P-Splines approaches for longitudinal data: Flexible modelling proposals and their advantages

Idoia García Ledo[1], *Vicente Núñez-Antón*[2], *Susan Orbe Mandaluniz*[3]

[1]igarcia253@ikasle.ehu.eus, Departamento de Economía Aplicada III (Econometría y Estadística), Universidad del País Vasco UPV/EHU

[2]vicente.nunezanton@ehu.eus, Departamento de Economía Aplicada III (Econometría y Estadística), Universidad del País Vasco UPV/EHU

[3]susan.orbe@ehu.eus, Departamento de Econometría Aplicada III (Econometría y Estadística), Universidad del País Vasco UPV/EHU

Abstract

The use of parametric models may not be appropriate in the analysis of longitudinal data. We proposed alternative methods based on penalized splines nonparametric smoothing techniques. We describe the proposed methodology and illustrate its use in the analysis of a real data set, the *cattle data*, making spacial emphasis on the limitations the proposed models have, as well as on their modelling flexibility.

Keywords: P-Splines; Mixed models; Longitudinal data.

1. Introduction

The usual complexity real data feature nowadays results in the fact that the use of complete parametric models may not be appropriate and, thus, authors have presented alternative nonparametric modelling proposals such as penalized splines (P-Splines) (Eilers and Marx, 1996), that have recently been shown to be popular modelling approaches (Ruppert et al., 2003, Durbán et al., 2004). This methodology proposes the development of statistical inferential procedures for which no hypothesis about the specific functional form modelling the relation between the response variable and the independent variables is not required.

We concentrate on the analysis of longitudinal data, a very common type of data used among researchers in the area, which results from measuring one or more response variables, together with some subject-specific or experimental unit-specific variables, along time. This is one of the reasons that motivates the use in practice of P-splines specified as mixed models in the context of longitudinal data analysis, because they allow the inclusion of both fixed and random effects in the proposed model, besides the model's usual error terms (Pinheiro and Bates, 2000). Moreover, given the existing correlation between measurements taken on the same individual or experimental unit in longitudinal data, we include the possibility of having correlated data, as well as having error variances that may be constant or not. That is, for example, if the model requires the possibility of allowing for having different variances for the different levels a given grouping variable may have.

2. Proposed methodology

Our modelling proposal for the analysis of longitudinal data motivates the use of mixed models, which will also include the possibility of having P-splines specified as mixed models in the last three proposed alternative models in the next subsections,

2.1. Model with random intercepts

An initial modelling approach in a fixed model setting to try to explain the individual's behavior in this type of data is the use of fixed intercepts, one for each individual in the data set under study. However, this model only includes a common slope. Moreover, it is also a model that is very difficult to estimate and interpret, which makes it a not so efficient modelling proposal. Therefore, we propose the use of a model that includes random intercepts, U_i's, so that $U_i \sim N(0, \sigma_U^2)$, with $\sigma_U^2 > 0$, for which only one parameter, representing the variance model component, should be estimated (i.e., σ_U^2), whereas, for the fixed intercepts model, too many parameters should be estimated. In addition, this modelling proposal allows for testing, by using the restricted maximum likelihood method (REML), if the included population randomness is statistically significant or not (i.e., if $\sigma_U^2 = 0$ or $\sigma_U^2 > 0$). In this way, this model assumes that *all population members have a linear and equal growth rate and that, in addition, the variability among individuals is modelled with the sole inclusion of the random effect* U_i, so that individual curves only differ from each other in their intercepts.

2.2. Mixed additive model

As we have described in the previous sections, we propose the use of nonparametric regression methods, so that a specific and maybe not appropriate parametric model is not assumed and, in some way, the funcional form of the nonparametric function $f(\cdot)$ will be given by the "data behavior" or will be indirectly specified as a result from a "data driven method." In this way, $f(\cdot)$ is a non-specified "smooth" function that requires to be estimated from the N pairs of data (X_{ij}, y_{ij}), with an estimation method given by the P-splines methodology, with the linear P-spline written as:

$$f(X_{ij}) = \sum_{p=1}^{P} u_p (X_{ij} - \omega_p)_+, \tag{1}$$

where $u_p \sim N(0, \sigma_u^2)$, with $\sigma_u^2 > 0$. Therefore, in order to estimate the nonparametric function $f(\cdot)$ we need to estimate the variance component parameter σ_u^2. Moreover, we can also test for the need to include the nonparametric smoothing function $f(\cdot)$ in the model (i.e., $\sigma_u^2 > 0$). The basis for the trimmed function that incorporates the estimation penalty term is given by $(X_{ij} - \omega_p)_+$, where $x_+ = x$ for positive x, and zero, otherwise, and the ω_p's are the different knots for the function being considered in the penalty term. In our case, in order to be able to determine the number of knots to be used, we follow the proposal in Ruppert (2002), where:

$$\textit{number of knots (P)= max (5, min [40, unique values of x / 40])} \tag{2}$$

Thus, in this model we also assume that all population members have equal growth rate, and that they differ from each other in their intercepts. However, this model includes the possibility that

population members do not have to necessarily feature a linear growth rate and, in addition, its linearity or not would depend on the function $f(\cdot)$, to be estimated by P-Splines.

2.3. Model with individual linear differences

The previous model may still have unnecessarily simplified the way relation between variables is modelled because it assumed that all individuals curves are parallel, with the same functional form and, in addition, not allowing for the possibility of modelling the different individual curves in an appropriate way. This is the reason why we propose the model with individual linear differences, so that they can also be given by the slope, thus allowing for individual curves to be different. In this model, instead of using the random intercepts U_i's, we propose to use the functional form $a_{i1} + a_{i2}X_{ij}$, where $(a_{i1}, a_{i2})' \sim N(0, \Sigma)$. By using a_{i1}, we allow for different individuals' random intercepts, whereas by using $a_{i2}X_{ij}$, we allow for different individuals' random slopes. Moreover, Σ is the variance-covariance matrix for the random intercepts and slopes, a_{i1}'s and a_{i2}'s.

Therefore, we consider the existing possibility of modelling individuals' heterogeneity not only with respect to their starting measurement (i.e., intercept), but also along time, because we are now assuming that *population members do not have a similar or linear growth rate, allowing for different functional forms that may be different for the different individuals.*

2.4. Model with specific individual curves

The previous model included the nonparametric smoothing function $f(\cdot)$ (mixed additive model and model with individual linear differences); that is, a global function that did not depend on individual characteristics. However, there are more flexible modelling proposals that allow for the possibility that the specific individual differences are also a nonparametric function, $g_i(\cdot)$, with both a linear and a nonlinear component (Ruppert et al., 2003), where both components are assumed to be random. That is,

$$g_i(X_{ij}) = a_{i1} + a_{i2}X_{ij} + \sum_{p=1}^{P} v_p(X_{ij} - \omega_p)_+, \tag{3}$$

where $(a_{i1}, a_{i2})' \sim N(0, \Sigma)$, and $v_p \sim N(0, \sigma_v^2)$, with $\sigma_v^2 > 0$.

As in previous models, we can also test for the statistical significance of the individual nonparametric function $g_i(\cdot)$ (i.e., $\sigma_v^2 > 0$). However and given that we use trimmed linear basis for the specific individual curves and that the number of measurements per individual may be small, we will use a smaller number of knots for the basis used for each individual as compared to the number of knots used with the function $f(\cdot)$ included in the two previous models.

3. Analysis of the *cattle data*

To illustrate the usefulness of the penalized splines as mixed models modelling approach proposed in the previous sections, we analyze the *cattle data*, which corresponds to an experiment described by Kenward (1987), in which cattle receiving two treatments for intestinal parasites were weighed 11 times over a 133-day period. For all of the aforementioned models and for modelling the

correlation between observations on the same subject, we use an *autoregressive structure of order two*, and we have also considered *different variances for the different treatments*. General conclusions from the analysis are as follows:

- We were not able to fit the model with different specific curves for each individual because the algorithm did not converge. We concluded that this is due to the fact that *the number of individuals, the number of measurements per individual and the observations in the data set under study are relevant factors to be able to estimate such a complex model*. Therefore, we have only considered and fitted the model with random intercepts, the mixed additive model and the model with individual linear differences. Model selection was based on three criteria: *AIC* (Akaike's Information Criterion), *BIC* (Bayesian Information Criterion), and *logLik* (logarithm of the restricted likelihood ratio method-REML). The best fitting and selected model was **the model with individual linear differences**.

- If we compare the results obtained to those from previous analyses of this data set (Zimmerman and Núñez-Antón, 2001), we conclude that *the model specification is not able to model changing means* and that, in addition, *the methodology used is not able to discriminate between the two different treatment groups because of the lack of flexibility the proposed models' specification has.*

4. Bibliography

[1] Durbán, M., Harezlak, J., Wand, M.P. and Carroll, R.J. (2004). *Simple fitting of subject-specific curves for longitudinal data*. Statistics in Medicine, **24(8)**, 1153-1167.

[2] Eilers, P. and Marx, B. (1996). *Flexible smoothing with B-splines and penalties*. Statistical Science, **11(2)**, 89-121.

[3] Kenward, M.C. (1987). *A method for comparing profiles of repeated measurements*. Applied Statistics, **36(3)**, 296-308.

[4] Pinheiro, J.C. and Bates, B.D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

[5] Ruppert, D. (2002). *Selecting the number of knots for penalized splines*. Journal of Computational and Graphical Statistics, **11(4)**, 735-757.

[6] Ruppert, D., Wand, M. and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

[7] Zimmerman, D.L. and Núñez-Antón, V. (2001). *Parametric modelling of growth curve data: An overview (invited paper with discussion)*. Test, **10(1)**, 1-73.