MASTER`S THESIS

# P-SPLINES FOR LONGITUDINAL DATA. APPLICATION TO THE STUDY OF CHILDREN GROWTH WITH PHENYLKETONURIA

IPEK GULER

Supervisors: César Andrés Sánchez Sellero, Carmen Cadarso Suárez

Master in Statistical Techniques

University of Santiago de Compostela

January 2013

El presente documento que tiene como título “P-splines for longitudinal data. Application to the study of children growth with Phenylketonuria” recoge el trabajo realizado por Ipek Guler como Proyecto Fin de Máster de Técnicas Estadísticas bajo la dirección de César Andrés Sánchez Sellero, Carmen Cadarso Suárez.

Fdo.: César Andrés Sánchez Sellero Fdo.: Carmen Cadarso Suárez

Fdo.: Ipek Guler

# 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.

# Acknowledgements

This study has been realized during my internship in Epidemiology Unit of Santiago de Compostela University Hospital.  I would like to express my appreciation to all those who provided me the possibility to complete this study. First of all special gratitude I give to my supervisors; to Carmen Cadarso Suárez for her guidance with my internship in Epidemiology Unit where I had a special experience for my future and also for all her advices and comments during my study, to César Andrés Sánchez Sellero for his useful comments, advices and support.

Furthermore, I would also like to give my special thanks to Francisco Gude Sampedro and Maria Jose Rodriguez Álvarez for their unique supervision during my internship and for all their sincere support, advices and comments on my master thesis. I also want to appreciate the comments and advices given by Maria Durban and Thomas Kneib with whose I had the chance to discuss my work.

Last but not least, I give my very special thanks to my family for their support from far away, and to my husband Borja F. Caamaño to make everything possible in my life.

## CONTENTS

# Chapter 1

# Introduction

## 1.1. Motivating study : Impact of Phenylketonuria in the growth of children

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.

While in HPA Benigna isn`t necessary to realize 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 are treated with a special diet and in case of having this treatment they have a favorable evolution.

The objective of this study is to model and to compare the 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 dataset contains observations from 109 children, followed in the center from 6 months up to 18 years from successive controls that were realized during the consultation and from the revision of the clinical histories of the patients.

The patients were classified depending on the level of their PAH, divided in three categories: Classical PKU(PAH > 1200 mol/L), Moderated PKU(PAH 360-1200 mol/L), HPA Benigna(PAH 120-360 mol/L). In the first two groups the patients follow a dietetic treatment restricted in natural proteins and the third one follow a normal diet. For the purpose of this study we compared the group of individuals with PKU (either classical or moderated) and the group HPA, in order to evaluate the effect of the pathology (HPA and PKU) on the growth of the children.

To model the growth curves of the children measured over time we need to use some flexible regression techniques to handle possible non linear effects.

In this study we used a flexible mixed model (Durban, 2005) to study the growth of children based on the representation of penalized-splines (Eilers and Marx, 1996) as mixed models (Currie,Durban, 2002). Also it has been applied subject specific curves for each child and factor by curve interactions to detect if there is any difference between the two types of pathologies.

## 1.2. Data Source

For each child included in the srudy, we have the following measurements;

***Height***: Height (kg) of the children measured every six months from 6 months to 18 years.

***Z-score***:  Measure that express the distance between an individual child's height and the average height of comparable children in the reference population.

***Age:*** Age of the children (6 months up to 18 years) when the measures were taken.

***Type of pathology***: Categorical covariate that indicates the type of pathology of the children (either HPA or PKU).

***Gender:*** Gender of the children (Female-Male).

## 1.3. Longitudinal study

Longitudinal studies are mostly designed to investigate changes in a characteristic during a period of time repeatedly for each study participant. Multiple measurements are obtained from each individual at different times and possibly under changing the experimental conditions. We cannot fully control all the conditions when the measurements are taken and so there may be considerable variation among individuals. This is mostly analyzed using some variant of a two-stage model **(Laird and Ware, 1982).** In this formulation we have fixed and random effects. The mixed effects models **(Pinheiro and Bates, 2000)** allow modeling and analysis of between and within individual variation.

However, in some circumstances in the analysis of longitudinal data, the parametric assumption in linear mixed models may not always be appropriate. It could be erroneous using traditional parametric regression techniques to model the growth curves measured over time as in our application study. So that it is fundamental to use flexible techniques such as smoothing splines **(Ruppert et al, 2003).**

For independent data, there is a rich literature on kernel and spline methods for nonparametric and semiparametric regression **(Green and Silverman, 1994; Speckman, 1988)**. **Wahba (1978)** proposed a Bayesian model for spline smoothing. However, only limited work has been done on nonparametric and semiparametric regression for correlated data.

There are several considered different types of smoothers (kernel, smoothing spline etc…) but they modeled the random effects by parametric functions. **Zhang et al** considered semiparametric models accounting for within subject correlation but they didn`t consider smooth curves for individual subjects. **Brumback and Rise (1998)** modeled both population mean and subject specific curves non-parametrically with smoothing splines and used their mixed model representation. However they ran into computational problems because they assumed fixed slopes and intercepts for the subject-specific curves. **Rice and Wu (2001)** partially solved this problem by modeling individual curves as spline functions with random coefficients. However, in their low-rank spline basis approach the number and location of the knots used to construct the basis became an important issue. Consequently, the fit of their models involved the use of some selection criteria to choose these parameters. More recently **Guo (2002)**, took a functional data analysis approach by introducing functional random effects which are modeled as realizations of a zero-mean stochastic process. He also used the connection between smoothing splines and mixed models for fitting and estimating this model, However Guo also faced computational problems due to large matrices (since smoothing splines use as many knots as data points) and consequently developed a sequential estimation procedure using Kalman filtering.

**Durban et al, (2005)** proposed an approach which is a trade-off between spline regression and smoothing splines. The equivalence between a penalized smoother and the optimal predictor in a mixed model is used to present an unified approach for model estimation. The penalty approach relaxes the importance of the number and location of knots and the use of a low-rank smoother solves the computational problems of other approaches when analyzing large data sets. In this study this approach is applied to model the growth of children using the penalized smoothing splines represented as mixed models.

The primary focus in designed experiments and longitudinal studies in medicine usually involves treatment comparisons, possibly factorial in nature. If a quantitative variable impacts on treatments, e.g. age in the longitudinal setting, the interaction of treatments with the quantitative variable is generally of interest. This study illustrate such situations to compare the two different pathological types in the growth of the children building and comparing a wide range of additive mixed models, from a simple additive mixed model up to an additive mixed model with subject specific curves and factor by curveinteractions.

# Chapter 2

# Mixed models for longitudinal data

Linear mixed models is one of the methodologies for analysis of the longitudinal data. Some of the advantages of this methodology is that it simplifies the complexity of typical longitudinal datasets and the existentce of a widely available software developments for this methodology.

The linear mixed models methodology also will be used in the following chapters to represent p-spline regression models. Therefore in this chapter we present a brief introduction to linear mixed models (**Laird and Ware , 1982)**, the covariance structures and the estimation of the fixed and random effects.

## 2.1 Fixed and Random Effects

The mixed model extends the fixed effects model by including random effects, random coefficients and/or covariance terms in the residual variance matrix. In longitudinal study the factors levels are randomly selected from a population of all possible factor levels. In our study the patients are the random factor levels. If we have also some fixed factors with the random factors we call the model as a mixed-effect model.

For a fixed-effects factor, we assume that there is a finite set of levels that contains all levels of interest for the study. For a random-effects factor we assume that there is an infinite set of levels (a population of levels) and we think about the levels present in the study as a sample from that population. The interest is in drawing inferences that are valid for the complete population of levels. As a simple example we could think about the patients in our application study. The appropriate model for our data would be a mixed-effect model with type of pathology as a fixed effect factor and patients as random-effect factor.

## 2.2 A two-stage analysis

In some researches, the data has repeated measures over time. For example, in studying the growth of the children, as in our application study, their height and z-scores are measured yearly for some specified number of years and we are interested the rate of change of this variable over age. One of the methods that can be used in this context can be a linear regression. The response, height or z.score of the children can be assumed to have a constant rate of change over age. This assumption may be unreasonable and variations on the basic linear regression method would be needed. Moreover a limitation of the usual linear regression model in this situation is that it ignores the fact that observations on the same subject are dependent.

If the quantities of interest are the average heights at each age point (without assuming constant rate of change), a repeated measures analysis of variance (ANOVA) may be used. Repeated measures ANOVA, under certain restrictive assumptions, does account for the within-subject dependence in the data. However, linear regression and ANOVA provide estimates of the population average of the rate of change or the means. No measure of between subject variability is given. Measurements on the same subject are much more similar than the measurements on different subjects (this is what is meant by "within-subject dependence"). Both simple linear model and the ANOVA method ignore this fact and average over subjects to obtain estimates.

In the linear mixed model extension **(Laird and Ware, 1982 ; Harville, 1977**) the repeated measurements are modeled using a linear regression model, with parameters which are allowed to vary over individuals, and which are therefore called random effects or subject-specific regression coefficients. The subject-specific regression parameters reflect the natural heterogeneity in the population and they are usually assumed to follow a Gaussian distribution.

Often, subject-specific longitudinal profiles can be well approximated by linear regression functions which lead to two-stage analyses **(Laird and Ware, 1982).** In this formulation, the probability distribution for the multiple measurements has the same form for each individual, but the parameters of that distribution vary over individuals. The distribution of these parameters, or “random effects” constitutes in the second stage of the model.

Such two-stage models have several desirable features;

- There is no requirement for balance data.

- They allow explicit modeling and analyses of between and within individual variation.

- They facilitate the study of the effects of background variables on the response of interest.

### Two-stage model formulation

Two stage random-effects models are based on explicit identification of individual and population characteristics, and their form extends to the unbalanced situation. Most of the two stage models in the literature can be described either as growth models or repeated measures models**.**

In this section we will introduce these two stage analysis formulation **(Laird and Ware, 1982).** Population parameters, individual effects and within-person variation are introduced in Stage 1, and between-person variation at Stage 2.

Before presenting the main ideas behind the two stage model, we introduce some notation;

- Let $\boldsymbol{y_i}$ be the response vector (*$n_i$x1*) where *$n_i$* is the number of observations in the *ith* individual *i=1,….N.*

- Let $\boldsymbol{\beta}$ denote a px1 vector of unknown population parameters and $\mathbf{X_i}$ be a known *$n_i$xp* design matrix linking $\boldsymbol{\beta}$ to $\boldsymbol{y_i}$.

- Let $\boldsymbol{u_i}$ denote a kx1 vector of unknown individual effects and $\mathbf{Z_i}$ a known *$n_i$xk* design matrix linking $\boldsymbol{u_i}$ to $\boldsymbol{y_i}$**.**

**Laird and Ware, (1982)** proposed the following model;

**Stage 1**

For each individual unit it is assumed that

$$\boldsymbol{y_i = X_i\beta + Z_iu_i + \varepsilon_i}$$

where $\boldsymbol{\varepsilon_i}$ is distributed as **N(0,R$_i$)** (normal with mean 0 and covariance matrix **R$_i$**.) Here **R$_i$** is an *$n_i$x$n_i$* positive definitive covariance matrix; it depends on *i* through its dimension *$n_i$* , but the set of unknown parameters in **R$_i$** will not depend upon *$n_i$*. At this stage, $\boldsymbol{\beta}$ and $\boldsymbol{u_i}$ are considered fixed, and the $\boldsymbol{\varepsilon_i}$ are assumed to be independent. It should be noted that, in this stage, $\boldsymbol{\beta}$ represents the population parameters and $\boldsymbol{u_i}$ $(\boldsymbol{i = 1, \ldots N})$ represents the individual effects.

**Stage 2**

In the second stage the $\boldsymbol{u_i}$ are assumed to be **N(0,G$_i$)** , independently of each other and of the $\boldsymbol{\varepsilon_i}$**.** Here **G$_i$** is a *kxk* positive definitive covariance matrix. The population parameters, $\boldsymbol{\beta}$**,** are treated as fixed effects.

The regression model defined in the second stage receives the name of Linear Mixed Model. Specifically this model can be expressed in matrix notation as;

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \qquad (\mathbf{2.1})$$

where;

- $\boldsymbol{y}$ is the vector of the observed responses $\begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_N \end{pmatrix} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \end{pmatrix}$

- $\boldsymbol{\beta}$ is the fixed effects parameters vector,

- $\boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \\ \vdots \\ \boldsymbol{\varepsilon_N} \end{pmatrix}$ is the residual vector

- $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z_1} & \dots & \mathbf{0} \\ \vdots & \dots & \vdots \\ \mathbf{0} & \dots & \boldsymbol{Z_N} \end{bmatrix}$

- $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X_1} \\ \vdots \\ \boldsymbol{X_N} \end{bmatrix}$

- $\boldsymbol{u} = \begin{bmatrix} \boldsymbol{u_1} \\ \vdots \\ \boldsymbol{u_N} \end{bmatrix}$

## 2.3 Covariance Matrix V

Marginally the covariance of $\boldsymbol{y}$, var($\boldsymbol{y}$)= $\boldsymbol{V}$, can be written as;

$$\boldsymbol{V} = \boldsymbol{var}(\boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}).$$

Since we assume that the random effects and the residuals are uncorrelated.

$$\boldsymbol{V} = \boldsymbol{var}(\boldsymbol{X\beta}) + \boldsymbol{var}(\boldsymbol{Zu}) + \boldsymbol{var}(\boldsymbol{\varepsilon})$$

Since $\boldsymbol{\beta}$ describes the fixed effects parameters, $\boldsymbol{var}(\boldsymbol{X\beta})$ **= 0.** Also, Z is a matrix of constants. Therefore;

$$\boldsymbol{V} = \boldsymbol{Zvar}(\boldsymbol{u})\boldsymbol{Z}' + \boldsymbol{var}(\boldsymbol{\varepsilon}).$$

If we let **G** denote **var($\boldsymbol{u}$)** , and since the random effects are assumed to follow normal distribution, $\boldsymbol{u}$ **~N(0,G)** and **var($\boldsymbol{\varepsilon}$)=R**. We have,

$$\boldsymbol{V} = \boldsymbol{ZGZ}' + \boldsymbol{R}$$

where

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G_1} & \dots & \boldsymbol{0} \\ \vdots & \dots & \vdots \\ \boldsymbol{0} & \dots & \boldsymbol{G_N} \end{bmatrix} \qquad \boldsymbol{R} = \begin{bmatrix} \boldsymbol{R_1} & \dots & \boldsymbol{0} \\ \vdots & \dots & \vdots \\ \boldsymbol{0} & \dots & \boldsymbol{R_N} \end{bmatrix}$$

## 2.4 The random effects model covariance structure

**The G Matrix**

**G$_i$** is a matrix with dimension *kxk* which is the dimension of random effects parameters. The **G$_i$** matrix has different forms depends on the model structure.

Usually the following structure is assumed;

$$\boldsymbol{G_i} = \begin{bmatrix} \boldsymbol{\sigma^2_{u_1}} & \dots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \dots & \boldsymbol{\sigma^2_{u_k}} \end{bmatrix}$$

where the $\boldsymbol{\sigma^2_{u_j}}$ *(j=1,..k)* is the variance of *jth* random effect parameter in $\boldsymbol{u_i}$.

**The R Matrix**

In linear mixed models the residuals are usually assumed to be uncorrelated, therefore we have a diagonal residual matrix **R**,

$$\boldsymbol{R_i = \sigma^2} \mathbf{I}_{\boldsymbol{n_i x n_i}}$$

$$\boldsymbol{R_i} = \begin{bmatrix} \boldsymbol{\sigma^2} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \boldsymbol{\sigma^2} \end{bmatrix}$$

**The covariance pattern model covariance structure**

Covariance patterns can be fitted into the linear mixed model definition using matrix notation. In covariance pattern models, the covariance structure of the data is defined by specifying a pattern for the covariance terms directly into **R**. Observations within a chosen blocking variable (e.g. patients) are allowed to be correlated and a pattern for their covariances is specified. This pattern is usually chosen to depend on a variable such as age. In such cases, **R** will have a block diagonal form and can be written as;

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R_1} & \dots & \mathbf{0} \\ \vdots & \boldsymbol{R_2} & \vdots \\ \mathbf{0} & \dots & \ddots \end{bmatrix}$$

The sub matrices $\boldsymbol{R_i}$**,** are covariance blocks corresponding to the *ith* blocking effect (for e.g. *ith* patient). They have dimension equal to the number of repeated measurements on each patient. The 0 represents the matrix blocks of zeros giving zero covariances for observations on different patients.

$\boldsymbol{R}$ matrix can have different forms as simple structure, compound symmetric structure, autoregressive structure depending on the correlation between the observations in each subject **(Brown and Prescott , 1962).**

## 2.5 Random Intercept and slope models

In this section we present two of the most important mixed effect models which are called the random intercept model and the random intercept and slope model. The random intercept model allows a random shift around the intercept resulting in "fitted" individual lines parallel to the population fitted line, whereas the random intercept and slope model also allows a change in the slopes of the fitted lines.

For the sake of illustration we consider here our application study, where we have only one continuous covariate (age), and the objective is to model the growth of the individuals along the time of the study.

Specifically, in this case, the random intercept model takes the form;

$$\boldsymbol{y_i = \beta_0 + \beta_1 t_i + u_i + \varepsilon_i} \qquad (\mathbf{2.2})$$

where $\boldsymbol{t_i} = (\boldsymbol{t_{i1}}, \dots, \boldsymbol{t_{in_i}})'$ denotes the ages at which the measurements were taken for individual *i* and $\boldsymbol{y_i}$ is the z-score of individual *i* at these ages.

Accordingly in this case,

$$\boldsymbol{X_i} = \begin{bmatrix} \mathbf{1} & \boldsymbol{t_{i1}} \\ \vdots & \vdots \\ \mathbf{1} & \boldsymbol{t_{in_i}} \end{bmatrix} \qquad \boldsymbol{Z_i} = \begin{bmatrix} \mathbf{1} \\ \vdots \\ \mathbf{1} \end{bmatrix} \qquad \boldsymbol{u_i} \sim \boldsymbol{N(0, \sigma_u^2)}$$

On the other hand the random intercept and slope model adopts this form;

$$\boldsymbol{y_i = \beta_0 + \beta_1 t_i + u_{i1} + u_{i2} t_i + \varepsilon_i} \qquad (\mathbf{2.3})$$

In this case the model matrices $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ have the form;

$$\boldsymbol{X_i} = \begin{bmatrix} \mathbf{1} & \boldsymbol{t_{i1}} \\ \vdots & \vdots \\ \mathbf{1} & \boldsymbol{t_{in_i}} \end{bmatrix} \qquad \boldsymbol{Z_i} = \begin{bmatrix} \mathbf{1} & \boldsymbol{t_{i1}} \\ \vdots & \vdots \\ \mathbf{1} & \boldsymbol{t_{in_i}} \end{bmatrix}$$

$$\boldsymbol{u_i} = (\boldsymbol{u_{i1}}, \boldsymbol{u_{i2}}) \sim \boldsymbol{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\sigma^2_{u_1}} & \boldsymbol{\rho} \\ \boldsymbol{\rho} & \boldsymbol{\sigma^2_{u_2}} \end{pmatrix}\right) \text{ where } \boldsymbol{\rho} = \boldsymbol{cov}(\boldsymbol{u_{i1}}, \boldsymbol{u_{i2}})$$

As can be observed, the only difference between the random intercept and the random intercept and slope model is the modification in matrix Z, which permits a slope variation between the subjects.

## 2.6 Model fitting methods

In this section the methodology to numerical methods for fitting mixed models is given. The model fitting process has three distinctive components: estimating fixed effects, estimating random effects, and estimating variance parameters.

### 2.6.1 The likelihood function

The mixed model can be fitted by maximizing the likelihood function. The likelihood function, L, measures the likelihood of the model parameters given the data and is defined using the density function of the observations. In models where the observations are assumed independent (e.g. fixed effects models), the likelihood function is simply the product of the density functions for each observations. However, observations in a mixed model are not independent and the likelihood function therefore needs to be based on a multivariate normal distribution for $\boldsymbol{y}$**.** As random effects have expected values of zero and therefore do not affect the mean, this distribution has a mean vector $\boldsymbol{X\beta}$ and a covariance matrix **V** which depends on several variance parameters. The likelihood function based on the multivariate normal density function is then

$$\boldsymbol{L}(\boldsymbol{y}) = \frac{\boldsymbol{exp}\left[-\frac{\mathbf{1}}{\mathbf{2}}(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{V}^{-\mathbf{1}}(\boldsymbol{y} - \boldsymbol{X\beta})\right]}{(\mathbf{2}\boldsymbol{\pi})^{(\mathbf{1}/\mathbf{2})\boldsymbol{n}}|\boldsymbol{V}|^{\mathbf{1}/\mathbf{2}}}$$

In practice, the log likelihood function is usually used in place of the likelihood function since it is simpler to work with and its maximum value coincides with that of the likelihood. The log likelihood is given by ; **(Brown and Prescott , 1962)**

$$\boldsymbol{log(L(y)) = K - \frac{1}{2}[log|V| + (y - X\beta)'V^{-1}(y - X\beta)]} \qquad (\mathbf{2.4})$$

where

$\boldsymbol{K = -\frac{1}{2}nlog(2\pi)}$ **(**a constant can be ignored in the maximization process).

$\boldsymbol{n}$= number of observations.

### 2.6.2 Estimation of Fixed Effects

The estimates of fixed effects parameters can be obtained by maximizing the likelihood by differentiating the log likelihood **(2.4)** with respect to $\boldsymbol{\beta}$ and setting the resulting expression to zero (see **Brown and Prescott, 1962**).

This leads to;

$$\boldsymbol{X'V^{-1}(y - X\beta) = 0}.$$

Rearranging,

$$\boldsymbol{\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y} \qquad (\mathbf{2.5})$$

and the variance of $\boldsymbol{\hat{\beta}}$ is obtained as,

$$\boldsymbol{var(\hat{\beta}) = (X'V^{-1}X)^{-1}}.$$

### 2.6.3 Estimation (or prediction) of random effects coefficients

In general, random effects coefficients are defined to have normal distributions with zero means and the specific values they take must be thought of as realizations of a sample from a distribution. Thus their expected values are, by definition, zero. But it is possible to obtain predictions of them.

The prediction of the $\boldsymbol{u}$ is (see **Brown and Prescott, 1962**),

$$\widehat{\boldsymbol{u}} = (\boldsymbol{Z}'\boldsymbol{R}^{-\mathbf{1}}\boldsymbol{Z} + \boldsymbol{G}^{-\mathbf{1}})^{-\mathbf{1}}\boldsymbol{Z}'\boldsymbol{R}^{-\mathbf{1}}(\boldsymbol{y} - \boldsymbol{X\beta})$$

In random effects models the **R** matrix is diagonal, **R**=$\boldsymbol{\sigma^2}$**,** so we can write alternatively;

$$\widehat{\boldsymbol{u}} = (\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G}^{-\mathbf{1}}/\boldsymbol{\sigma}^{\mathbf{2}})^{-\mathbf{1}}\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X\beta})$$

Also recalling the equation

$$\boldsymbol{V} = \boldsymbol{ZGZ}' + \boldsymbol{R}$$

$$\widehat{\boldsymbol{u}} = \boldsymbol{GZ}'\boldsymbol{V}^{-\mathbf{1}}(\boldsymbol{y} - \boldsymbol{X\beta}) \qquad (\mathbf{2.6})$$

And the variance of $\widehat{\boldsymbol{u}}$

$$\boldsymbol{var}(\widehat{\boldsymbol{u}}) = \boldsymbol{GZ}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{ZG} - \boldsymbol{GZ}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{ZG}$$

It should be noted that in order to estimate $\boldsymbol{\beta}$ and $\boldsymbol{u}$ (see equations **(2.5)** and **(2.6)**) we need an estimate of $\boldsymbol{V}$ and therefore of the covariance components involve in its definition denoted by $\boldsymbol{\gamma}$. There are several approaches to estimate $\boldsymbol{V}(\boldsymbol{\gamma})$ which are all based (directly and indirectly) on maximizing the likelihood function.

## 2.7 Estimation of Variance Components

### 2.7.1 Maximum likelihood (ML) and Restricted Maximum likelihood Estimation (REML)

This method is based on the concept of maximizing the log likelihood **(2.4)** with respect to the variance parameters while treating the fixed effects, $\boldsymbol{\beta},$ as constants. This approach would yield an estimator for the covariance matrix of $\boldsymbol{u_i}$ which would take into account the extra variability due to the estimation of $\boldsymbol{\beta}.$

Restricted maximum likelihood estimation was first suggested by **Patterson and Thompson (1971).** In this approach the parameter $\boldsymbol{\beta}$ is eliminated from the log likelihood so that it is defined only in terms of the variance parameters.

First we obtain a likelihood function based on the residual terms, $\boldsymbol{y - X\widehat{\beta}}.$ This contrasts with the likelihood which is based directly on the observations, $\boldsymbol{y}.$ We notice that these residuals differ from the ordinary residuals which has to be $\boldsymbol{e = y - X\widehat{\beta} - Z\widehat{u}},$ in that $\boldsymbol{Z\widehat{u}}$ is not deducted.

**Brown and Prescott (1962)** explained that also referring to the $\boldsymbol{y - X\widehat{\beta}}$ as residuals is not unreasonable since they can be considered as error terms that include sources of random variation (residual and random effects). They referred to $\boldsymbol{y - X\widehat{\beta}}$ as the full residuals in order to differentiate them from the ordinary residuals.

The REML described by **Brown and Prescott (1962)** as;

$$\boldsymbol{L(\gamma; y - X\widehat{\beta})} \propto \left|\boldsymbol{X'V^{-1}X}\right|^{-\frac{1}{2}} |\boldsymbol{V}|^{-\frac{1}{2}}\, \boldsymbol{exp}\left(\frac{1}{2}(\boldsymbol{\widehat{\beta} - \beta})'\, \boldsymbol{XV^{-1}X}(\boldsymbol{\widehat{\beta} - \beta})\right)$$

Accordingly the REML log likelihood is described as;

$$\boldsymbol{log\big(L(\gamma; y - X\widehat{\beta})\big) = K - \frac{1}{2}\Big[log|V| - log\left|X'V^{-1}X\right|^{-1} + \big(y - X\widehat{\beta}\big)'V^{-1}(y - X\widehat{\beta})\Big]}$$

where the $\boldsymbol{\gamma}$ are the variance parameters.

We see that the differences between the REML log likelihood and the ordinary log likelihood (**2.4)** is caused by the extra term $\boldsymbol{log\left|X'V^{-1}X\right|^{-1}}$ which is the log determinant of $\boldsymbol{var(\hat{\beta})}$**.**

REML is sometimes referred to as a “marginal method” because it takes account of the fact that $\boldsymbol{\beta}$ is a parameter and not a constant, therefore the resulting variance parameter estimates are unbiased.

### 2.7.2 Comparasion between ML and REML

Maximum likelihood estimation and restricted maximum likelihood estimation both have the same merits of being based on the likelihood principle which leads to useful properties such as;

- Consistency
- Asymptotic normality
- Efficiency

ML estimation also provides estimators of the fixed effects, while REML estimations, on itself, does not. On the other hand, for balanced mixed models, the REML estimates for the variance components are identical to classical ANOVA-type estimates obtained from solving the equations which set mean squares equal to their expectations, which have optimal minimum variance properties and which do not rely on any normality assumption since only moment assumptions are involved **(Veerbeke and Molenberghs, 1997).**

## 2.8 Model testing

It is shown in section 2.7.1 that the parameters in mixed models can be estimated by maximum likelihood. **Ruppert et. al (2003),** explained that the likelihood ratio procedure also can be used for model testing.

Consider that we want to test the following hypothesis;

$$\boldsymbol{H_0 : \beta_i = 0 \qquad versus \qquad H_1 : \beta_i \neq 0}$$

where the $\boldsymbol{\beta_i}$ is the *i*th entry of $\boldsymbol{\beta}$.

Let $\boldsymbol{L_0}(\boldsymbol{y})$ be the likelihood function under the null hypothesis and $\boldsymbol{L_1}(\boldsymbol{y})$ under the alternative hypothesis. The likelihood ratio statistic for testing a null restricted model against an alternative restricted model is,

$$\boldsymbol{LR}(\boldsymbol{y}) = \boldsymbol{L_0}(\boldsymbol{y})/\boldsymbol{L_1}(\boldsymbol{y})$$

It is more common to work with log likelihood functions,

$$-\boldsymbol{2log}\big(\boldsymbol{LR}(\boldsymbol{y})\big) = -2\{\boldsymbol{log}(\boldsymbol{L_0}(\boldsymbol{y})) - \boldsymbol{log}(\boldsymbol{L_1}(\boldsymbol{y}))\}$$

The classical result for determining the significance of the observed value of $\boldsymbol{LR}(\boldsymbol{y})$ is one that states, under $\boldsymbol{H_0}$ (see **Ruppert et al, 2003**)

$$-\boldsymbol{2log}\{\boldsymbol{LR}(\boldsymbol{y})\} \sim \boldsymbol{\chi^2_v} \qquad (\mathbf{2.7})$$

Where the right hand side is the chi-squared distribution with $\boldsymbol{v}$ degrees of freedom and

$$\begin{aligned}\boldsymbol{v} = \boldsymbol{number\ of\ independent\ parameters\ in\ unrestricted\ model}\\ -\ \boldsymbol{number\ of\ independent\ parameters\ in\ null\ model}\end{aligned}$$

where the $\boldsymbol{v} = \mathbf{1}$ for the hypothesis test $\boldsymbol{H_0} : \boldsymbol{\beta_i} = \mathbf{0}\ \ \boldsymbol{versus}\ \ \boldsymbol{H_1} : \boldsymbol{\beta_i} \neq \mathbf{0}.$

In some circumstances we are also interested in hypothesis test for the covariance matrix parameters to investigate whether the corresponding random effects are necessary for the model. For instance, in the random intercept model **(2.2)** we could be interested in the following hypothesis to test if the variance parameter of the random intercept is zero therefore it would be not appropriate for the model;

$$\boldsymbol{H_0}\text{: } \boldsymbol{\sigma^2_u} = \mathbf{0} \qquad \boldsymbol{versus} \qquad \boldsymbol{H_1}\text{: } \boldsymbol{\sigma^2_u} > 0 \qquad (\mathbf{2.8})$$

For this hypothesis the approximation **(2.7)** may not be appropriate because it assumes that the parameter of interest is not on the boundary of its parameter space. Since the parameter space for $\boldsymbol{\sigma^2_u}$ is [0,∞), this assumption is violated. **(Ruppert et al , 2003)**

**Stram and Lee (1994)** proved that the likelihood ratio test L for testing this hypothesis **(2.8)** has a $\boldsymbol{\chi_s^2} + \frac{1}{2}\boldsymbol{\chi_{s+1}^2}$ asymptotic distribution where s is the number of fixed effects parameters constrained under the null hypothesis. For the null hypothesis in **(2.8),** s=0 and therefore,

$$\boldsymbol{-2\log\{LR(y)\} \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2}$$

This means that the log likelihood function has an approximate density function equal to a 50:50 mixture of the $\boldsymbol{\chi_0^2}$ and $\boldsymbol{\chi_1^2}$ densities.

Furthermore **Fabian Scheipl and Ben Bolker (2013)** have realized some simulations of restricted log likelihood value and implemented to R-project. ( Package "RLRsim"). We have used this package and *exactRLRT()* function to compare two models with more than one random effects.

# Chapter 3

# P-splines

In Chapter 2 we have presented the linear mixed model methodology to afterwards represent the p-spline regressions using these techniques.

In our application part, to model the growth curves of the children we have to adopt non parametric model to allow sufficient flexibility. For that reason, in this section an introduction for P-splines and the representation of a P-splines regression models as a mixed model are given to obtain more flexible regression models. **(Brumback et al., 1999; Currie and Durban, 2002).**

## 3.1 Introduction to P-splines

In statistic research the smoothing term is frequently used in many areas and two factors made these smoothing techniques to become recently so popular. One of them is the increasing complexity of the data that is used and the advances in computing that has facilitated these kinds of models reducing significantly the computational cost.

There are several ways of smoothing as regression splines, smoothing splines or penalized smoothing splines **(Green and Silverman, 1994**). Splines is a method of smoothing which are polynomials pieces that are joined at points, called “knots”.In regression splines the smoothing function depends on the election of knots and this can be complicated to compute in multidimensional situations. In smoothing splines the number of knots are equal to the number of observations which can be also a problem when we have large datasets.

Splines with penalizations or P-Splines **(Eilers and Marx, 1996)** combine the best part of these two methods: it uses less parameters than smoothing splines and the selection of the knots are not so determinant like in regression splines. They are low range splines in which the number of knots are much lower than the dimension of the data. Also P-Splines relax the importance of the localization and the number of knots. Finally the representation of a P-Spline regression model as mixed models permits, in some case, the use of the methodology of the mixed models.

Consider a flexible regression model

$$\boldsymbol{y_i} = \boldsymbol{f}(x_i) + \ \boldsymbol{\varepsilon_i} \qquad \boldsymbol{\varepsilon_i} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma^2}), \qquad \boldsymbol{i} = \boldsymbol{1}, \dots \boldsymbol{N} \qquad\qquad (\boldsymbol{3.1})$$

where $\boldsymbol{y_i}$ is the response variable for the observations $\boldsymbol{i} = \boldsymbol{1}, \dots \boldsymbol{N}$ and $\boldsymbol{f}(.)$ is smooth function of covariate $x$**.**

To estimate the function $\boldsymbol{f}(.)$ it is assumed that this function can be represented by a linear combination of **d** known basis functions $\boldsymbol{B_j}$**.** Therefore

$$\boldsymbol{f}(x) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{d}} \boldsymbol{B_j}\,(x)\boldsymbol{\theta_j}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \dots \boldsymbol{\theta_d})'$ is a vector of unknown regression coefficients. Under this representation of model **(3.1)** is parametric and it can be easily estimated using ordinary least squares;

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{B'B})^{-1}\boldsymbol{B'y}$$

where

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B_1}(x_1) & \dots & \boldsymbol{B_d}(x_1) \\ \vdots & \ddots & \vdots \\ \boldsymbol{B_1}(x_N) & \dots & \boldsymbol{B_d}(x_N) \end{pmatrix}$$

and $\boldsymbol{y} = (\boldsymbol{y_1}, \dots, \boldsymbol{y_n})$ **.**

## 3.2 Basis and knots

There are several alternatives for the choice of the basis functions $\boldsymbol{B_j}$ such as; truncated polynomials, thin plate splines **(Wood, 2003)** and B-splines **(de Boor, 2001).**

Truncated polynomial bases are useful for understanding the mechanics of spline-based regression and they can be used if the knots are selected carefully or a penalized fit is used **(Durban et al , 2005).** However, the truncated power bases have the disadvantage that they are far from orthogonal. Therefore it is more recommendable to use equivalent bases with more table numerical properties. This study is focused on
B-splines basis as a recommendation of **Durban et al (2005)** who indicate that these basis are more stable basis for P-splines.

With respect of the number of knots, in most of the situations, the suggestion is to use a moderately large number of equally-spaced knots. The first goal for any algorithm for selecting number of knots (K) is to make certain that K is sufficiently large to fit the data. The second goal is to choose K not so large that computation time is excessive **(Durban et al , 2005).** As regards this issue, **Ruppert et. al (2003)** advise the following equation for the selection of the number of knots ,

$$\boldsymbol{number\ of\ knots = min\{40, the\ unique\ values\ of\ x/4\}}$$

## 3.3. Penalizations

In regression spline, the number and also the location of the knots have a great impact on the final estimates. When the number of knots increases the estimated curve (in comparison with the true curve ) becomes too wiggly, meaning that the data are over fitted**.** The proposed methods for the selection of knots **(Fried and Silverman, 1989**) have the disadvantage of being computationally intensive. Therefore, to solve this problem, **O`Sullivan (1986)** introduced the idea of penalized splines, where a smoothness penalty is added to the least squares criterion when estimating the regression coefficients $\boldsymbol{\beta}$**.**

For the sake of simplicity, we first present the idea of the penalization based on the second derivative function of $\boldsymbol{f}(.)$. In this case of penalized splines, model **(3.1)** is fitted by minimizing the penalized sum of squares:

$$\boldsymbol{(y - B\theta)'(y - B\theta) + \lambda \int f''(x)^2\, d_x} \qquad \boldsymbol{(3.2)}$$

where $\boldsymbol{\lambda}$ is the smoothing parameter which controls the trade-off between fidelity to the data and roughness of the function estimate.

Later **Eilers and Marx (1996)** proposed a penalization based on the q order differences between the coefficients of the B-splines basis functions which is a more flexible method and independent from the degree of the polynomial that is used to construct the B-splines. In this approach the penalty becomes $\boldsymbol{\lambda\theta' P\theta}$ where $\boldsymbol{P = D'D,}$ and $\boldsymbol{D}$ is a known *(d-q)xd* matrix whose elements depend on the chosen q-order difference.

For example for a second order penalty, the D matrix has the form;

$$\boldsymbol{D} = \begin{bmatrix} \mathbf{1} & \mathbf{-2} & \mathbf{1} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{1} & \mathbf{-2} & \mathbf{1} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{-2} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

As can be observed, this penalization is discrete and penalizes the coefficients instead of the whole curve as in $(\mathbf{3.2})$ which reduces the dimensional problem.

## 3.4. Smoothing parameter $\boldsymbol{\lambda}$ selection

The critical point of the penalized spline smoothing is the choice of the smoothing parameter $\boldsymbol{\lambda}$. If a large smoothing parameter is chosen the resulting curve is very smooth, but if we choose a small smoothing parameter the resulting estimate becomes too wiggly. There are several methods to choose the optimal value of $\boldsymbol{\lambda}$ such as Generalized Cross Validation **(Hastie and Tibshirani, 1990 ; Wood, 2006)** or Akaike Information Criterion **(Wood, 2008)**. A different approach to choose the optimal smoothing parameter $\boldsymbol{\lambda}$ comes from the fact that a P-spline regression model can be reformulated as a linear mixed model **(Brumback et al, 1999 ; Currie and Durban, 2002).** In this study we will focus on this later approach.

## 3.5. Penalized Splines Mixed Model Representation

The representation of penalized splines as mixed models has several advantages in estimating a semi parametric or non parametric models **(Brumback et al., 1999; Currie and Durban, 2002).**

- It allows to use the methodology and the several softwares for mixed models.

- It is a model-based approach to smoothing that uses two basic principles of statistics: Maximum likelihood and best prediction. The incorporation of likelihood based models for complications such as dependence, measurement error, and missing data is more straightforward.

- It comes equipped with an automatic smoothing parameter choice that corresponds to maximum likelihood and/or restricted maximum likelihood estimation of variance components. Their availability in software packages makes ML and REML smoothing parameter selection quite attractive.

The interest of this representation is coming from the difficulties of identification of an additive model. P-Splines as mixed models modify the basis which can be decomposed as a sum of a polynomial component and a non-polynomial component.

One of the most attractive parts of this representation is that the smoothing parameter, becomes the ratio between the variance of residuals and the variance of the random effects $\boldsymbol{\lambda} = \boldsymbol{\sigma}^2/\boldsymbol{\sigma}_u^2$ (which will be defined in this Section). Therefore, the selection of smoothing parameter $\boldsymbol{\lambda}$ which becomes a problem of variance components estimation.

Recall that we are interested on estimating the model,

$$\boldsymbol{y} = \boldsymbol{B\theta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma}^2\boldsymbol{I_n}) \qquad (\mathbf{3.3})$$

The aim is to formulate the P-spline model into a mixed model $(\mathbf{2.1})$. This reformulation can be viewed as a reparametrization of the original non parametric model, for which we transform the model B-spline basis into a new model basis **(Lee, 2010),**

$$\boldsymbol{B} \to [\boldsymbol{X}\colon \boldsymbol{Z}]$$

This representation decomposes the fitted values as the sum of a unpenalized part $(\boldsymbol{X\beta})$ and a non linear penalized part $(\boldsymbol{Zu})$ smooth term.

We follow the approach of **Lee (2010)** and use the B-spline basis and the usual penalty $\boldsymbol{P} = \boldsymbol{D}'\boldsymbol{D}$ to reparametrize the original model into a mixed model. Let

$$\boldsymbol{D}'\boldsymbol{D} = \boldsymbol{U\Sigma U}'$$

be the singular value decomposition (SVD) of the penalty matrix P, where U is a matrix that contains the eigenvectors of the SVD and $\boldsymbol{\Sigma}$ is a diagonal matrix containing the eigenvalues, with q null eigenvalues.

Then, we can compose the penalty as (see **Lee (2010)**)

$$\boldsymbol{D}'\boldsymbol{D} = (\boldsymbol{U_n}\colon \boldsymbol{U_s})\begin{pmatrix} \boldsymbol{0}_q & \boldsymbol{0} \\ \boldsymbol{0} & \widetilde{\boldsymbol{\Sigma}} \end{pmatrix}\begin{pmatrix} \boldsymbol{U_n} \\ \boldsymbol{U_s} \end{pmatrix}$$

where $\boldsymbol{0_q}$ is square matrix of zeroes of order **q** and $\widetilde{\boldsymbol{\Sigma}}$ are the (c - q) positive eigenvalues and $\boldsymbol{U_n}$ contains the eigenvector associated with the null eigenvalues and $\boldsymbol{U_s}$ (with dimension dxq) contains the non null part of the decomposition (eigenvectors of the non null eigenvalues).

Let $\boldsymbol{D}'\boldsymbol{D} = \boldsymbol{U\Sigma U}'$ be the singular value decomposition of the penalty matrix, the matrix of the eigen values can be splitted by,

$$\boldsymbol{U} = (\boldsymbol{U_n}\colon \boldsymbol{U_s})$$

.Given the singular value decomposition of $\boldsymbol{D}'\boldsymbol{D}$ , the fixed and random effects matrices are (see **Lee (2010)**)

$$\boldsymbol{X} = \boldsymbol{BU_n} \qquad (\boldsymbol{dxq})$$

$$\boldsymbol{Z} = \boldsymbol{BU_s} \qquad (\boldsymbol{dx}(\boldsymbol{d}-\boldsymbol{q}))$$

where n=dxq (d=basis dimension and q=orden of the penalization) and s=dx(d-q).

And the coefficients,

$$\boldsymbol{\beta} = \boldsymbol{U_n}'\boldsymbol{\theta} \qquad (\boldsymbol{qx1})$$

$$\boldsymbol{u} = \boldsymbol{U_s}'\boldsymbol{\theta} \qquad ((\boldsymbol{d}-\boldsymbol{q})\boldsymbol{x1})$$

Therefore the model **(3.3)** becomes as,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \qquad (\boldsymbol{3.4})$$

$$\boldsymbol{u} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G}) \;\; \boldsymbol{and} \;\; \boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma}^2 \boldsymbol{I_n})$$

With the new reparametrization, the variance components matrix, $\boldsymbol{G}$ becomes

$$\boldsymbol{G} = \boldsymbol{\sigma}^2(\boldsymbol{\lambda}\widetilde{\boldsymbol{\Sigma}})^{-1} = \frac{\boldsymbol{\sigma}^2}{\boldsymbol{\lambda}}\widetilde{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{\sigma}_u^2\widetilde{\boldsymbol{\Sigma}}^{-1}$$

Therefore $\boldsymbol{\lambda}$ becomes,

$$\boldsymbol{\lambda} = \frac{\boldsymbol{\sigma}^2}{\boldsymbol{\sigma}_u^2}$$

It should be noted that the fixed parameters $\boldsymbol{\beta}$ in model **(3.4)** are unpenalised. Accordingly **(see Lee, 2010)** the fixed effect matrix $\boldsymbol{X} = \boldsymbol{B}\boldsymbol{U_n}$may be replaced by any sub-matrix such that:

- The composed matrix **[X : Z]** has full rank. This also implies that both **X** and **Z** have full column rank.
- **X** and **Z** are orthogonal, i.e. $\boldsymbol{X}'\boldsymbol{Z} = \boldsymbol{0}$

Assuming a second order penalty, i.e. q = 2, the diagonal matrix of eigenvalues has two zeroes and d-2 positive eigenvalues. Then, the fixed effects matrix can be taken as:

$$\boldsymbol{X} = [\boldsymbol{1}\!:\!\boldsymbol{x}]$$

where 1 is a vector of ones and **x** is the covariate vector.

Morever, by taking the random effect matrix as $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{U_s}\widetilde{\boldsymbol{\Sigma}}^{-\mathbf{0.5}}$ (instead of $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{U_s}$) we obtain that $\boldsymbol{G} = \boldsymbol{\sigma}_{\boldsymbol{u}}^{\mathbf{2}}\boldsymbol{I}_{(\boldsymbol{d}-\boldsymbol{q})}$ **(Lee and Durban, 2009).** This new feature has attractive properties from a computational point of view, specially it allows to use standard software.

## Chapter 4

# Flexible mixed models

As we mentioned before, in many longitudinal studies the response variable should be modeled as a non-linear function of age for each individual. In some circumstances the parametric assumption of the linear mixed models may not be appropriate.

In Chapter 3, we introduced P-splines to obtain more flexible regression models. The mixed model representation of the P-splines regression models allows us to take advantage of the methodology and software existent for mixed model analysis and makes possible a simple implementation of otherwise complicated models.

In this chapter we introduce how to include the P-spline methodology that we defined in Chapter 3, in to usual mixed model framework. For the sake of illustration, we will present the models that will be applied to our dataset from the simplest linear mixed model which is a random intercept model, up to a more complicated model with interaction between age and the type of pathology and with individual curves for each subject.

The variables of the dataset are;

***Height***: Height (kg) of the children measured in each six months from 6 months to 18 years.

***Z-score***:  Measure that express the distance between an individual child's height and the average height of comparable children in the reference population.

***Type of pathology***: The categorical covariate that indicates the type of pathology of the children (either HPA or PKU).

***Age:*** Age of the children (6 months up to 18 years) when the measures are taken.

***Gender:*** Gender of the children (Female-Male).

The aim of the study is to model the z-score of the children and to evaluate the long-term pathology type effect on children`s z-score. Figure 4 represents the z-score measurements of each individual divided by gender and the two types of pathology; PKU and HPA.

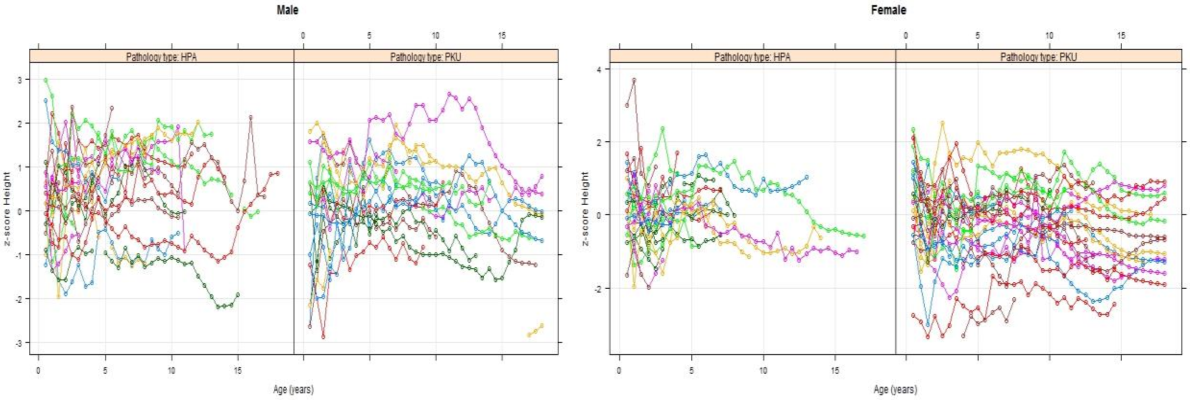

**Figure 4.1**: The dataset that corresponds to z-score measurements of 109 children that are taken between the age 6 months up to 18 years.

As can be observed in Figure 4.1 the measurements of z-scores for each individual behaves different during the period of the study and as we see in the graphic the parametric assumption of the traditional linear mixed models may not be appropriate to study these individual trajectories. Furthermore we will discuss this behavior applying different kind of linear and flexible models.

## 4.1. Random intercept model

Let $\boldsymbol{y_i}$ denote the z-score of a child *i*, *i=1,…N*, the first model that we can propose for the data may be a linear mixed model,

$$\boldsymbol{y_i = \beta_0 + \beta_1 t_i + U_i + \varepsilon_i} \qquad \boldsymbol{(4.1)}$$

Where $\boldsymbol{y_i}$ is the response variable vector $\boldsymbol{(y_{i1}, \dots y_{in_i})'}$ which is z-scores in our application, $\boldsymbol{t_i}$ is the vector of covariates (in this case the covariate is the age variable) $\boldsymbol{(t_{i1}, \dots t_{in_i})'}$ . $\boldsymbol{\beta_0}$ is the overall mean and $\boldsymbol{U_i}$ is a random intercept for the child *i* , which is treated as a random sample from the population of children and requires just a single variance parameter, $\sigma_U^2$.

The model can be written in matrix notation as;

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}$$

Where $\boldsymbol{Y}$ is the response vector which is z-scores in our example and $\boldsymbol{X}$ is the vector of covariates. and $\boldsymbol{u} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$ and $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{R})$**.** The matrices in this model have the form of,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{1} & \boldsymbol{t_1} \\ \vdots & \vdots \\ \boldsymbol{1} & \boldsymbol{t_N} \end{bmatrix} \qquad \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{1_1} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{1_2} & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \boldsymbol{1_N} \end{bmatrix}$$

$$\boldsymbol{u} = [\boldsymbol{U_1}, \dots \boldsymbol{U_N}]$$

$$\boldsymbol{\beta} = [\boldsymbol{\beta_0}, \boldsymbol{\beta_1}]' \quad \boldsymbol{G} = \boldsymbol{\sigma_U^2 I_N}$$

This model is a random intercept model which we have already introduced in Chapter 2 **(2.2)**, that allows a random shift over the intercept and assume that the age effect is lineal ($\boldsymbol{\beta_1 t_i}$).

## 4.2 Random Intercept and Slope Model

As we see in Figure 4.1 the individual trajectories are not only change over intercept but also in slope. Therefore we define a new model which is explained in Chapter 2 **(2.3)** where the extension is to assume that the subject-specific differences are straight lines;

$$\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 t_i} + \boldsymbol{u_{i1}} + \boldsymbol{u_{i2} t_i} + \boldsymbol{\varepsilon_i} \qquad (\boldsymbol{4.2})$$

$$\boldsymbol{\varepsilon_i} \sim \boldsymbol{N}\left(\boldsymbol{0}, \boldsymbol{\sigma^2 I_{n_i}}\right) \qquad (\boldsymbol{u_{i1}}, \boldsymbol{u_{i2}})' \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

where the $\boldsymbol{\Sigma}$ is a unstructured 2x2 matrix for the intercept and slope $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_{u_1} & \rho \\ \rho & \sigma^2_{u_2} \end{pmatrix}$.

Here, the random part of the model is represented as $\boldsymbol{u_{i1} + u_{i2}t_i}$ which allows us to model individual straight lines not just changes over the intercept $\boldsymbol{\beta_0}$ but also changes in slope $\boldsymbol{\beta_1}$.

In this model, the random part have the variance parameters; $\boldsymbol{\sigma^2_{u_1}}$ which corresponds the random intercept and $\boldsymbol{\sigma^2_{u_2}}$ which corresponds the random slope.

In matrix notation,

$$\boldsymbol{Y = X\beta + Zu + \varepsilon}$$

The matrix Z now incorporates the subject-specific lines,

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & t_1 \\ \vdots & \vdots \\ \mathbf{1} & t_N \end{bmatrix} \quad \boldsymbol{Z} = \begin{bmatrix} \mathbf{1} & X_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & X_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1} & \vdots & \vdots & \vdots & X_N \end{bmatrix} = \begin{bmatrix} \mathbf{1}t_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}t_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \mathbf{1}t_N \end{bmatrix}$$

$$\boldsymbol{u} = \ [\boldsymbol{u_{11}, u_{12} \dots, u_{N1}, u_{N2}}]'$$

$$\boldsymbol{G} = \begin{pmatrix} G_1 & \dots & \dots \\ \vdots & \ddots & \vdots \\ \dots & \dots & G_N \end{pmatrix} \quad \boldsymbol{G_i} = \begin{pmatrix} \sigma^2_{u_1} & \rho \\ \rho & \sigma^2_{u_2} \end{pmatrix} \qquad \boldsymbol{\rho = cov(u_{i1}, u_{i2})}$$

Model **(4.2)** assumes a linear trend for the z-score along age but this linear assumption may not be appropriate for our data as we can see the behavior of the z-score measurements over the age (Figure 4.1). Therefore we can use P-splines smoothing to model the effect of age on the response variable z-score.

## 4.3. Non linear effect of age plus random intercept and slope

$$\boldsymbol{y_i = f(t_i) + u_{i1} + u_{i2}X_i + \varepsilon_i} \qquad (\mathbf{4.3})$$

where;

$y_i$ = z-score of the patient *i*, *i*=1,…N. $\boldsymbol{f}(.)$ is a smooth function which reflects the overall increasing trend of z-score overall age. $\boldsymbol{f}(.)$ is estimated by penalized splines $\boldsymbol{f(t_i) = B^i\theta}$ where $\boldsymbol{B^i}$ is the B-spline basis for the function $\boldsymbol{f(t_i)}$**.** As we described the representation of the P-splines regression model as mixed models **(3.4)** a penalized spline model for **(4.3)** is,

$$\boldsymbol{y_i \;=\; \underbrace{X_i\,\widetilde{\beta} + \;Z_i\widetilde{u}}_{f(t_i)} + \;u_{i1} + u_{i2}X_i + \varepsilon_i}$$

$$\boldsymbol{\varepsilon_i \sim N\left(0, \sigma^2 I_{n_i}\right) \qquad (u_{i1}, u_{i2})' \sim N(0, \Sigma) \quad \widetilde{u} \sim N\left(0, \sigma^2_{\widetilde{u}} I_{d-2}\right)}$$

Where $\boldsymbol{X_i} = [\mathbf{1}\colon \boldsymbol{t_i}]$ , $\boldsymbol{Z_i = B^i U_s \widetilde{\Sigma}^{-0.5}}$ , $\boldsymbol{\widetilde{\beta}} = (\boldsymbol{\widetilde{\beta}_1, \widetilde{\beta}_2})'$ (we use q=2 by assuming a second order penalty.

In matrix notation,

$$\boldsymbol{Y = X\beta + Zu + \varepsilon}$$

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{t_1} \\ \vdots & \vdots \\ \mathbf{1} & \boldsymbol{t_N} \end{bmatrix} \qquad \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z_1} & \boldsymbol{X_1} & \mathbf{0} & \dots & \mathbf{0} \\ \boldsymbol{Z_2} & \mathbf{0} & \boldsymbol{X_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Z_m} & \vdots & \vdots & \vdots & \boldsymbol{X_N} \end{bmatrix}$$

The random parts of the model are the following;

$$\boldsymbol{u} = \ [\widetilde{\boldsymbol{u}}, \boldsymbol{u}_{11}, \boldsymbol{u}_{12} \ldots, \boldsymbol{u}_{N1}\ , \boldsymbol{u}_{N2}]'$$

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{\sigma}^2_{\tilde{\mathrm{u}}}\boldsymbol{I} & & \boldsymbol{0} & \\ & \boldsymbol{\Sigma} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \vdots & \ddots & \vdots \\ & \boldsymbol{0} & \cdots & \boldsymbol{\Sigma} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\sigma}^2_{\tilde{\mathrm{u}}}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{blockdiagonal}\ \boldsymbol{\Sigma} \end{pmatrix}$$

Here the random part of the smooth function which reflects the overall increasing trend of the z-score over time has the following distribution;

$$\widetilde{\boldsymbol{u}} \sim \boldsymbol{N}\left(\boldsymbol{0}, \boldsymbol{\sigma}^2_{\tilde{\mathrm{u}}}\boldsymbol{I}_{\boldsymbol{d}-\boldsymbol{2}}\right)$$

## 4.4. Subject Specific Curves

To adjust a more flexible model, that allows subject-specific ltrends to be non linear the individual trends can be smoothed by using penalized splines. In this case. We assume the following regression model,

$$\boldsymbol{y}_{\boldsymbol{i}} \ = \boldsymbol{f}(\boldsymbol{t}_{\boldsymbol{i}}) + \boldsymbol{g}_{\boldsymbol{i}}(\boldsymbol{t}_{\boldsymbol{i}}) + \boldsymbol{\varepsilon}_{\boldsymbol{i}} \qquad\qquad (\boldsymbol{4}.\boldsymbol{4})$$

$\boldsymbol{y_i}$ = z-score of the patient *i*, *i*=1,…N. $\boldsymbol{g}_{\boldsymbol{i}}(.)$ is a smooth function which reflects the individual curves. $\boldsymbol{g}_{\boldsymbol{i}}(.)$ is estimated by penalized splines $\boldsymbol{g}_{\boldsymbol{i}}(\boldsymbol{t}_{\boldsymbol{i}}) = \boldsymbol{B}^{\boldsymbol{i}}\boldsymbol{\theta}_{\boldsymbol{i}}$. Following the ideas presented in Chapter 2 and 3, the mixed model representation of $(\boldsymbol{4}.\boldsymbol{4})$ becomes:

$$\boldsymbol{y}_{\boldsymbol{i}} \ = \ \underbrace{\boldsymbol{X}_{\boldsymbol{i}}\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z}_{\boldsymbol{i}}\tilde{\mathbf{u}}}_{\boldsymbol{f}(\boldsymbol{t}_{\boldsymbol{i}})} + \underbrace{\boldsymbol{X}_{\boldsymbol{i}}\boldsymbol{\beta}^{\boldsymbol{i}} + \boldsymbol{Z}_{\boldsymbol{i}}\tilde{\mathbf{u}}^{\boldsymbol{i}}}_{\boldsymbol{g}_{\boldsymbol{i}}(\boldsymbol{t}_{\boldsymbol{i}})} + \boldsymbol{\varepsilon}_{\boldsymbol{i}}$$

$$\tilde{\mathbf{u}} \sim \boldsymbol{N}\left(\boldsymbol{0}, \boldsymbol{\sigma}^2_{\tilde{\mathrm{u}}}\boldsymbol{I}_{\boldsymbol{d}-\boldsymbol{2}}\right) \quad \boldsymbol{\varepsilon}_{\boldsymbol{i}} \sim \boldsymbol{N}\left(\boldsymbol{0}, \boldsymbol{\sigma}^2\boldsymbol{I}_{\boldsymbol{n}_{\boldsymbol{i}}}\right) \quad \tilde{\mathbf{u}}^{\boldsymbol{i}} \sim \boldsymbol{N}\left(\boldsymbol{0}, \boldsymbol{\sigma}^2_{\tilde{\mathrm{u}}^*}\boldsymbol{I}_{\boldsymbol{d}-\boldsymbol{2}}\right)$$

where $\boldsymbol{X_i} = [\boldsymbol{1}\colon \boldsymbol{t_i}]$, $\boldsymbol{Z_i} = \boldsymbol{B^i U_s \widetilde{\Sigma}^{-0.5}}$ $\boldsymbol{f(t_i)} = \boldsymbol{B^i \theta_i}$, $\boldsymbol{g_i(t_i)} = \boldsymbol{B^i \theta_i}$ and $\boldsymbol{\widetilde{\beta}} = (\boldsymbol{\widetilde{\beta}_1}, \boldsymbol{\widetilde{\beta}_2})'$, .

In model **(4.3),** the individual trajectories were lineal, $\boldsymbol{u_{i1} + u_{i2} x_{ij}}$ **,** therefore in the model **(4.4),** we have individual curves (subject specific curves) which have two components: a linear part $\boldsymbol{X_i \beta^i}$ (similar in model **4.3**), and a non linear part $\boldsymbol{Z_i \widetilde{u}^i}$ , which allows more flexibility in the model. Both components are random part of the model as described by **Durban et al (2005).**

In matrix notation,

$$\boldsymbol{Y = X\beta + Zu + \varepsilon}$$

The complex model **(4.4),** can be easily described in the mixed model framework as ;

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{1} & \boldsymbol{X_1} \\ \vdots & \vdots \\ \boldsymbol{1} & \boldsymbol{X_N} \end{bmatrix} \qquad \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z_1} & \boldsymbol{X_1} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{Z_1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{Z_2} & \boldsymbol{0} & \boldsymbol{X_2} & \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z_2} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Z_N} & \vdots & \vdots & \vdots & \boldsymbol{X_N} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{Z_N} \end{bmatrix}$$

$$\boldsymbol{u} = [\boldsymbol{\widetilde{u}}, \boldsymbol{\beta^1}, \ldots \boldsymbol{\beta^N}, \boldsymbol{\widetilde{u}^1}, \ldots \boldsymbol{\widetilde{u}^N}]'$$

$$\boldsymbol{G} = \boldsymbol{Cov(u)} = \begin{pmatrix} \boldsymbol{\sigma^2_{\widetilde{u}} I_{d-2}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{blockdiagonal\,\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{blockdiagonal\,\sigma^2_{\widetilde{u}^*} I_{d-2}} \end{pmatrix}$$

As we can see in the mixed model framework representation, the random part of the model includes, $\boldsymbol{\beta^1}, \ldots \boldsymbol{\beta^N}, \boldsymbol{\widetilde{u}^1}, \ldots \boldsymbol{\widetilde{u}^N}$ which means that both the linear part and the non-linear part of the subject specific curves are taken as random. The G matrix has an additional term in comparison with model **(4.3)** which is the variance parameter associated with the penalized/non linear part of the subject specific curves. As can be observed, the same amount of smoothing was assumed for each individual $\boldsymbol{\sigma^2_{\widetilde{u}^*} I_{d-2}}$**.**

$\mathbf{Z}$ matrix have 3 components; the first component $(\boldsymbol{Z_1}, \boldsymbol{Z_2}, \dots \boldsymbol{Z_N})'$ represents the $\boldsymbol{f(t_i)}$ which is the population effect of age, the second matrix is the random intercept and slope part of the model and the last matrix is the penalized/non linear part for each individual.

## 4.5. Factor by Curve Interactions

As in the application part of this study, one of the purposes of the analyses is to compare the long-term effects of two different pathology types. So we are interested in fitting separate curves for each pathology type.

$$\boldsymbol{y_i = \alpha_0 + \alpha_1 pth_{iPKU} + f(t_i) + h(t_i) pth_{iPKU} + g_i(t_i) + \varepsilon_i} \qquad \mathbf{(4.5)}$$

where $\boldsymbol{pth_{iPKU}} = \mathbf{1}$ if the pathology of individual i is PKU and 0 otherwise. Based on the previous parameterization $\boldsymbol{f}(.)$ represents the smooth effect of age in the group with HPA, whereas $\boldsymbol{h}(.)$ represents the deviation from this effect for those individuals with PKU. Accordingly, the age effect in an individual with PKU is $\boldsymbol{f(t_i) + h(t_i)}$. Therefore if $\boldsymbol{h(t)} = \mathbf{0}\ \forall \boldsymbol{t}$ , it means that no interaction between age and type of pathology is present.

The mixed model representation of model $(\mathbf{4.5})$ is,

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

where $\boldsymbol{\widetilde{X}_i} = \begin{bmatrix} \boldsymbol{t_{i1}} \\ \vdots \\ \boldsymbol{t_{in_i}} \end{bmatrix}$. It should be noted that the column of ones in $\boldsymbol{X_i}$ has been removed in order to identify the model and

$$\boldsymbol{\widetilde{u} \sim N\left(0, \sigma^2_{\widetilde{u}} I_{d-2}\right), w \sim N(0, \sigma^2_w I_{d-2}), \beta^i \sim N(0, \Sigma), \varepsilon_i \sim N(0, \sigma^2_\varepsilon), \widetilde{u}^i \sim N\left(0, \sigma^2_{\widetilde{u}^*} I_{d-2}\right)}$$

In matrix notation,

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}$$

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

where $\boldsymbol{n_{HPA}}$ is the number of individuals with HPA and $\boldsymbol{\beta} = (\boldsymbol{\alpha_0}, \boldsymbol{\alpha_1}, \boldsymbol{\widetilde{\beta}}, \boldsymbol{\gamma})$

$$\boldsymbol{Z} = \begin{bmatrix} Z_1 & 0 & X_1 & 0 & \dots & \dots & \dots & 0 & Z_1 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & 0 & X_2 & 0 & \dots & \dots & \vdots & 0 & Z_2 & 0 & \dots & \dots & \dots \\ Z_{n_{HPA}} & 0 & 0 & 0 & \ddots & 0 & \dots & \vdots & \vdots & 0 & \ddots & 0 & \dots & \vdots \\ Z_{n_{HPA}}+1 & Z_{n_{HPA}}+1 & \vdots & \vdots & 0 & \ddots & 0 & \vdots & \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 & \ddots & 0 & \vdots & \vdots & \vdots & 0 & \ddots & 0 \\ Z_{102} & Z_{102} & 0 & 0 & \vdots & \vdots & 0 & X_N & 0 & \dots & 0 & \dots & 0 & Z_N \end{bmatrix}$$

$$\boldsymbol{u} = \left[\widetilde{\boldsymbol{u}}, \boldsymbol{w}, \widetilde{\boldsymbol{\beta}}^1, \dots \widetilde{\boldsymbol{\beta}}^N, \widetilde{\boldsymbol{u}}^1, \dots \widetilde{\boldsymbol{u}}^N\right]'$$

$$\boldsymbol{G} = \begin{pmatrix} \sigma_{\tilde{u}}^2 I & 0 & \dots & 0 \\ \vdots & \sigma_w^2 I & \dots & \dots \\ \vdots & 0 & blockdiagonal\,\Sigma & 0 \\ 0 & 0 & 0 & block\ diagonal\ \sigma_{\tilde{u}^*}^2 I \end{pmatrix}$$

In this model, the random part of the deviation of HPA from PKU has the following distribution;

$$\boldsymbol{w} \sim \boldsymbol{N}(\boldsymbol{0}, \sigma_w^2 \boldsymbol{I}_{d-2})$$

Furthermore, we will discuss if this random part is necessary for the model, by investigating whether this variance parameter $\sigma_w^2$ is zero.

## CHAPTER 5

# Application to real data

### 5.1 Children Growth Data

In Chapter 4, we have introduced some flexible mixed models with P-spline regression models to furthermore apply to the data of children growth. The data contains the patients 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 which are controlled with a dietetic treatment. The data has 102 children followed up from 6 months up to 18 years in the center.

For the correct physical development of children with PKU, it is important to have a special diet while for those with HPA it isn`t necessary a medical intervention. The objective of this study is to model the growth of these children with PKU and HPA and see if there is any difference between these two types of pathologies during a period of time. (from 6 months to 18 years).

The variables are;

***Height***: Height (kg) of the children measured in each six months from 6 months to 18 years.

***Z-score***: Measure that express the distance between an individual child's height and the average height of comparable children in the reference population.

***Type of pathology***: The categorical covariate that indicates the type of pathology of the children (either HPA or PKU).

***Age:*** Age of the children (from 6 months up to 18 years) when the measures are taken.

***Gender:*** Gender of the children (Female-Male).

The characteristics of the dataset are,

The dataset has 75 columns in total which are,

- “Patient Id” indicates the id number of the children
- “Type of Pathology” indicates if the child has HPA or PKU.
- “Gender” indicates the gender of each child
- (“Heigth 0.5” , “Height 1” , “Height 1.5”, … , “Height 18” ) which indicate the measures of height for each child at different ages.
- (“Z-score 0.5” , “Z-score 1” , “Z-score 1.5”, … , “Z-score 18” ) which indicate the measures of z-scores for each child at different ages.

It is necessary to reshape the data to obtain a longitudinal format in which we have the z-scores, type of pathology, id number and the age variable for each child.

```
data.long <- reshape(data, varying = list(names(data)[seq(4,74,2)],
+names(data)[seq(5,75,2)]), direction = "long", v.names = c("Height", "z_score"),
+idvar = "Patient Id", ages = seq(0.5, 18, by = 0.5))
```

```
> summary(data.long)
```

| Patient.Id | Type.of.Pathology | Gender | time | Height | z_score | Patient Id |
|---|---|---|---|---|---|---|
| Min. : 1 | HPA :1980 | Male :1584 | Min. : 0.500 | Min. : 59.5 | Min. :-21.8294 | Min. : 1.0 |
| 1st Qu.: 28 | PKU :1944 | Female:2340 | 1st Qu.: 4.875 | 1st Qu.: 94.0 | 1st Qu.: -0.5900 | 1st Qu.: 36.0 |
| Median : 55 | NA's:1188 | NA's :1188 | Median : 9.250 | Median :116.5 | Median : 0.1100 | Median : 71.5 |
| Mean : 55 | | | Mean : 9.250 | Mean :117.7 | Mean : -0.0152 | Mean : 71.5 |
| 3rd Qu.: 82 | | | 3rd Qu.:13.625 | 3rd Qu.:140.0 | 3rd Qu.: 0.6700 | 3rd Qu.:107.0 |
| Max. :109 | | | Max. :18.000 | Max. :182.5 | Max. : 3.6800 | Max. :142.0 |
| NA's :1188 | | | | NA's :2971 | NA's :2967 | |

As can be observed, we create a sequence from 0.5 up to 18 by 0.5 which indicates the variable “Age”. This new shape of the dataset has a longitudinal data format.

We want to apply the models will be defined in Chapter 4 for both gender separately, therefore we divide the dataset in two groups (male, female).

In the following tables, the mean, standart deviation of z-scores with the p-values of Kruskal-Wallis test for two groups for each age is presented to see descriptively if there are any differences between the z-score means of the children with PKU and with HPA in any age.

| | **HPA** | **PKU** | **p-value** |
|---|---|---|---|
| **6 months** | 0.31 (1.001) - 23 | -0.092 (1.404) - 21 | 0.768 |
| **1 years** | 0.301 (1.053) - 23 | -0.174 (1.18) - 21 | 0.253 |
| **1.5 years** | 0.151 (1.072) - 23 | 0.231 (1.261) - 21 | 0.671 |
| **2 years** | 0.255 (0.967) - 22 | 0.062 (0.929) - 21 | 0.535 |
| **2.5 years** | 0.498 (1.027) - 22 | 0.026 (0.709) - 20 | 0.075 |
| **3 years** | 0.324 (0.891) - 22 | 0.123 (0.697) - 19 | 0.44 |
| **3.5 years** | 0.425 (0.926) - 19 | 0.317 (0.764) - 19 | 0.588 |
| **4 years** | 0.522 (0.85) - 19 | 0.189 (0.593) - 19 | 0.095 |
| **4.5 years** | 0.695 (0.664) - 18 | 0.359 (0.552) - 19 | 0.131 |
| **5 years** | 0.632 (0.843) - 17 | 0.436 (0.772) - 19 | 0.333 |
| **5.5 years** | 0.667 (1.007) - 18 | 0.429 (0.708) - 19 | 0.301 |
| **6 years** | 0.468 (0.981) - 17 | 0.402 (0.695) - 19 | 0.399 |
| **6.5 years** | 0.605 (0.972) - 16 | 0.606 (0.813) - 18 | 0.704 |
| **7 years** | 0.568 (1.085) - 16 | 0.434 (0.842) - 18 | 0.501 |
| **7.5 years** | 0.745 (1.075) - 16 | 0.475 (0.737) - 16 | 0.273 |
| **8 years** | 0.533 (1.05) - 16 | 0.388 (0.802) - 16 | 0.346 |
| **8.5 years** | 0.696 (0.957) - 14 | 0.39 (0.949) - 15 | 0.213 |
| **9 years** | 0.624 (1.049) - 13 | 0.413 (0.844) - 16 | 0.456 |
| **9.5 years** | 0.501 (1.069) - 12 | 0.347 (0.765) - 15 | 0.66 |
| **10 years** | 0.479 (1) - 12 | 0.285 (0.788) - 14 | 0.718 |
| **10.5 years** | 0.448 (1.026) - 12 | 0.333 (0.846) - 12 | 0.862 |
| **11 years** | 0.413 (1.143) - 10 | 0.32 (1.039) - 10 | 0.649 |
| **11.5 years** | 0.696 (1.11) - 8 | 0.346 (1.032) - 10 | 0.328 |
| **12 years** | 0.7 (1.347) - 7 | 0.375 (1.031) - 10 | 0.494 |
| **12.5 years** | 0.387 (1.387) - 6 | 0.265 (1.148) - 10 | 0.588 |
| **13 years** | 0.318 (1.467) - 6 | 0.182 (1.081) - 10 | 0.48 |
| **13.5 years** | -0.118 (1.476) - 5 | 0.228 (1.011) - 9 | 0.894 |
| **14 years** | -0.242 (1.314) - 5 | 0.168 (0.881) - 9 | 0.639 |
| **14.5 years** | -0.625 (1.173) - 4 | -0.009 (0.927) - 8 | 0.395 |
| **15 years** | -0.763 (1.021) - 3 | -0.169 (0.828) - 8 | 0.347 |
| **15.5 years** | 0.24 (0.391) - 3 | -0.161 (0.73) - 8 | 0.539 |
| **16 years** | 0.717 (1.231) - 3 | -0.15 (0.658) - 8 | 0.357 |
| **16.5 years** | 0.23 (0.225) - 3 | -0.159 (0.567) - 8 | 0.303 |
| **17 years** | 0.405 (0.134) - 2 | -0.472 (1.032) - 9 | 0.125 |
| **17.5 years** | 0.824 (NA) - 1 | -0.484 (1.015) - 9 | 0.116 |
| **18 years** | 0.851 (NA) - 1 | -0.388 (1.032) - 8 | 0.12 |

**Table 5.1**: Means, standart deviations, p-values of Kruskal-Wallis test for two groups and frequencies for each age of the male individuals and for each type of pathology.

|  | HPA | PKU | p-value |
|---|---|---|---|
| **6 months** | 0.374 (0.833) - 32 | 0.412 (1.099) - 33 | 0.849 |
| **1 years** | 0.257 (1.007) - 32 | -0.05 (0.996) - 32 | 0.412 |
| **1.5 years** | -0.025 (0.718) - 29 | -0.381 (1.055) - 33 | 0.268 |
| **2 years** | -0.261 (0.767) - 28 | -0.33 (0.884) - 30 | 0.809 |
| **2.5 years** | -0.062 (0.8) - 23 | -0.079 (1.043) - 30 | 0.879 |
| **3 years** | -0.019 (0.855) - 21 | -0.102 (1.066) - 28 | 0.952 |
| **3.5 years** | 0.096 (0.574) - 18 | -0.257 (0.997) - 26 | 0.176 |
| **4 years** | 0.302 (0.629) - 17 | -0.246 (1.03) - 28 | 0.084 |
| **4.5 years** | 0.235 (0.417) - 16 | -0.349 (0.89) - 28 | 0.01 |
| **5 years** | 0.166 (0.529) - 15 | -0.331 (1.019) - 28 | 0.062 |
| **5.5 years** | 0.316 (0.707) - 14 | -0.29 (0.995) - 28 | 0.033 |
| **6 years** | 0.249 (0.707) - 14 | -0.234 (0.898) - 28 | 0.142 |
| **6.5 years** | 0.322 (0.606) - 13 | -0.201 (0.961) - 28 | 0.133 |
| **7 years** | 0.221 (0.739) - 10 | -0.154 (1.08) - 26 | 0.447 |
| **7.5 years** | 0.171 (0.775) - 7 | -0.206 (0.937) - 26 | 0.402 |
| **8 years** | 0.143 (0.822) - 6 | 0.024 (0.842) - 25 | 0.707 |
| **8.5 years** | -0.102 (0.885) - 5 | -0.048 (0.873) - 25 | 0.759 |
| **9 years** | -0.128 (0.981) - 5 | -0.017 (0.846) - 24 | 0.685 |
| **9.5 years** | -0.09 (0.837) - 4 | -0.124 (0.804) - 24 | 1 |
| **10 years** | -0.022 (1.007) - 4 | -0.192 (0.797) - 24 | 0.767 |
| **10.5 years** | -0.1 (1.005) - 4 | -0.205 (0.881) - 23 | 0.973 |
| **11 years** | 0.005 (0.889) - 4 | -0.089 (1.01) - 23 | 0.891 |
| **11.5 years** | -0.213 (0.993) - 4 | -0.246 (1.016) - 22 | 0.972 |
| **12 years** | -0.162 (0.956) - 4 | -0.336 (1.002) - 21 | 0.738 |
| **12.5 years** | -0.254 (1.028) - 4 | -0.417 (0.986) - 21 | 0.795 |
| **13 years** | -0.198 (0.959) - 4 | -0.548 (1.097) - 19 | 0.516 |
| **13.5 years** | -0.477 (0.382) - 3 | -0.647 (1.101) - 17 | 0.711 |
| **14 years** | -0.643 (0.326) - 3 | -0.732 (1.089) - 17 | 0.751 |
| **14.5 years** | -0.67 (0.354) - 2 | -0.715 (1.021) - 17 | 0.74 |
| **15 years** | -0.715 (0.46) - 2 | -0.678 (0.882) - 16 | 0.778 |
| **15.5 years** | -0.79 (0.453) - 2 | -0.683 (0.866) - 16 | 0.888 |
| **16 years** | -0.74 (0.311) - 2 | -0.667 (0.872) - 16 | 0.725 |
| **16.5 years** | -0.775 (0.318) - 2 | -0.605 (0.861) - 15 | 0.94 |
| **17 years** | -0.58 (NA) - 1 | -0.52 (0.89) - 14 | 0.907 |
| **17.5 years** | NaN (NA) - 0 | -0.556 (0.924) - 13 | NA |
| **18 years** | NaN (NA) - 0 | -0.541 (0.958) - 13 | NA |

**Table 5.2**: Means, standart deviations, p-values of Kruskal-Wallis test for two groups and frequencies for each age of the female individuals and for each type of pathology

The Kruskal-Wallis test results shows us that there isn`t any differences between the z-score means of the two type of pathologies in any age except the 4.5 and 5.5 years old females.

In this Chapter we will present the application of the regression models presented in Chapter 4 to analyze the children growth data.

First of all we present the R code of the function to create the P-spline regression basis,

```
>bspline <-function(X., XL., XR., NDX., BDEG.){
  dx <- (XR. - XL.)/NDX.
  knots <- seq(XL. - BDEG.*dx, XR. + BDEG.*dx, by=dx)
  B <- spline.des(knots, X., BDEG.+1, 0*X.)$design
  B
}

>MM.basis <- function (x, xl, xr, ndx, bdeg, pord, decom = 2) {
  B = bspline(x,xl,xr,ndx,bdeg)
  m = ncol(B)
  n = nrow(B)
  D = diff(diag(m), differences=pord)
  P.svd = svd(crossprod(D))
  U = (P.svd$u)[,1:(m-pord)] # eigenvectors
  d = (P.svd$d)[1:(m-pord)]  # eigenvalues
  Delta = diag(1/sqrt(d))
  Z = B%*%U%*%Delta
  X = NULL
  for(i in 0:(pord-1)){
    X = cbind(X,x^i)
  }
 > list(X = X, Z = Z, d = d, B = B)
}
```

As it can be observed, in this code we calculated $\boldsymbol{Z}$ as $\boldsymbol{BU_s\widetilde{\Sigma}^{-0.5}}$ as it mentioned in Section 3.5.

Before applying the defined regression models presented in Chapter 4 to our data we introduce some basic code.

```
>attach(data.male)

>Id <- factor(rep(1, length = length(z_score)))

>K <- max(5,min(floor(length(unique(age))/4),40))

>MM = MM.basis(age, min(age)- 0.5, max(age) + 0.5, K, 3, 2)

>Z = MM$Z
```

```
>MM.subject <- MM.basis(age, min(age)- 0.5, max(age) + 0.5, K, 3, 2)

>Z.subject <- MM.subject$Z

>n = length(z_score)
```

where ***K*** is the optimum number of knots defined in Section 3.2.

## 5.2 The model with factor by curve interactions

The model that we applied for the data was with the subject specific curves and factor by curve interactions **(4.5),**

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

The aim of the study is too see the long-term effect of the two types of pathologies so that we have used an interaction model in which a categorical factor (pathology type) interacts with a continuous factor (age).

We introduce matrices defined for the model (**4.5)** in Chapter 4 to the model,

***R Code***

```
>X <- model.matrix(z_score ~ Type.of.Pathology*age)

>Z.interact <- model.matrix(z_score ~ Z*Type.of.Pathology)[,-c(1,ncol(Z)+2)]

>Z.1 <- Z.interact[,1:ncol(Z)]

>Z.2 <- Z.interact[,(ncol(Z)+1):(2*ncol(Z))]

>Z.block <- list(Id = pdIdent(~Z.1-1), Id = pdIdent(~Z.2-1), Patient.Id =
pdSymm(~age), +Patient.Id = pdIdent(~Z.subject - 1))
```

where ***Z.block*** is the design matrix for the model **(4.5),** and have four parts,

1. ***Id = pdIdent(~Z.1-1)*** indicates $\boldsymbol{f}(.)$
2. ***Id = pdIdent(~Z.2-1)*** indicates $\boldsymbol{h}(.)$
3. ***Patient.Id = pdSymm(~age)*** indicates the linear part of the $\boldsymbol{g_i}(.)$
4. ***Patient.Id = pdIdent(~Z.subject - 1))*** indicates the non-linear part of the $\boldsymbol{g_i}(.)$

And ***pdIdent*** specifies that the variance structure of the random effects is a multiple of the identitiy matrix, ***pdSymm*** specifies covariance structure for the random intercept and slope.

It is necessary to create a structure for the data for the reason of working with matrices not with the variables.

```
>data.fr <- groupedData(z_score ~ X[,-1]|rep(1, length = length(z_score)), data =
+data.frame(z_score, X, Z.1, Z.2, Z.subject, Patient.Id, Type.of.Pathology))
```

```
>fit1 <- lme(z_score ~ X[,-1], random = Z.block, data = data.fr)
```

We applied the model **(4.5)** according to steps in R code for both males and females. The numerical output for the applied **(4.5)** model for the boys shows that the estimated variance parameter of non-linear time effect is $\boldsymbol{\sigma^2_{\tilde{u}}} = (\mathbf{0.1196})^{\mathbf{2}}$ and the variance parameter of the deviation is $\boldsymbol{\sigma^2_w} = (\mathbf{0.000224})^{\mathbf{2}}$**.**

The variance parameter of the non-linear part of the individual curves $\boldsymbol{\sigma^2_{\tilde{u}^*}}$ is estimated $(\mathbf{0.3722})^{\mathbf{2}}$ **.**

$\boldsymbol{\sigma^2_{\tilde{u}_{11}}} = (\mathbf{0.9141})^{\mathbf{2}}$ which is the variance parameter of intercept and $\boldsymbol{\sigma^2_{\tilde{u}_{22}}} = (\mathbf{0.0476})^{\mathbf{2}}$ is the variance parameter of the slope. Also there is a correlation between the random intercept and slope, which is estimated,$\boldsymbol{\gamma} = -\mathbf{0.332}$**.** The Akaike Information Criterion for this model is 1317.416.

The numerical output for the adjusted **(4.5)** model for the girls shows that the variance parameters are the following, the estimated variance parameter of non-linear time effect is $\boldsymbol{\sigma^2_{\tilde{u}}} = (\mathbf{0.8128})^{\mathbf{2}}$ , and the variance parameter of the deviation is $\boldsymbol{\sigma^2_w} = (\mathbf{0.4529})^{\mathbf{2}}$ .

The variance parameter of the non-linear part of the individual curves is estimated $\boldsymbol{\sigma^2_{\tilde{u}^*}} = (\mathbf{0.5180})^{\mathbf{2}}$ and the intercept and slope variance parameters are $\boldsymbol{\sigma^2_{\tilde{u}_{11}}} = (\mathbf{0.7249})^{\mathbf{2}}$**,** $\boldsymbol{\sigma^2_{\tilde{u}_{22}}} = (\mathbf{0.03689})^{\mathbf{2}}$ and the correlation is $\boldsymbol{\gamma} = -\mathbf{0.32}$**.** The Akaike Information Criterion for this model is 1521.956.

The estimated curves for PKU and HPA and the difference between these curves with the confidence intervals obtained with bootstrap method are represented in Figure 5.1 for boys and in Figure 5.2 for girls.

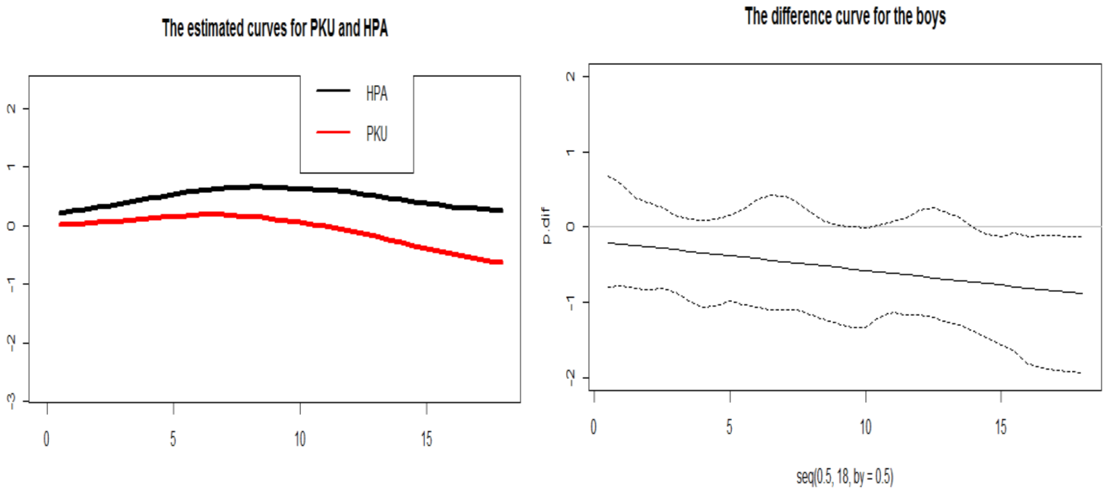

**Figure 5.1** The estimated curves for two types of pathology for the boys (left) and the difference between these two curves (right).

As we can observe in Figure 5.1 the male individuals with PKU has a lower z-scores which means the male individuals with HPA have a favorable evaluation especially during 5-15 years. This is observed from the graphic of the complete model but furthermore we will discuss whether the random parts of these model is necessary for the dataset.

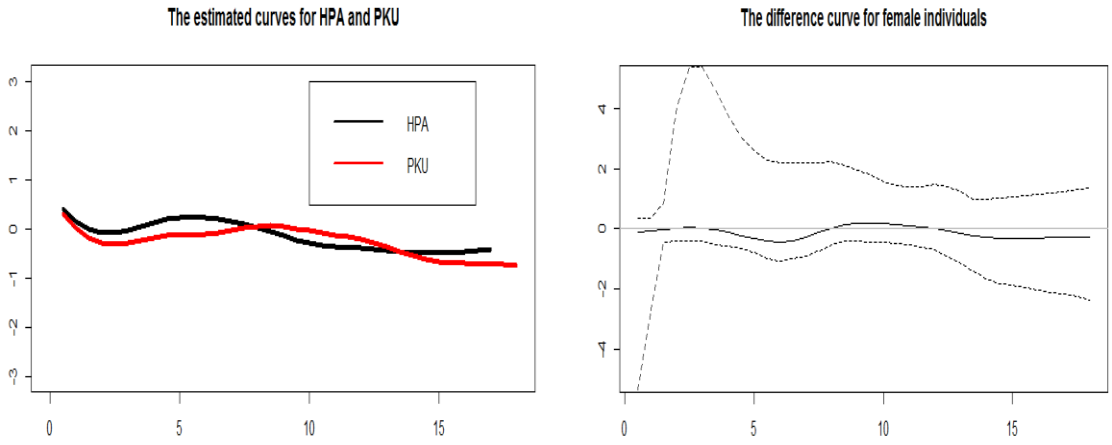

**Figure 5.2** The estimated curves for two types of pathology for the girls (left) and the difference between these two estimated curves (right)

In Figure 5.2 we observe that the children with PKU and the children with HPA have different behavior along the age which they begin with a higher z-scores then 0 and changes during the time of the study. We can observe that for the female individuals the children with HPA have a more favorable evolution then those with PKU until approximately 7 years old and then they become nearly equal to each other. After 14 years old the children with HPA becomes having a more favorable evolution.

Furthermore we will discuss if the observed differences between these two groups are statistically significant.

To see the differently estimated individual curves, in Figure 5.3 we present the estimated curves for some male and female individuals.

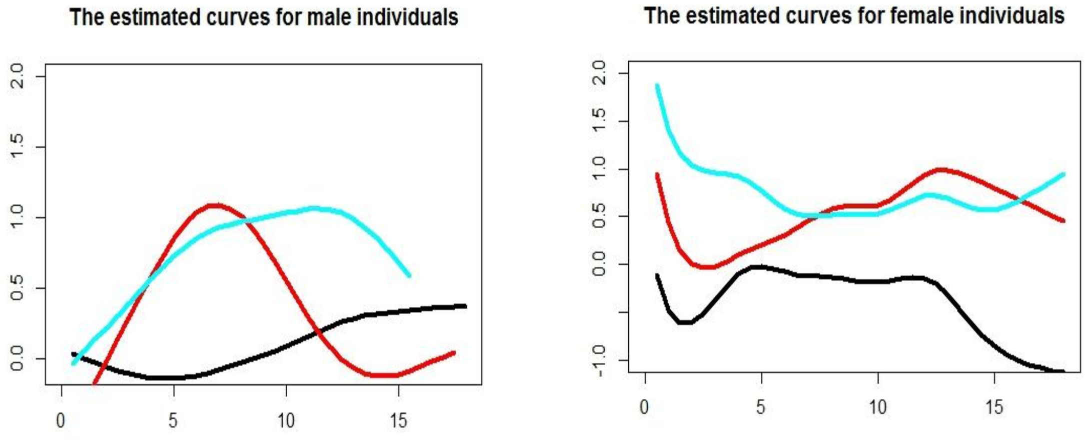

Figure 5.3 Estimated individual curves for patients with patient ID 2,15,43,53, 63, 82

## 5.3 The model without the non-linear part of the subject specific curves

We want to see if the assumption of individual curves is appropriate for our data. For this objective we extended the model **(4.3)** with the model **(4.5)** to obtain the model without assuming presence of the individual curves.

$$\boldsymbol{y_i = \alpha_0 + \alpha_1 pth_{iPKU} + f(t_i) + h(t_i)pth_{iPKU} + u_{i1} + u_{i2}X_i + \varepsilon_i} \qquad (\mathbf{5.1})$$

and we compare this model with the model **(4.5)** to test whether the random part of the subject specific curves is necessary for the model. To test this assumption we investigate if the variance parameter of this random part ($\boldsymbol{\sigma}^{2}_{\tilde{u}^*}$) is zero or not**,** which is equivalent to testing,

$$\boldsymbol{H_0}: \boldsymbol{\sigma}^{2}_{\tilde{u}^*} = \boldsymbol{0} \; ; \; \boldsymbol{H_1}: \boldsymbol{\sigma}^{2}_{\tilde{u}^*} > 0$$

To fit the given model ($\mathbf{5.1}$) we introduce the matrices,

***R Code***

```
>X <- model.matrix(z_score ~ Type.of.Pathology*age)

>Z.interact <- model.matrix(z_score ~ Z*Type.of.Pathology)[,-c(1,ncol(Z)+2)]

>Z.1 <- Z.interact[,1:ncol(Z)]

>Z.2 <- Z.interact[,(ncol(Z)+1):(2*ncol(Z))]

>Z.block <- list(Id = pdIdent(~Z.1-1), Id = pdIdent(~Z.2-1), Patient.Id =
pdSymm(~age))

>data.fr <- groupedData(z_score ~ X[,-1]|rep(1, length = length(z_score)), data =
>data.frame(z_score, X, Z.1, Z.2, Patient.Id, Type.of.Pathology))

>fit2 <- lme(z_score ~ X[,-1], random = Z.block, data = data.fr)
```

The difference between the model **(5.1)** and **(4.5)** is that we have removed the non-linear part of the individual curves, which is indicated by ***Patient.Id = pdIdent(~Z.subject - 1))*** in R code.

The numerical output for the applied **(5.1)** model for the boys shows that the estimated variance parameter $\boldsymbol{\sigma}^{2}_{\tilde{u}} = (\mathbf{0.1473})^{\mathbf{2}}$ and $\boldsymbol{\sigma}^{2}_{w} = (\mathbf{0.10124})^{\mathbf{2}}$ . The variance parameters, $\boldsymbol{\sigma}^{2}_{\tilde{u}_{11}} = (\mathbf{0.8709})^{\mathbf{2}}$ which is the variance parameter of intercept and $\boldsymbol{\sigma}^{2}_{\tilde{u}_{22}} = (\mathbf{0.0669})^{\mathbf{2}}$ is the variance parameter of the slope. Also there is a correlation between the random intercept and slope, which is estimated,$\boldsymbol{\gamma} = \mathbf{-0.332.}$ The Akaike Information Criterion for this model is 1550.223.

The numerical output for the adjusted **(5.1)** model for the girls shows that the variance parameters are the following, $\boldsymbol{\sigma_{\tilde{u}}^2} = (\mathbf{0.7298})^{\mathbf{2}}$ , $\boldsymbol{\sigma_w^2} = (\mathbf{0.4413})^{\mathbf{2}}$ and the intercept and slope variance parameters are $\boldsymbol{\sigma_{\tilde{u}_{11}}^2} = (\mathbf{0.7840})^{\mathbf{2}}$**,** $\boldsymbol{\sigma_{\tilde{u}_{22}}^2} = (\mathbf{0.1046})^{\mathbf{2}}$ and the correlation is $\boldsymbol{\gamma = -0.337}$**.** The Akaike Information Criterion for this model is 1818.985.

The estimated curves for PKU and HPA for the model **(5.1)** for both males and females are presented in Figure 5.3 and Figure 5.4

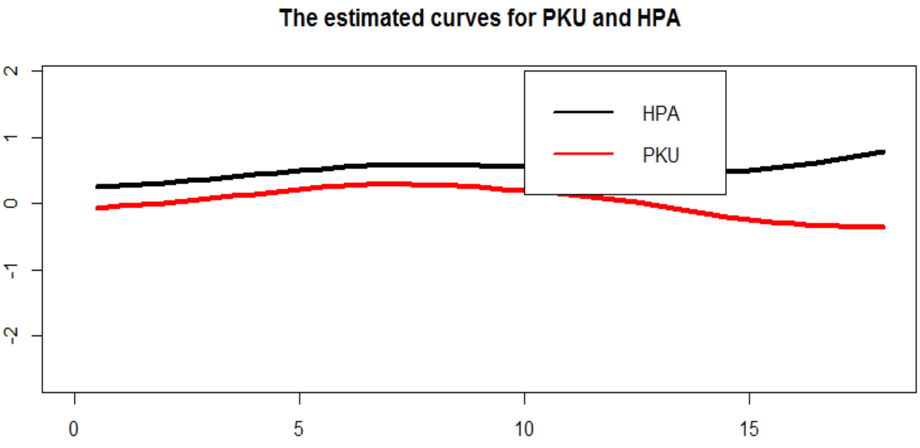

Figure 5.4 Estimated curves for males for the model (5.1)

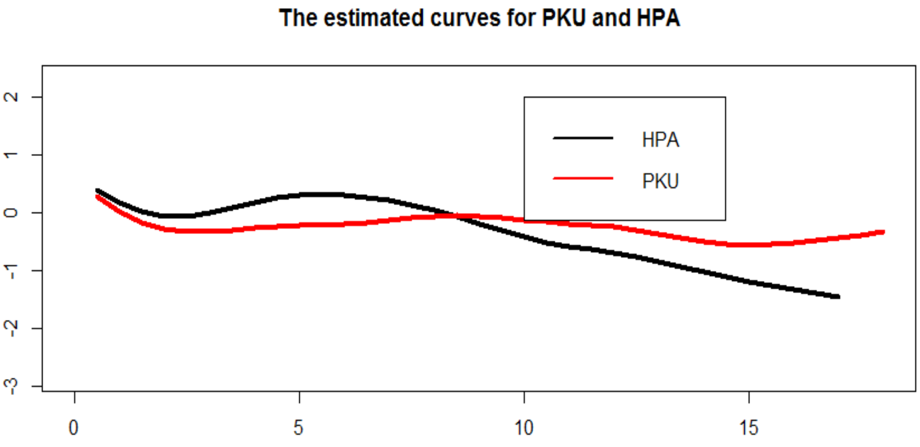

Figure 5.5 Estimated curves for females for the model (5.1)

As we can observe the model without different individual curves changes the estimated curves for each type of pathology because a same curve for all individuals is estimated instead of different smoothness for each of them.

**Model Comparison**

As it is described in Section 2.7 (Hypothesis testing in mix models) , the likelihood ratio test can be used for hypothesis tests for covariance matrix parameters.

$$\boldsymbol{H_0}: \boldsymbol{\sigma^2_{\tilde{u}^*}} = \boldsymbol{0} \ ; \ \boldsymbol{H_1}: \boldsymbol{\sigma^2_{\tilde{u}^*}} > 0$$

The Restricted log-likelihood value (RLRT) is 234.8164 and the p-value is < 2.2e-16 based on 10.000 simulated values for the male individuals and RLRT is 299.0287 with the p-value < 2.2e-16 again based on 10.000 simulated values for the female individuals (see Annex 1).

For both of the models there is an enough evidence to suggest that the null hypothesis is false which shows that the subject specific curves would be appropriate.

## 5.4 The model without the non-linear part of the factor by curve interactions

In model **(4.5)** it is assumed that the curves describing the effect of the pathology have different smoothness.

$$\boldsymbol{y_i = \alpha_0 + \alpha_1 pth_{iPKU} + f(t_i) + g_i(t_i) + \varepsilon_i} \qquad (\mathbf{5.2})$$

To test this assumption we compare the models **(4.5)** with **(5.2)** we investigate whether the random part of the subject specific curves is necessary for the model If the variance parameter of this random part ($\boldsymbol{\sigma^2_{\tilde{u}^*}}$) is zero it would mean that the random part of the deviation may not be necessary for the model, which means that the different smooth curves for each pathology type would not be appropriate to study the dataset.

So the hypothesis is;

$$\boldsymbol{H_0}: \boldsymbol{\sigma^2_w = 0} \qquad \boldsymbol{H_1}:\ \boldsymbol{\sigma^2_w} > 0$$

To apply the given model $(\mathbf{5.2})$ we introduce the matrices,

***R Code***

```
>X = model.matrix(z_score ~ age*Type.of.Pathology)

>Z.block = list(list(Id = pdIdent(~Z-1)),list(Patient.Id = pdSymm(~age)),list(Patient.Id
= +pdIdent(~Z.subject-1)))

>Z.block <- unlist(Z.block, recursive=FALSE)

>data.fr <- groupedData(z_score~ X[,-1]|Id, data = data.frame(z_score, X, Z,
Z.subject, +Patient.Id))

>fit3 <- lme(z_score ~ X[,-1], data = data.fr, random = Z.block)
```

The only difference between the model (4.5) and (5.2) is that we removed the Z.interact matrix from the model which indicates the different smooth curves for each type of pathology.

The numerical output for the applied **(5.2)** model for the boys shows that the estimated variance parameter $\boldsymbol{\sigma^2_{\tilde{u}}} = (\mathbf{0.1196})^{\mathbf{2}}$ . The variance parameter of the non-linear part of the individual curves is estimated $\boldsymbol{\sigma^2_{\tilde{u}^*}} = (\mathbf{0.3722})^{\mathbf{2}}$**.**The variance parameters, $\boldsymbol{\sigma^2_{\tilde{u}_{11}}} = (\mathbf{0.914})^{\mathbf{2}}$ which is the variance parameter of intercept and $\boldsymbol{\sigma^2_{\tilde{u}_{22}}} = (\mathbf{0.047})^{\mathbf{2}}$ is the variance parameter of the slope. Also there is a correlation between the random intercept and slope, which is estimated,$\boldsymbol{\gamma} = -\mathbf{0.332}$**.** The Akaike Information Criterion for this model is 1315.416.

The numerical output for the adjusted **(5.2)** model for the girls shows that the variance parameters are the following, $\boldsymbol{\sigma^2_{\tilde{u}}} = (\mathbf{0.77721})^{\mathbf{2}}$ , $\boldsymbol{\sigma^2_{\tilde{u}^*}} = (\mathbf{0.5430})^{\mathbf{2}}$ and the intercept and slope variance parameters are $\boldsymbol{\sigma^2_{\tilde{u}_{11}}} = (\mathbf{0.7224})^{\mathbf{2}}$**,** $\boldsymbol{\sigma^2_{\tilde{u}_{22}}} = (\mathbf{0.0364})^{\mathbf{2}}$ and the correlation is $\boldsymbol{\gamma} = -\mathbf{0.313}$**.** The Akaike Information Criterion for this model is 1523.141.

**Model Comparison**

The hypothesis test is

$$\boldsymbol{H_0}: \boldsymbol{\sigma_w^2 = 0} \qquad \boldsymbol{H_1}:\ \boldsymbol{\sigma_w^2} > 0$$

The Restricted log-likelihood value (RLRT) is 1 and the p-value is 1 based on 10.000 simulated values for the male individuals and RLRT is 3.1845 with the p-value 0.0186 again based on 10.000 simulated values for the female individuals. (see Annex 1)

For the model which is applied to male individuals there is no evidence to suggest that the null hypothesis is false which shows that the assumption of the different variance parameters for the deviation and for the time-effect of PKU would not be appropriate. In which we prefer the model without the random part of the deviation.

For the female individuals` model has a p-value 0.0186, therefore we don`t have a strong evidence to accept the null hypothesis. It shows us that the deviation of the children with HPA from the children with PKU have a variance parameter different form zero which says that the random part of the smoothing would be appropriate for this model.

## 5.5 The model without the random part of the non linear effect of age

As the random part of the deviation of the complete model is appropriate for the female individuals, but for the male ones not, now we have a different models for each gender. For the male ones we remove the random part of the deviation and for the female ones we keep it.

In this section we want to test if the smooth assumption of the effect of the age is appropriate for the model of the male and female individuals. We compare the model **(5.2)** with **(5.3)** for male individuals,

$$\boldsymbol{y_i = \alpha_0 + \alpha_1 pth_{iPKU} + \beta_0 + \beta_1 t_i + g_i(t_i) + \varepsilon_i} \qquad (\mathbf{5.3})$$

And for the female individuals we compare **(5.1)** with **(5.4)**

$$y_i = \alpha_0 + \alpha_1 pth_{iPKU} + \beta_0 + \beta_1 t_i + h(t_i) pth_{iPKU} + g_i(t_i) + \varepsilon_i \qquad (5.4)$$

The random part of the non linear effect of age has the following distribution;

$\tilde{u} \sim N(0, \sigma^2_{\tilde{u}})$ so we have to test whether this variance parameter is zero, which is equivalent to test the hypothesis,

$$H_0: \sigma^2_{\tilde{u}} = 0 \, , H_1: \sigma^2_{\tilde{u}} > 0$$

To apply the model **(5.3)**,

***R Code***

```
>X = model.matrix(z_score ~ age*Type.of.Pathology)

>Z.block = list(list(Patient.Id = pdSymm(~age)),list(Patient.Id = pdIdent(~Z.subject-1)))

>Z.block <- unlist(Z.block, recursive=FALSE)

>data.fr <- groupedData(z_score~ X[,-1]|Id, data = data.frame(z_score, X, Z, Z.subject, Patient.Id))

>fit4 <- lme(z_score ~ X[,-1], data = data.fr, random = Z.block)
```

To apply the model **(5.4)**,

```
>X <- model.matrix(z_score ~ Type.of.Pathology*age)

>Z.interact <- model.matrix(z_score ~ Z*Type.of.Pathology)[,-c(1,ncol(Z)+2)]

>Z.1 <- Z.interact[,1:ncol(Z)]

>Z.2 <- Z.interact[,(ncol(Z)+1):(2*ncol(Z))]

>Z.block <- list(Id = pdIdent(~Z.2-1), Patient.Id = pdSymm(~age), +Patient.Id = pdIdent(~Z.subject - 1))

>data.fr <- groupedData(z_score~ X[,-1]|Id, data = data.frame(z_score, X, Z, Z.subject, Patient.Id))
```

***>fit5 <- lme(z_score ~ X[,-1], data = data.fr, random = Z.block)***

The only difference between the model **(5.2)** and **(5.3)** is that we removed the non-linear part of the age effect which is indicated by ***Id = pdIdent(~Z-1))*** in R code. And the difference between **(5.1)** and **(5.4)** is that we removed the same part by ***Id = pdIdent(~Z.1-1).***

The numerical output for the applied **(5.3)** model for the boys shows that the estimated variance parameter of the non-linear part of the individual curves is estimated $\boldsymbol{\sigma^2_{\tilde{u}^*}} = (\mathbf{0.3954})^{\mathbf{2}}$.The variance parameters, $\boldsymbol{\sigma^2_{\tilde{u}_{11}}} = (\mathbf{0.9239})^{\mathbf{2}}$ which is the variance parameter of intercept and $\boldsymbol{\sigma^2_{\tilde{u}_{22}}} = (\mathbf{0.0479})^{\mathbf{2}}$ is the variance parameter of the slope. Also there is a correlation between the random intercept and slope, which is estimated,$\boldsymbol{\gamma = -0.336.}$ The Akaike Information Criterion for this model is 1316.336.

The numerical output for the applied **(5.4)** model for the girls shows that the variance parameters are the following ,$\boldsymbol{\sigma^2_w} = (\mathbf{1.095})^2$ $\boldsymbol{\sigma^2_{\tilde{u}^*}} = (\mathbf{0.5750})^{\mathbf{2}}$ and the intercept and slope variance parameters are $\boldsymbol{\sigma^2_{\tilde{u}_{11}}} = (\mathbf{0.7228})^{\mathbf{2}}$, $\boldsymbol{\sigma^2_{\tilde{u}_{22}}} = (\mathbf{0.035})^{\mathbf{2}}$ and the correlation is $\boldsymbol{\gamma = -0.3.}$ The Akaike Information Criterion for this model is 1553.136.

**Model Comparison**

$$\boldsymbol{H_0}: \boldsymbol{\sigma^2_{\tilde{u}}} = \mathbf{0}\,, \boldsymbol{H_1}: \boldsymbol{\sigma^2_{\tilde{u}}} > 0$$

The Restricted log-likelihood value (RLRT) is 2.92 and the p-value is 0.0319 based on 10.000 simulated values for the male individuals and RLRT is 33.1793 with the p-value 2.2e-16 again based on 10.000 simulated values for the female individuals. (see Annex 1)

For the both models for male and female individuals there is an evidence to reject the null hypothesis, therefore the non linear effect of age would be appropriate for models.

## 5.6 The model without the linear part of the interaction between age and type of pathology

In Section 5.4 according to the likelihood ratio test the assumption of factor by curve interactions would not be appropriate, therefore we removed the non linear part of the factor by curve interactions from the model for the male individuals. In this section, we will remove the linear part of this interaction from the model to test if the linear part of the interaction between age and pathology is appropriate for the data.

We remove the linear part of the interaction from the model (5.2) that we have fitted for the male individuals,

$$\boldsymbol{y_i = f(t_i) + g_i(t_i) + \varepsilon_i} \qquad (\mathbf{5.5})$$

```
> X = model.matrix(z_score ~ age+Type.of.Pathology)

> Z.block = list(list(Patient.Id = pdSymm(~age)),list(Patient.Id = pdIdent(~Z.subject-1)))

> Z.block <- unlist(Z.block, recursive=FALSE)

> data.fr <- groupedData(z_score~ X[,-1]|Id, data = data.frame(z_score, X, Z, Z.subject, Patient.Id))

> fit4.1 <- lme(z_score ~ X[,-1], data = data.fr, random = Z.block)
```

As can be observed we only changed the model matrix, and we removed the interaction between age and type of pathology.

We compare the two models with the Akaike Information Criterion,

AIC of model (5.2) = 1315.416

AIC of model (5.4) = 1310.022

We choose the model without the linear part of the interaction between the age and the type of pathology for the male individuals.

## 5.7 Conclusions

We have fitted four different models to model the z-scores of the children and we obtained the following results,

| | $\sigma^2_{\tilde{u}}$ | $\sigma^2_{w}$ | $\sigma^2_{\tilde{u}^*}$ | $\sigma^2_{\tilde{u}_{11}}$ | $\sigma^2_{\tilde{u}_{22}}$ | $\gamma$ | AIC |
|---|---|---|---|---|---|---|---|
| **Model 4.5** | $(0.1196)^2$ | $(0.000224)$ | $(0.3722)^2$ | $(0.9141)^2$ | $(0.047)^2$ | $-0.332$ | 1317.416 |
| **Model 5.1** | $(0.147)^2$ | $(0.01012)^2$ | - | $(0.8709)^2$ | $(0.0669)^2$ | $-0.32$ | 1550.223 |
| **Model 5.2** | $(0.1196)^2$ | - | $(0.3722)^2$ | $(0.914)^2$ | $(0.047)^2$ | $-0.332$ | 1315.416 |
| **Model 5.3** | - | - | $(0.395)^2$ | $(0.9239)^2$ | $(0.047)^2$ | $-0.336$ | 1316.336 |
| **Model 5.5** | $(0.1236)^2$ | - | $(0.3699)^2$ | $(0.924)^2$ | $(0.050)^2$ | $-0.339$ | 1310.022 |

**Table 5.3 :** The variance parameters and Akaike Information Criterion values of the each models fitted for male individuals.

| | $\sigma^2_{\tilde{u}}$ | $\sigma^2_{w}$ | $\sigma^2_{\tilde{u}^*}$ | $\sigma^2_{\tilde{u}_{11}}$ | $\sigma^2_{\tilde{u}_{22}}$ | $\gamma$ | AIC |
|---|---|---|---|---|---|---|---|
| **Model 4.5** | $(0.8128)^2$ | $(0.4529)^2$ | $(0.5180)^2$ | $(0.7249)^2$ | $(0.0368)^2$ | $-0.32$ | 1521.956 |
| **Model 5.1** | $(0.7298)^2$ | $(0.4413)^2$ | - | $(0.7840)^2$ | $(0.1046)^2$ | $-0.337$ | 1818.985 |
| **Model 5.2** | $(0.7772)^2$ | - | $(0.5430)^2$ | $(0.7224)^2$ | $(0.0364)^2$ | $-0.313$ | 1523.141 |
| **Model 5.3** | - | - | $(0.644)^2$ | $(0.7067)^2$ | $(0.034)^2$ | $-0.242$ | 1580.407 |
| **Model 5.4** | | $(1.095)^2$ | $(0.5750)^2$ | $(0.7228)^2$ | $(0.035)^2$ | $-0.3$ | 1553.136 |

**Table 5.4 :** The variance parameters and Akaike Information Criterion values of the each models fitted for female individuals.

To test the non-linear parts of the models we used the likelihood ratio test to contrast if the variance parameters are different from zero. Table 5.5 and 5.6 shows the results for each model testing,

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

**Table 5.5** Likelihood ratio test statistics values for each model testing for the male individuals.

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

**Table 5.6** Likelihood ratio test statistics values for each model testing for the female individuals.

The p-values of RLRT test are obtained using "RLRsim" package implemented in R by **Fabian Scheipl and Ben Bolker (2008)** based on 10.000 simulations.

As conclusion, the appropriate model (5.5) for the male individuals includes just one curve over time for the mean constructed by subject specific curves for each individual and its results are the following;

$$\boldsymbol{y_i} \ = \boldsymbol{f(t_i)} + \boldsymbol{g_i(t_i)} + \boldsymbol{\varepsilon_i} \qquad (\mathbf{5.5})\ .$$

| | **Value** | **Std.Error** | **DF** | **t-value** | **p-value** |
|---|---|---|---|---|---|
| **(Intercept)** | 0.4536873 | 0.19720838 | 902 | 2.300548 | 0.0216 |
| **Time** | -0.0175212 | 0.01333553 | 902 | -1.313873 | 0.1892 |
| **Type.of.PathologyPKU** | -0.3759156 | 0.27041493 | 42 | -1.390144 | 0.1718 |

**Table 5.7** Estimations of the fixed effects from the model (5.5) for male individuals.

As we observe the fixed effect of the time and type of pathology are not statistically significant, but we keep them in the model because the corresponding random parts are necessary for the model.

And the appropriate model (4.5) for the female individuals includes different curves for each type of pathology and also subject specific curves and its results are;

$$\boldsymbol{y_i = \alpha_0 + \alpha_1 pth_{iPKU} + f(t_i) + h(t_i) pth_{iPKU} + g_i(t_i) + \varepsilon_i} \qquad (\mathbf{4.5})$$

|  | Value | Std.Error | DF | t-value | p-value |
|---|---|---|---|---|---|
| **(Intercept)** | 0.9138604 | 0.21225067 | 1125 | 4.305571 | 0.0015 |
| **Time** | -0.0896243 | 0.01892408 | 1125 | -4.735995 | 0.0398 |
| **Type.of.PathologyPKU** | -0.1465574 | 0.19003366 | 63 | -0.771218 | 0.8889 |
| **Type.of.PathologyPKU:time** | -0.017258 | 0.03790231 | 1124 | -0.455337 | 0.6490 |

**Table 5.8** Estimations of the fixed effects from the model (5.1) for female individuals.

We observe almost the same results for the female individuals that the fixed part of the effect of the type of pathology and its interaction with time are not significant but we keep them because their corresponding random effects are necessary for the model.

## REFERENCES

1. Brumback BA, Rice JA. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* 1998; 93:961—994.
2. Brumback BA, Ruppert D, Wand MP. Commenton Variable selection and function estimation in additive nonparametric regression using a data-based prior. Journal of the American Statistical Association 1999; 94:794—797.
3. Brown, H. Prescott, R. Applied Mixed Models in Medicine. John Wiley&Sons 2006.
4. Currie ID, Durbán M. Flexible smoothing with P-splines: a unified approach. Statistical Modelling 2002; 2:333—349.
5. Durbán, M., Harezlak, J., Wand, M. and Carroll, R. Simple Fitting of Subject Specific Curves for Longitudinal Data. Statistics in Medicine 2005; 24:1153-1162
6. De Boor, Carl. A practical guide to splines. Springer 2001
7. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. Statistical Science 1996; 11:89—121.
8. Friedman, J. H. & B. W. Silverman. Flexible parsimonious smoothing and additive modeling. *Technometrics* 1989; 31: 3–21
9. Green P.J., Silverman B.W. Non parametric regression and Generalized Linear Models. Chapmann&Hall 1994.
10. Guo W. Functional mixed effects models. Biometrics 2002; 58:121—128.
11. Harville, D.A. Maximum likelihood approaches to variance components estimation and to related problems. Journal of *the American Statistical Association* 1977; 72:320‑340.
12. Hastie, T.J. and Tibshirani, R.J. *Generalized Additive Models*, New York: Chapman and Hall 1990.
13. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982; 38:963—974.
14. Lee, D-J Smoothing mixed models for spatial and spatio-temporal data. P.h.d Thesis 2010.
15. Lee, D-J. and Durbán, M. Smooth-Car mixed models for spatial count data Computational Statistics and data Analysis 2009; 53, 2968-2977.
16. Patterson, H. and Thompson, R. Recovery of interblock information when the block sizes are unequal., *Biometrika* 1971; **58** 545-554.
17. Pinheiro Jose C., Bates Douglas M. Mixed-effects models in S and S-plus. Springer, 2000

18. Rice JA, Wu CO. Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 2001; 57:253—259.
19. Ruppert D, Wand MP, Carroll RJ. Semiparametric Regression. Cambridge University Press: Cambridge, 2003.
20. Scheipl, F. RLRsim: Exact (Restricted) Likelihood Ratio tests for mixed and additive models. R package RLRsim 2008; 2.0-2.
21. Stram DO, Lee JW. Variance components testing in the longitudinal mixed effects model. Biometrics 1994; 50:1171—1177.
22. Speckman, P. Kernel smoothing in partially linear models. *Journal of the Royal Statistical Society* 1988; Series B 50, 413-436.
23. Verbeke, G. Molenbergs G. Linear Mixed Models in practice Springer 1997.
24. Wahba, G. Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression. *Journal of the Royal Statistical Society* 1978; 40: 364-372
25. Wood, N. Thin plate splines regression. *Journal of the Royal Statistical Society* 2003; 65(1):95-114
26. Wood, S.N. Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American Statistical Association* 2004; 99, 673-686
27. Wood, S.N. Fast stable direct fitting and smoothness selection for Generalized Additive Models. *Journal of the Royal Statistical Society* 2008; 70(3), 495-518.
28. Zhang D, Lin X, Raz J, Sowers M. Semiparametric stochastic mixed models for longitudinal data. Journal of the American Statistical Association 1998; 93:710—719.

# Annex 1

## *R Code for the RLRT test.*

## *Package RLRsim*

```
####################################################
##    FOR MALE  AND FEMALE INDIVIDUALS
####################################################
> attach(data.male) /  attach(data.female)
>Id <- factor(rep(1, length = length(z_score)))
# Numero de nodos optimo
>K <- max(5,min(floor(length(unique(time))/4),40))
>MM = MM.basis(time, min(time)- 0.5, max(time) + 0.5, K, 3, 2)
>Z = MM[[2]]
>MM.subject <- MM.basis(time, min(time)- 0.5, max(time) + 0.5, K, 3, 2)
>Z.subject <- MM.subject[[2]]
>n = length(z_score)
########################################
#######################################


#  FIRST HYPOTHESIS:
#  H0: Model without the non linear part of the subject specific curves
#  H1: Complete Model
#  m : Model with only the fixed effects and the random part of the subject specific curves
#############
##COMPLETE MODEL
############
>X <- model.matrix(z_score ~ Type.of.Pathology*time)
>XX <- X[,-1]
>Z.interact <- model.matrix(z_score ~ Z*Type.of.Pathology)[,-c(1,ncol(Z)+2)]
>Z.1 <- Z.interact[,1:ncol(Z)]
```

```
>Z.2 <- Z.interact[,(ncol(Z)+1):(2*ncol(Z))]

>Z.block <- list(Id = pdIdent(~Z.1-1), Id = pdIdent(~Z.2-1), Patient.Id = pdSymm(~time), Patient.Id =
+pdIdent(~Z.subject - 1))

>data.fr <- groupedData(z_score ~ XX|rep(1, length = length(z_score)), data = data.frame(z_score, X, Z.1, Z.2,
+Z.subject, Patient.Id, Type.of.Pathology))


>fit1 <- lme(z_score ~ XX, random = Z.block, data = data.fr)

>summary(fit1)

############

##MODEL WITHOUT SUBJECT SPECIFIC CURVES

############

>X <- model.matrix(z_score ~ Type.of.Pathology*time)

>XX <- X[,-1]

>Z.interact <- model.matrix(z_score ~ Z*Type.of.Pathology)[,-c(1,ncol(Z)+2)]

>Z.1 <- Z.interact[,1:ncol(Z)]

>Z.2 <- Z.interact[,(ncol(Z)+1):(2*ncol(Z))]


>Z.block <- list(Id = pdIdent(~Z.1-1), Id = pdIdent(~Z.2-1), Patient.Id = pdSymm(~time))

>data.fr <- groupedData(z_score ~ XX|rep(1, length = length(z_score)), data = data.frame(z_score, X, Z.1, Z.2,
+Patient.Id, Type.of.Pathology))

>fit2 <- lme(z_score ~ XX, random = Z.block, data = data.fr)

>summary(fit2)

#############

##MODEL "m"

############

>X <- model.matrix(z_score ~ Type.of.Pathology*time)

>XX <- X[,-1]

>Z.block <- list(Patient.Id = pdIdent(~Z.subject - 1))

>data.fr <- groupedData(z_score ~ XX|rep(1, length = length(z_score)), data = data.frame(z_score, X, Z.subject,
+Type.of.Pathology))

>fitm <- lme(z_score ~ XX, random = Z.block, data = data.fr)

>summary(fitm)
```

```
###############################################

###  RLRT TEST  ########################

###############################################


>library(RLRsim)

>exactRLRT(m=fitm, mA=fit1, m0=fit2)


############################################################################
#########

############################################################################
######

#  SECOND HYPTOHESIS:

#  H0: Model without the non linear part of the factor by curve interactions

#  H1: Complete Model

#  m : Model with only the fixed effects and the random part of the factor by curve interactions


############################################################################
#########

#############

##COMPLETE MODEL

############

>X <- model.matrix(z_score ~ Type.of.Pathology*time)

>XX <- X[,-1]

>Z.interact <- model.matrix(z_score ~ Z*Type.of.Pathology)[,-c(1,ncol(Z)+2)]

>Z.1 <- Z.interact[,1:ncol(Z)]

>Z.2 <- Z.interact[,(ncol(Z)+1):(2*ncol(Z))]

>Z.block <- list(Id = pdIdent(~Z.1-1), Id = pdIdent(~Z.2-1), Patient.Id = pdSymm(~time), Patient.Id =
+pdIdent(~Z.subject - 1))

>data.fr <- groupedData(z_score ~ XX|rep(1, length = length(z_score)), data = data.frame(z_score, X, Z.1, Z.2,
+Z.subject, Patient.Id, Type.of.Pathology))

>fit1 <- lme(z_score ~ XX, random = Z.block, data = data.fr)

>summary(fit1)
```

```
#############

##MODEL WITHOUT FACTOR BY CURVE INTERACTIONS

############

>X = model.matrix(z_score ~ time*Type.of.Pathology)

>XX <- X[,-1]

>Z.block = list(list(Id = pdIdent(~Z-1)),list(Patient.Id = pdSymm(~time)),list(Patient.Id = pdIdent(~Z.subject-1)))

>Z.block <- unlist(Z.block, recursive=FALSE)

>data.fr <- groupedData(z_score~ XX|Id, data = data.frame(z_score, X, Z, Z.subject, Patient.Id))

>fit2 <- lme(z_score ~ XX, data = data.fr, random = Z.block)

#############

##MODEL "m"

############

>X <- model.matrix(z_score ~ Type.of.Pathology*time)

>XX <- X[,-1]

>Z.block <- list(Id = pdIdent(~Z.2-1))

>data.fr <- groupedData(z_score ~ XX|rep(1, length = length(z_score)), data = data.frame(z_score, X, Z.1,Z.2,
+Patient.Id, Type.of.Pathology))

>fitm <- lme(z_score ~ XX, random = Z.block, data = data.fr)

>summary(fitm)

###############################################

###  RLRT TEST  ########################

###############################################

>library(RLRsim)

>exactRLRT(m=fitm, mA=fit1, m0=fit2)

########################################

#######################################
```

```
#  THIRD HYPOTHESIS:

#  H0: Model without the non linear part of the f(time)

#  H1: Complete Model

#  m : Model with only the fixed effects and the random part of the f(time)

#############################################################################
########

#############

##COMPLETE MODEL

############

>X = model.matrix(z_score ~ time*Type.of.Pathology)

>XX <- X[,-1]

>Z.block = list(list(Id = pdIdent(~Z-1)),list(Patient.Id = pdSymm(~time)),list(Patient.Id = pdIdent(~Z.subject-1)))

>Z.block <- unlist(Z.block, recursive=FALSE)

>data.fr <- groupedData(z_score~XX|rep(1, length = length(z_score)), data = data.frame(z_score, X, Z, Z.subject,
+Patient.Id))

>fit1 <- lme(z_score ~ XX, data = data.fr, random = Z.block)

>summary(fit1)

#############################################################################
######

#############

##MODEL WITHOUT f(time)

############

>X = model.matrix(z_score ~ time*Type.of.Pathology)

>XX <- X[,-1]

>Z.block = list(list(Patient.Id = pdSymm(~time)),list(Patient.Id = pdIdent(~Z.subject-1)))

>Z.block <- unlist(Z.block, recursive=FALSE)

>data.fr <- groupedData(z_score~ XX|rep(1, length = length(z_score)), data = data.frame(z_score, X, Z, Z.subject,
+Patient.Id))

>fit2 <- lme(z_score ~ XX, data = data.fr, random = Z.block)
```

```
#############

##MODEL "m"

############

>X <- model.matrix(z_score ~ time*Type.of.Pathology)

>XX <- X[,-1]

>Z.block <- list(list(Id = pdIdent(~Z-1)))

>Z.block <- unlist(Z.block, recursive=FALSE)

>data.fr <- groupedData(z_score ~ XX|Id, data = data.frame(z_score, X, Z))

>fitm <- lme(z_score ~ XX, random = Z.block, data = data.fr)

>summary(fitm)

###############################################

###  RLRT TEST  ########################

###############################################

>library(RLRsim)

>exactRLRT(m=fitm, mA=fit1, m0=fit2)

############################################################################
######

###### TESTING NON LINEAR PART OF THE INTERACTION ##########

>X = model.matrix(z_score ~ time*Type.of.Pathology)

>XX <- X[,-1]

>Z.block = list(list(Id = pdIdent(~Z-1)),list(Patient.Id = pdSymm(~time)),list(Patient.Id = pdIdent(~Z.subject-1)))

>Z.block <- unlist(Z.block, recursive=FALSE)

>data.fr <- groupedData(z_score~XX|rep(1, length = length(z_score)), data = data.frame(z_score, X, Z, Z.subject,
+Patient.Id))

>fit1 <- lme(z_score ~ XX, data = data.fr, random = Z.block)

>summary(fit1)

########################

>X = model.matrix(z_score ~ time+Type.of.Pathology)

>XX <- X[,-1]

>Z.block = list(list(Id = pdIdent(~Z-1)),list(Patient.Id = pdSymm(~time)),list(Patient.Id = pdIdent(~Z.subject-1)))

>Z.block <- unlist(Z.block, recursive=FALSE)
```

```
>data.fr <- groupedData(z_score~XX|rep(1, length = length(z_score)), data = data.frame(z_score, X, Z, Z.subject,
+Patient.Id))

>fit2 <- lme(z_score ~ XX, data = data.fr, random = Z.block)

>summary(fit2)

>anova(fit1,fit2)

###################################################
```