SELF-MODELING REGRESSION WITH RANDOM EFFECTS

1560 -M **March 2001**

Naomi S. Altman
and
Julio C. Villarreal

Keywords: longitudinal data, semi-parametric, smoothing, penalized spline, random effects, functional data

Abstract:

In many longitudinal studies, the response can be modeled as a (discretely sampled) curve over time for each subject. Often these curves have a common shape function and individual subjects differ from the common shape by a transformation of the time and response scales. Lindstrom (1995) represented the common shape by a free-knot regression spline, and used a parametric random effects model to represent the differences between curves. We extend Lindstrom's work by representing the common shape by a penalized regression spline, and use a parametric random effects model to represent the differences between curves. The use of penalized regression splines allows for a generalization in the modeling, estimation, and testing of parameters and is easily implemented. An iterative two-step algorithm is proposed for fitting the model.

Conditional on the fitted common shape model, it is possible to fit and test nonlinear mixed effects using standard methods. While the sieve parametric form of the model suggests that a conditional likelihood ratio test should be available for testing whether the shape varies with a time invariant covariate, the null distribution of the likelihood ratio test may not be chi-squared.

1560

Self-Modeling Regression with Random Effects

Naomi S. Altman Julio C. Villarreal
Cornell University EdVISION Corporation

Acknowledgements: *This work was partially supported by NSF grant* DMS #9625350

Abstract

In many longitudinal studies, the response can be modeled as a (discretely sampled) curve over time for each subject. Often these curves have a common shape function and individual subjects differ from the common shape by a transformation of the time and response scales. Lindstrom (1995) represented the common shape by a free-knot regression spline, and used a parametric random effects model to represent the differences between curves. We extend Lindstrom's work by representing the common shape by a penalized regression spline, and use a parametric random effects model to represent the differences between curves. The use of penalized regression splines allows for a generalization in the modeling, estimation, and testing of parameters and is easily implemented. An iterative two-step algorithm is proposed for fitting the model.

Conditional on the fitted common shape model, it is possible to fit and test nonlinear mixed effects using standard methods. While the sieve parametric form of the model suggests that a conditional likelihood ratio test should be available for testing whether the shape varies with a time invariant covariate, the null distribution of the likelihood ratio test may not be chi-squared.

Keywords: longitudinal data; semi-parametric; smoothing; penalized spline; random effects; functional data

1. Introduction

In many longitudinal studies, the response to be modeled is a continuous curve measured over time. In this paper we will consider the Self- Modeling Regression (SEMOR) Model introduced by Lawton, Sylvestre and Maggio (1972) :

$$Y_{ij} = \phi_i\{ \mu_0 [\kappa_i(t_{ij})]\} + \varepsilon_{ij}$$

where Y_{ij} is the response for curve i, i=1, ... ,m, measured at n_i times, t_{ij}. ϕ_i (x) is a monotone inverse link transforming the regression function and κ_i (x) is a monotone transformation of the time axis. μ_0 is a shape function that is common to all the curves, and ε_{ij} are errors. This paper will focus on nonparametric modeling of μ_0 and parametric modeling of ϕ_i (x) and κ_i (x) with known correlation structure for ε_{ij}.

We give special attention Shape Invariant Model (SIM) (Lawton et. al.1972)

$$Y_{ij} = \alpha_{0i} + A_{1i}\ \mu_0 (\beta_{0i} + B_{1i}\ t_{ij}) + \varepsilon_{ij} \qquad (1)$$

Since A_{1i} and B_{1i} should be positive, we express them as $A_{1i} = \exp(\alpha_{1i})$ and $A_{1i} = \exp(\beta_{1i})$

If one has physical or theoretical justification to pre-specify μ_0 (t) parametrically, this is just a special case of nonlinear regression. The semi-parametric SEMOR model allows flexible modeling by estimating μ_0 (t) nonparametrically.

Several different approaches have been studied in fitting the SIM model. Lawton et. al 1972, Kneip and Gasser 1988, Kneip and Engel 1995 considered $\theta_i = (\alpha_{0i}, \alpha_{1i}, \beta_{0i}, \beta_{1i})^T$ to be fixed effects. We will follow Lindstrom (1995) in modeling θ_i as mixed effects.

We model the regression function μ_0 (t) by a penalized regression spline (p-spline) implemented as a linear mixed effects model. Estimation of the regression function μ_0, mixed effects, and variance components will be done via a two-step iterative algorithm consisting of a linear mixed effects step (p-spline fitting) and a nonlinear mixed effects step (mixed effects and variance component estimation).

2. P-Splines and the GML Method

The regression spline model of order $p \geq 1$ is:

$$\mu_o(t;\boldsymbol{\beta}) = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_p t^p + \beta_{p+1}(t-\tau_1)_+^p + \ldots + \beta_{p+k}(t-\tau_k)_+^p$$

where the parameters to be estimated are $\beta=(\beta_0, \ldots, \beta_{p+k})$ and $\{\tau_1, \ldots, \tau_k\}$ are k fixed knots with $a \leq \tau_1 < \ldots < \tau_k \leq b$ and $(x)^p_+ = x^p I_{\{x \geq 0\}}$. We use a relatively large number of knots with a quadratic penalty function (Ruppert and Carroll 1997) to reduce overfitting. The estimator β_λ^* is defined as the minimizer of

$$\sum_{i=1}^{n} \{Y_i - \mu_o(t;\boldsymbol{\beta})\}^2 + \lambda \sum_{j=1}^{k} \beta_{p+j}^2 \qquad (2)$$

where λ is the smoothing parameter.

This can be reformulated as a linear mixed model: $\mathbf{Y} = \mathbf{X}^*\boldsymbol{\gamma} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$ where

$$x_i^* = \left(1, t_i, t_i^2, \ldots t_i^p\right) \qquad (3)$$
$$z_i = \left((t_i-\tau_1)_+^p, (t_i-\tau_2)_+^p, \ldots (t_i-\tau_k)_+^p\right)$$

and ε is distributed multivariate normal with mean equal to the zero vector and covariance matrix $\sigma_\varepsilon^2 \mathbf{I}$, γ is the fixed effect and u is the random effect, independent of ε, with u~i.i.d $N(0, \sigma_u^2 \mathbf{I})$. Comparing the likelihood and parameters of this model with (2), we see that the smoothing parameter λ is replaced by the ratio of the variance components $\sigma_\varepsilon^2/\sigma_\gamma^2$, and that fitting can be done using standard linear mixed effects software. This is the GML method of Wahba, 1985.

3. Model Formulation

We extend SIM model (1) to consider mixed effects for the scaling and shifting parameters.

$$Y_{ij} = \alpha_0 + a_{0i} + \exp[(\alpha_1 + a_{1i})]\ \mu_0\ [\beta_0 + b_{0i} + \exp[(\beta_1 + b_{1i})]\ t_{ij}] + \varepsilon_{ij} \qquad (4)$$

for i=1, ... ,m; j=1, ... ,n_i, t_{ij} in (a,b) where (α_0, α_1, β_0, β_1) are fixed effects and (a_{0i}, a_{1i}, b_{0i}, b_{1i}) are random effects. To impose identifiability in the SIM model, we constrain (α_0, α_1, β_0, β_1) to be (0,0,0,0), the mixed model equivalent to setting the sum of each parameter equal to zero as suggested by Kneip and Engel (1995). The random effects (a_{0i}, a_{1i}, b_{0i}, b_{1i}, ε_{ij}) are modeled as independently distributed multivariate normal:

$$\varepsilon_i \sim i.i.d.N\left(0_{n_i}, I_{n_i}\sigma_\varepsilon^2\right) \qquad u_i \sim i.i.d.N\left(0, \sigma_u^2\right)$$
$$a_{0i} \sim i.i.d.N\left(0, \sigma_{a_0}^2\right) \qquad a_{1i} \sim i.i.d.N\left(0, \sigma_{a_1}^2\right)$$
$$b_{0i} \sim i.i.d.N\left(0, \sigma_{b_0}^2\right) \qquad b_{1i} \sim i.i.d.N\left(0, \sigma_{b_1}^2\right)$$

From the model we have formulated, we can compute maximum likelihood estimators, but this is computationally difficult. Instead, we handle this estimation in a two step approach: estimating the parameters for the p-spline that model μ_0, and then the variance components, fixed effects and the Best Unbiased Predictors (BUPS) which enter the model nonlinearly. Note that we can also generalize Equation (4) to the Bayesian context and allow for non-normal priors to be placed on the scale and shift parameters. We can readily extend to parametric structure for error variance.

Equation (1) suggests the following algorithm. Let

$$\boldsymbol{\theta}_i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i})$$
$$= (\alpha_0 + a_{0i}, \alpha_1 + a_{1i}, \beta_0 + b_{0i}, \beta_1 + b_{1i})$$

and

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_m^{\mathrm{T}}).$$

SIM Algorithm

Step 0 Choose initial estimates of $\theta_i^{(0)} = (0,0,0,0)$. Set k=0.

Step 1 Transform data and time with

$$Y^*_i{}^{(k)} = (Y_i - \theta_{0i}{}^{(k)})/\exp(\theta_{1i}{}^{(k)}) \text{ and } t_i{}^{*(k)} = \theta_{3i}{}^{(k)} + \exp(\theta_{4i}{}^{(k)})\ t_i$$

Step 2 Using LME, estimate $\gamma^{(k)}$ and $u^{(k)}$ by fitting:

$$Y^{*(k)} = X(t_i^{*(k)})\,\gamma^{(k)} + Z(t_i^{*(k)})\,u^{(k)}$$

Step 3 Using NLME, estimate $\theta^{(k+1)}$ by fitting the model

$$Y_{ij} = \alpha_0 + a_{0i} + \exp(\alpha_1 + a_{1i})\{X(\beta_0 + b_{0i} + \exp(\beta_1 + b_{1i})\,t_{ij})\,\gamma^{(k)} + Z(\beta_0 + b_{0i} + \exp(\beta_1 + b_{1i})\,t_{ij})\,u^{(k)}\} + \varepsilon_{ij}$$

Check for convergence. Else normalize the parameters and set
$\theta_i^{(k+1)} = (a_{0i}^{*(k+1)}, a_{1i}^{*(k+1)}, b_{0i}^{*(k+1)}, b_{1i}^{*(k+1)})$ Go to step 1.

There are several convergence criteria that can be used to terminate the algorithm. Lawton et al (1972), terminated the fitting algorithm when MSE of the parameters converged. We terminated the algorithm when the change in the log-likelihood in the nonlinear mixed effects step converges.

Although the theoretical rate of convergence of the p-spline as a sieve estimator has not to our knowledge been established, it should be similar to the rate for B-splines, which was determined by Shen and Wong (1994). Simulation results in Villarreal (2001) support the convergence of the p-spline with GML selection of the smoothing parameter if the number of knots increases with the sample size.

To determine the efficacy of the computational method, we performed a small simulation study, using 3 underlying curves for the common shape, two levels of number of data points per curve, three levels of numbers ofcurves and two levels of variance components.

Fitting was done using the `lme` and `nlme` procedures in Splus. Convergence of the algorithm depends on convergence of all steps: `lme` to fit the shape, `nlme` to fit the variance components, and the iterations between `lme` and `nlme`. The `lme` step always converged. The `nlme` step was more problematic. There were frequent failures, even when the iterative procedure appeared to be close to convergence. However, when the `lme` and `nlme` both converged, the algorithm usually converged quickly.

The MISE decreases as a function of both the number of data points per curve and the number of curves per set. The variance components for the location parameters are estimated quite well. There appears to be a small downwards bias. There is some downwards bias in the estimating the variance components for the scale parameters.

5. Application to Spirometer Data

For a real application, we considered the spirometer data used by Lawton et. al. (1972) and Lindstrom (1995). We fitted the SIM model, using 10 equally spaced knots. The rescaled data are displayed in Figure 1, with the fit of the common shape. The lower right panel is a plot of all the rescaled data on the common time and response scale.

The final estimates for the regression mean function fit the data very closely. Lindstrom (1995) also analyzed these data.

6. More Complex Modeling

Often the set of curves may be collected as part of a designed experiment, or there may be a time-invariant covariate which may explain differences among curves. In this case, we may wish to consider a model in which the curve is a function of the design or covariate. For example, we may look at drug uptake curves as a function of delivery method or dose.

The SEMOR model allows us to simply formulate tests of the covariate effect on θ. After estimating the shape function μ_0, we can condition on the estimate. We then fit a model involving the covariates as if μ_0 were a known parametric function. All the tools for nonlinear parametric mixed models are then available for conditional tests and confidence intervals.

If, instead, we are interested in determining whether the shape of the curve is affected by a covariate, we can use the estimated parameters, θ, to align the curves. We can then test for equality of the aligned curves across treatments, or for different values of the covariates. While it is tempting to base a likelihood ratio test on the linear mixed model fitting method, simulation studies by Altman as well as others (D. Ruppert, personal communication) show that the null distribution does not have the mixture of chi-squared distribution that would be suggested by statistical theory.

The SEMOR methodology has been shown to be very feasible in a mixed model setting. Many problems in growth modeling, pharmacokinetics, materials science other fields involve curves with similar shape. The SEMOR method provides a powerful tool for modeling these curves.

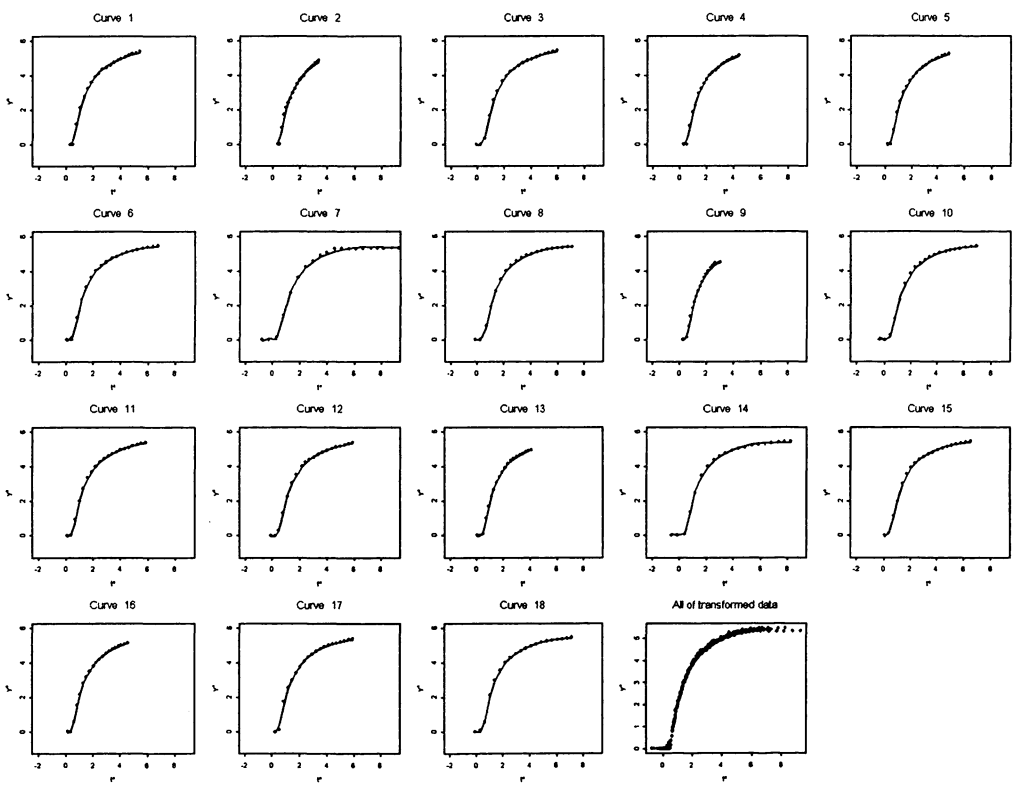

Figure 1: *Scaled and shifted spirometer data (...) with individual estimated shifted curve (-)).*

7. References

Eilers, P.H.C. and Marx, B.D. (1999), "Generalized Linear Models on Sampled Signals and Curves: A P-Spline Approach",*Technometrics*, **41**, 1-13.

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

Kneip, A. and Engel, J., (1995), "Model estimation in non-linear regression under shape invariance", *The Annals of Statistics*, **25**, 551-532.

Lawton, W.H., Sylvestre, E.A. and Maggio, M.S., (1972), "Self modeling non-linear 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.

Ruppert, D., Carroll, R.J., (1997) *Penalized Regression Splines*, tech. report, Operations Research, Cornell University.

Shen, X. and Wong, W. H., (1994), "Convergence Rate of Sieve Estimates", *The Annals of Statistics*, **22**, 580--615.

Wahba, G. (1985) "A Comparison of GCV and GML for Choosing the Smoothing Parameters in the Generalized Spline Smoothing Problem", *The Annals of Statistics*, **4**, 1378-1402.

Villareal, J. (2001) *Self-modeling Nonlinear Regression with Random Shifts and a Penalized Regression Spline Shape Function* M.S. Thesis, Cornell University.