XI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 24-25-26 de outubro de 2013

P-Splines for Longitudinal Data. Application to the Study of Children Growth with Phenylketonuria

Ipek Guler[1], Cesar Sanchez Sellero[2], Carmen Cadarso Suarez[3], Francisco Gude Sampedro[4], María Xosé Rodríguez Álvarez[5]

1.ipek__guler@hotmail.com Unidad de Epidemiología Clínica **2**.Universidad de Santiago de Compostela **3**.Universidad de Santiago de Compostela **4**.Unidad de Epidemiología Clínica **5**.Unidad de Epidemiología Clínica

Abstract

Phenylketonuria (PKU) is a rare disease that affects the growth of children and Hyperphenylalaninemia (HPA) is a medical condition which mostly results in PKU. The objective of this study is to compare the long-term growth of children with PKU and HPA. To this aim, we model the growth of children on the basis of a flexible mixed model including subject - specific curves and factor by curve interactions (Durban et al, 2005).

Keywords: Phenylketonuria, logitudinal study, penalized splines.

1. Introduction

Phenylketonuria (PKU) is a rare metabolic disorder that affects the way the body breaks down protein. If not treated shortly after birth, PKU can be destructive to the nervous system, causing mental retardation. PKU is caused by a mutation in a gene on chromosome 12. The gene codes for a protein called PAH (phenylalanine hydroxylase), an enzyme in the liver. This enzyme breaks down the amino acid phenylalanine into other products the body needs. When this gene is mutated, the shape of the PAH enzyme changes and it is unable to properly break down phenylalanine. Phenylalanine builds up in the blood and poisons nerve cells (neurons) in the brain.

Hyperphenylalaninemia (HPA) is a medical condition characterized by mildly or strongly elevated levels of the amino acid phenylalanine in the blood. Severe hyperphenylalaninemia results in phenylketonuria (PKU). While in Benign HPA isn't necessary to make a therapeutic intervention, in PKU, for the correct physical and mental development of the patients, it is fundamental to have a treatment. The patients of PKU should be treated with a special diet in order to have a favourable evolution.

The objective of this study is to compare the long-term growth of children with PKU and HPA followed up in "Unidad de Diagnóstico y Tratamiento de Enfermedades Congénitas del Metabolismo del Hospital Clínico Universitario in Santiago de Compostela" between the years 1978 and 2011. The children were classified depending on the PAH level: Classical PKU (PAH > 1200 mol/L), Moderated PKU (PAH 360-1200 mol/L); and Benign HPA (PAH 120-360 mol/L). In the first two groups, the children followed a dietetic treatment restricted in natural proteins, whereas in the third one, they followed a normal diet. For the purpose of this study we compared the group of individuals with PKU (either classical or moderated) and the group with HPA.

Longitudinal data is nowadays an active topic of current statistical research. An example of such data can be the growth curves measured over time. In many practical situations, it could be erroneous to model these curves using traditional parametric regression techniques. Misleading conclusions may be reached if the time effect is incorrectly specified. This problem can be overcome by using more general flexible regression techniques, such as Penalized Splines (P-Splines, Eilers and Marx, 1996). This technique has the advantage of not assuming a parametric form for the effect of time, and avoids the need to impose functional assumptions. The only assumption required is that the time effect follows a smooth function.

In this study we model the growth of children, on the basis of a semi parametric model proposed by Durbán et al. (2005). In this approach, the growth curve is modelled as P-Splines, and model's estimation is based on the mixed model representation of a P-Spline (Currie and Durbán, 2002). Moreover, the proposal by Durbán et al. (2005) also allows to model the subject - specific curves as P-Splines with random coefficients, and to incorporate factor by curve interaction. In our study, it would allow to study and compare the long-term growth of children with PKU and HPA.

2. Application to the Study of Children Growth with Phenylketonuria

2.1 Data Source

The dataset contains observations from 102 children (43 boys and 59 girls), followed in the *Hospital Clínico Universitario in Santiago de Compostela* from 6 months up to 18 years. The measurements were recorded from successive controls that were carried out during the consultation or from the revision of the clinical histories of the patients. Specifically, the dataset contains the following information:

- **Age**: Age of the children when the measurements were taken (6 months up to 18 years).
- **Height**: Height (kg) of the children at each measuring time
- **Z-score**: Measure that express the distance between an individual child's height and the average height of comparable children in a reference population.
- **Type of pathology**: Categorical covariate that indicates the type of pathology of the children (either HPA or PKU).
- **Gender**: Female - Male.

2.2 Statistical Analysis

For the sake of illustration, we present in this work the results for the analysis of the long-term growth based on the z-score. The analyses were done separately in boys and girls. Let $y_i = (y_{i1},...,y_{in_i})^T$ denote the z-scores of child *i* at ages $t_i = (t_{i1},...,t_{in_i})^T$. The following complete model was assumed:

$$y_i = \alpha_0 + \alpha_1 pth_{iPKU} + f(t_i) + h(t_i) pth_{iPKU} + g_i(t_i) + \varepsilon_i,$$

where $pth_{iPKU} = 1$ if the pathology of individual *i* is PKU and 0 otherwise, $f(\cdot)$ represents the smooth effect of age in the group with HPA, $h(\cdot)$ represents the deviation from this effect for those individuals with PKU, and $g_i(\cdot)$ is the subject - specific smooth effect of age on individual i. On the basis of this model, the age effect in an individual with PKU is therefore $f(t) + h(t)$. Accordingly, if $h(t) = 0, \forall t$, then no interaction between age and type of pathology is present.

To estimate this model, functions $f(\cdot), h(\cdot)$ and $g(\cdot)$ were represented by cubic B-splines with 10 knots and second order penalties. The mixed model representation of the model is (see Lee and Durbán, 2011):

$$y_i = \alpha_0 + \alpha_1 pth_{iPKU} + \underbrace{\tilde{X}_i \beta + Z_i \tilde{u}}_{f(t_i)} + \underbrace{\left(\tilde{X}_i \gamma + Z_i w\right) pth_{iPKU}}_{h(t_i)} + \underbrace{X_i \beta^i + Z_i \tilde{u}^i}_{g_i(t_i)} + \varepsilon_i,$$

where

$$\tilde{u} \sim N\left(\vec{0}, \sigma_{\tilde{u}}^2 I_{10}\right), w \sim N\left(\vec{0}, \sigma_w^2 I_{10}\right), \beta^i \sim N\left(\vec{0}, \Sigma_{2x2}\right), \tilde{u}^i \sim N\left(\vec{0}, \sigma_{\tilde{u}^*}^2 I_{10}\right), \varepsilon_i \sim N\left(\vec{0}, \sigma_\varepsilon^2 I_{n_i}\right),$$
$$\tilde{X}_i = t_i, X_i = [1 : t_i], Z_i = B^i U_s \tilde{\Sigma}^{-0.5},$$

where B^i denotes the B-spline design matrix for individual *i*, and U_s and $\tilde{\Sigma}$ come from the Singular Value Decomposition of the penalty matrix D:

$$D^T D = \left(U_n : U_s\right) \begin{pmatrix} 0_2 & 0 \\ 0 & \tilde{\Sigma} \end{pmatrix} \begin{pmatrix} U_n \\ U_s \end{pmatrix}.$$

In matrix notation,

$$Y = X\theta + Zu + \varepsilon,$$

with

$$X = \begin{bmatrix} 1 & 0 & t_1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & t_{n_{HPA}} & 0 \\ 1 & 1 & t_{n_{HPA}+1} & t_{n_{HPA}+1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & t_N & t_N \end{bmatrix}$$

where n_{HPA} is the number of individuals with HPA and $\theta = \left(\alpha_0, \alpha_1, \beta, \gamma\right)^T$ and

$$Z = \begin{bmatrix} Z_1 & 0 & X_1 & 0 & \cdots & 0 & Z_1 & 0 & \ldots & 0 \\ Z_2 & 0 & 0 & X_2 & \cdots & 0 & 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_{n_{HPA}} & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ Z_{n_{HPA}+1} & Z_{n_{HPA}+1} & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_N & Z_N & 0 & 0 & \ldots & X_N & 0 & 0 & \ldots & Z_N \end{bmatrix}$$

$$u = \left(\tilde{u}^T, w^T, \tilde{\beta}^{1T}, ..., \tilde{\beta}^{NT}, \tilde{u}^{1T}, ..., \tilde{u}^{NT}\right)^T,$$

and the random effect covariance matrix G is

$$G = \begin{pmatrix} \sigma_{\tilde{u}}^2 I_{10} & 0 & 0 & 0 \\ 0 & \sigma_w^2 I_{10} & 0 & 0 \\ 0 & 0 & \textit{block diagonal } \Sigma_{2x2} & 0 \\ 0 & 0 & 0 & \textit{block diagonal } \sigma_{\tilde{u}^*}^2 I_{10} \end{pmatrix}.$$

3. Conclusions

Results of the fitted models are presented in Figures 1 and 2. Figure 1 shows, for boys and girls, the estimated long-term growth profiles for individuals with HPA and PKU, as well as the difference between these two curves. In Figure 2, the estimated subject - specific profile for six different individuals (three boys and three girls) is shown. Finally, we performed the likelihood ratio test to investigate whether the different random effects incorporated in the model were necessary. Table 1 and 2 shows the results of the restricted likelihood test (RLRT) for each variance parameter of the random effects.The results show that the subject - specific curves and the non linear part of the time effect are appropriate for each gender. However, we finally have obtained two different appropriate models for boys and girls, there is no evidence to suggest that long-term differences between the two types of pathologies are present for male individuals but we keep this long-term differences between the two types of pathologies for the female ones.

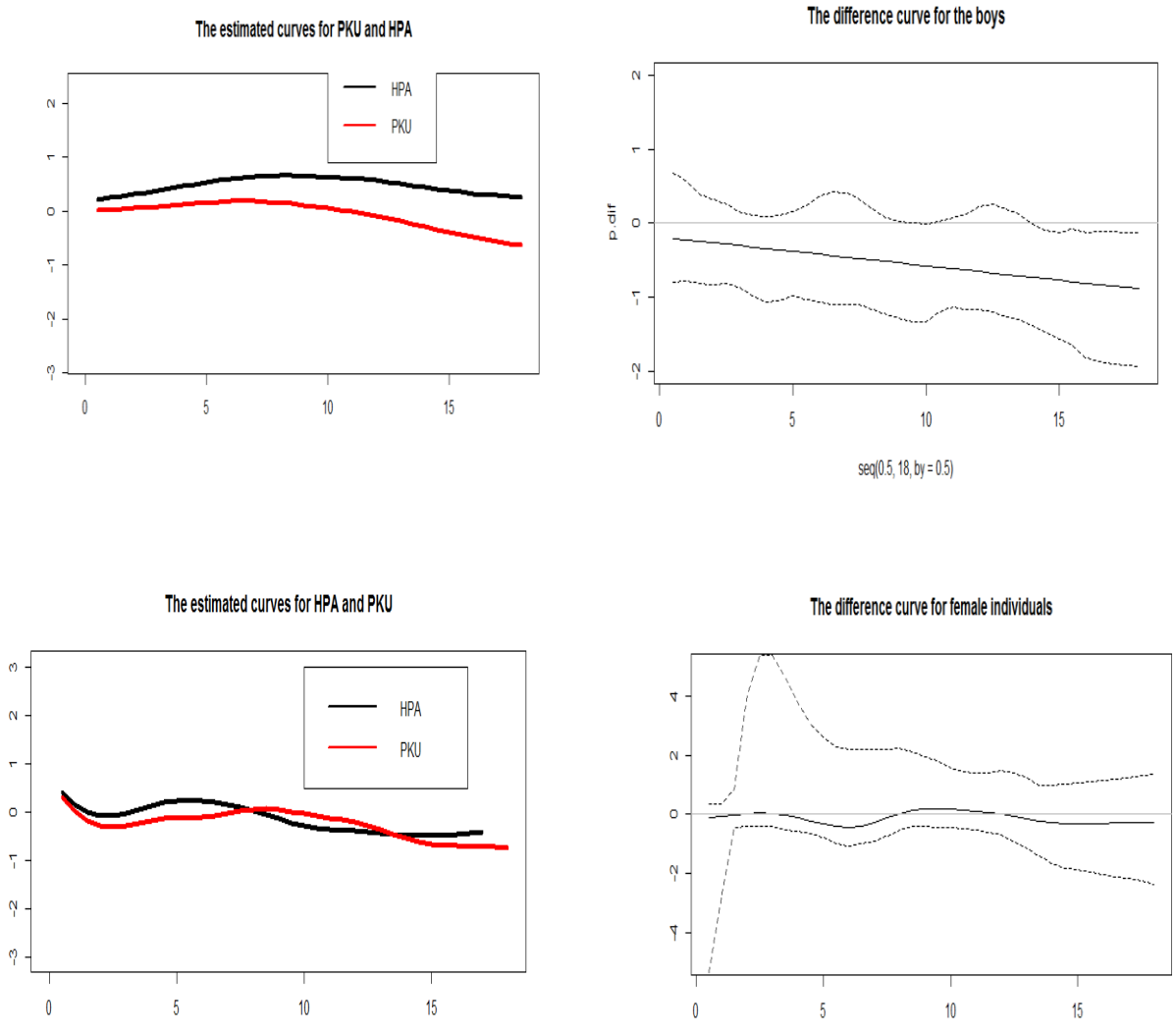

Figure 1. Left: estimated long-term growth profiles for individuals with HPA and PKU. Right: estimated difference between HPA and PKU curves. Top: Boys. Bottom: girls.

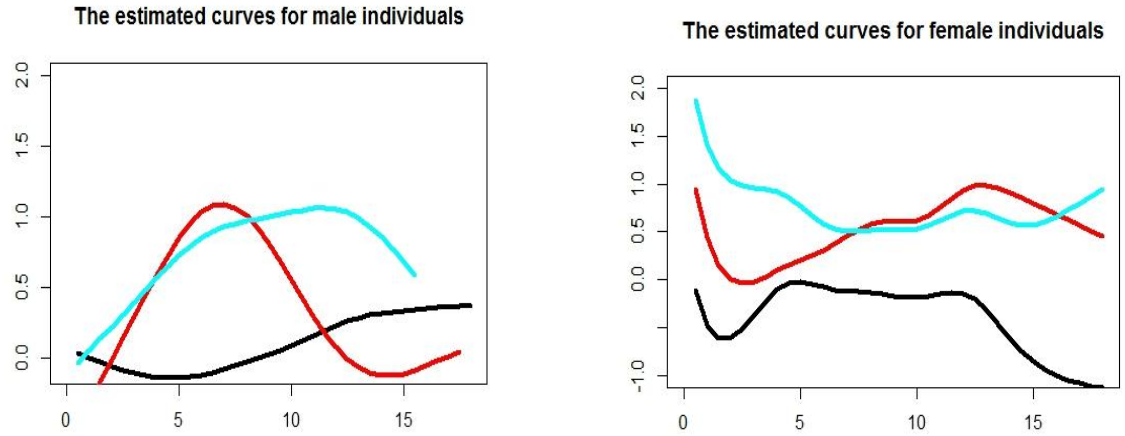

Figure 2: Estimated individual curves for patients with patient ID 2,15,43,53, 63, 82

HYPOTHESIS	RLRT	p-value
$H_0: \sigma^2_{\tilde{u}\cdot} = 0 \ ; \ H_1: \sigma^2_{\tilde{u}\cdot} > 0$	234.8164	< 2.2e-16
$H_0: \sigma^2_w = 0 \ ; \ H_1: \sigma^2_w > 0$	0	1
$H_0: \sigma^2_{\tilde{u}} = 0 \ ; \ H_1: \sigma^2_{\tilde{u}} > 0$	2.92	0.025

Table 1: Likelihood ratio test statistics values for each model testing for the male individuals.

HYPOTHESIS	RLRT	p-value
$H_0: \sigma^2_{\tilde{u}\cdot} = 0 \ ; \ H_1: \sigma^2_{\tilde{u}\cdot} > 0$	299.0287	< 2.2e-16
$H_0: \sigma^2_w = 0 \ ; \ H_1: \sigma^2_w > 0$	3.1845	0.0186
$H_0: \sigma^2_{\tilde{u}} = 0 \ ; \ H_1: \sigma^2_{\tilde{u}} > 0$	33.1793	< 2.2e-16

Table 2: Likelihood ratio test statistics values for each model testing for the female individuals.

BIBLIOGRAPHY

[1] Durbán, M., Harezlak, J., Wand, M.P. and Carroll, R.J. (2005). Simple fitting of subject-specific curves for longitudinal data. Statistics in Medicine, 24, 1153 – 1167.

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

[3] Currie, I.D. and Durbán M. (2002). Flexible smoothing with P-splines: a unified approach. Statistical Modelling, 2, 333 – 349.

[4] Lee, D.-J. and Durbán, M. (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. Statistical Modelling, 11, 49 – 69.