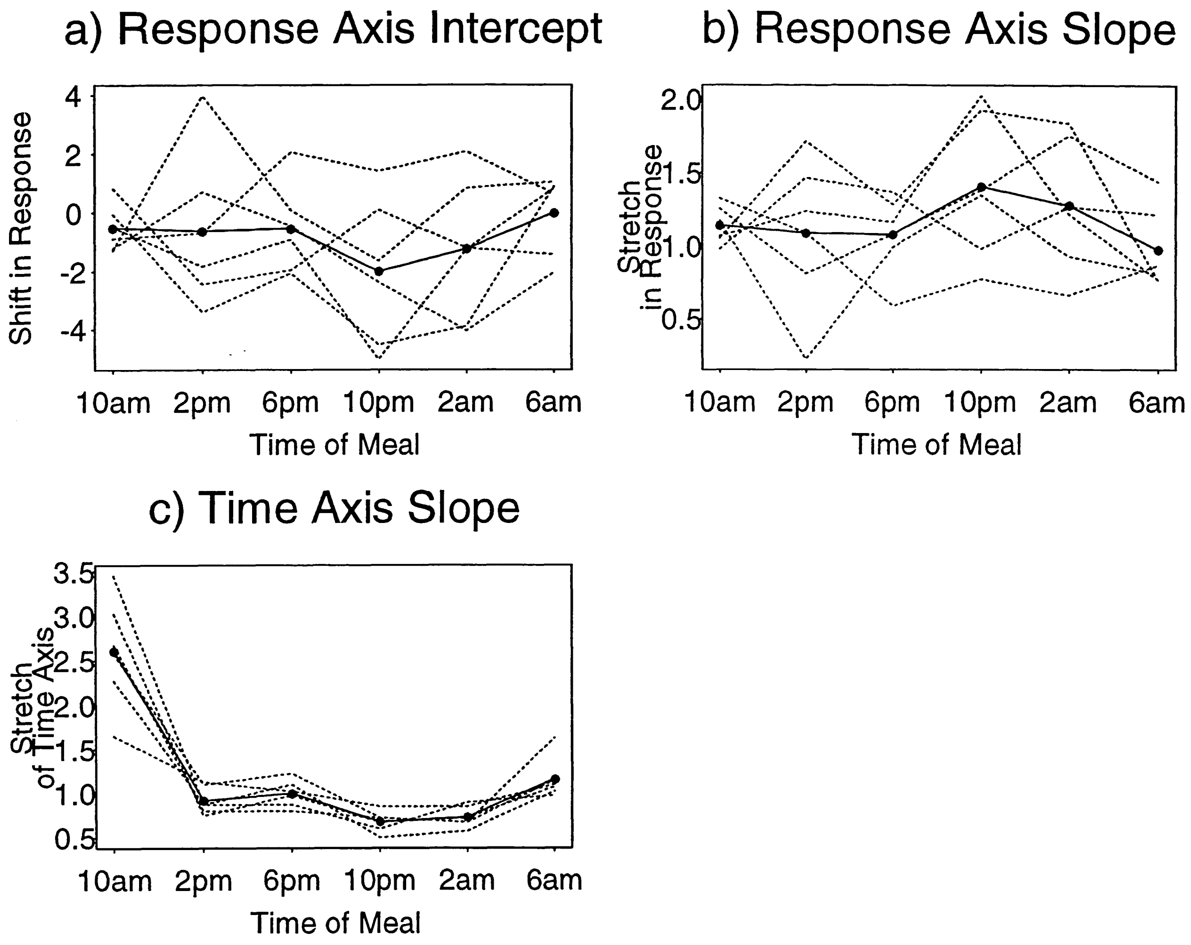

Figure 5: *Estimated parameter values for each volunteer plotted against mealtime. The solid line is the mean. There is some suggestion of a mealtime effect for all of the parameters, but the estimates of the response axis parameters are highly variable. However, the stretch of the time axis shows a very consistent pattern for all subjects.*

pattern, the model

$$\mu_{ij}(t) = \alpha_{i0} + \alpha_{i1}\mu_0\big((\gamma_{i0} + \gamma_{i1} cos\frac{2\pi(j-3)}{24})t\big)$$

was fitted using algorithm 1, where i refers to subject and j refers to mealtime. Notice that this model has 24 parameters (4 per subject) which is very parsimonious compared to the full fixed effect model, which has 108 parameters (3 per subject per mealtime). Although this model cannot pick up the sharp peak in β_{i1} at 10 a.m., it does capture the basic pattern of variation over time. The only "significant" effect detected by the model was the sinusoid. Approximate testing was done using rank tests.

An even more parsimonious model can be developed by treating the parameters as random. In this case, α_{i0}, α_{i1}, γ_{i0} and γ_{i1} are assumed to come from a distribution with unknown mean and correlation structure. The methods of Lindstrom (1995) can be used to simultaneously estimate μ_0 and the distribution of the parameters.

7 Discussion

Although the SEMOR model was proposed in 1972 (Lawton et al, 1972), it seems to have received scant attention in the statistical literature until the late 1980's. However, the statistical user community has shown greater interest. Of the 35 citations listed in the *Science Citation Index* as of summer, 1996, 26 were in applications journals.

More work needs to be done to improve SEMOR as a tool for longitudinal data analysis. In particular, more work needs to be done on fixed effect inference in small samples, where small means either few time points or few experimental units or both. Choice of smoothing parameter will probably turn out to be critical for accurate assessment of the curve, but less important for parameter estimation. Lindstrom (1995) demonstrates a method for fitting random effects to the parametric component. Work on understanding the properties of the estimator is still progressing.

A valuable contribution of the mixed linear model to longitudinal data analysis is the ability to combine time-varying and time-invariant covariates in the same model. Time-varying covariates have the potential to change the shape of the response curve, whereas similarity of the curves over subject and treatment is critical to fitting the SEMOR model. It will therefore take some clever insight to find a way to include time-varying covariates. However, the seminal paper by Lawton et al (1972) does include some suggestions for special cases.

SEMOR has many strengths even at its present state of development. It provides a flexible model for longitudinal data analysis with many advantages over nonparametric and fully parametric modeling, particularly when the shape of the response curve is not specified by an a priori scientific model. The nonparametric part of the model can be used to recover shape information which may provide insight into the underlying mechanisms governing the response. As part of the fitting process, SEMOR can provide graphical diagnostics of goodness of the "same shape" hypothesis. The parametric part of the model can be used to model treatment and covariate effects in a mixed model setting.

Bibliography

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) **The new S language: A programming environment for data analysis and graphics**. Wadsworth:CA

Grenander, U. (1981) **Abstract Inference** John Wiley & Sons: New York.

Härdle, W. and Marron, J.S. (1990) "Semiparametric comparison of regression curves," *The Annals of Statistics*, **18** 63-89.

Jones, M. C. and Rice, J. A. (1992) "Displaying the important features of large collections of similar curves," *The American Statistician*, **46**, 140-145.

Kneip, A. and Engel, J. (1995) "Model Estimation in Nonlinear Regression under Shape Invariance," *The Annals of Statistics*, **23** 551-570.

Kneip, A. and Gasser, T., (1988) "Convergence and consistency results for self-modeling nonlinear regression," *The Annals of Statistics*, **16**, 82-112.

Lawton, W. H., Sylvestre, E.A., Maggio, M.S. (1972) "Self modeling nonlinear regression," *Technometrics.* **14** 513-532.

Lindstrom, M. J. (1995) "Self-modeling with Random Shift and Scale Parameters and a Free-Knot Spline Shape Function". *Statistics in Medicine* **14** 2009-2021.

Lindstrom, M.J. and Bates, D.M. (1990) "Nonlinear mixed effects models for repeated measures data" *Biometrics*, **46** 673-687.

Rice, J. A. and Silverman, B. W. (1981) "Estimating the mean and covariance structure nonparametrically when the data are curves," *Journal of the Royal Statistical Society, Series B*, **53**, 233-243.