Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Nordhausen et al., J Biom Biostat 2015, 6:3
DOI: 10.4172/2155-6180.1000239

Research Article Open Access

Mixed-Effects Regression Splines to Model Myopia Data

Nordhausen K[1,2*], Oja H[1] and Pärssinen O[3]

[1]Department of Mathematics and Statistics, University of Turku 20014 Turku, Finland
[2]School of Health Sciences, University of Tampere 33014 Tampere, Finland
[3]Ophthalmic Department, Central Hospital of Central Finland 40620 Jyväskylä, Finland

Abstract

Myopia is a disorder of ocular refraction with varying rates of progression. Although the disorder has a dynamic nature, prospective longitudinal studies with long term follow-ups have been remarkably few. In this paper, we show how mixed-effects regression splines with different choices of basis functions can be used to model myopia progression data in a flexible way. We show how the estimated model may be used to find prediction curves with corresponding confidence and tolerance intervals for a new myopic subject. We discuss alternative choices of the basis functions such as the truncated polynomial spline functions (2 types) and B-spline functions. Principal component functions may be used for an analysis of the variation of the curves in the population. The theory is collected together and presented in a coherent way as well as illustrated with a careful analysis of myopia progression data from a Finnish myopia study.

Keywords: Basis function; B-spline; Linear mixed model; Prediction curve; Principal curve; Progression; Truncated polynomial spline

Introduction

Myopia, also known as nearsightedness, is defined as that state of ocular refraction in which parallel rays of light entering the eye at rest are brought to focus in front of the retina. In this situation, distant objects cannot be perceived distinctly. Myopia can occur at any age but in most cases it appears at school age and usually progresses several years, in few cases progression continues throughout life. In practice myopia is treated either by prescribing corrective lenses, such as glasses or contact lenses, or by performing refractive surgery. For people with myopia and for those who treat myopia and prescribe glasses, reliable predictions of myopia progression as early as possible would be extremely valuable.

The prevalence of myopia has markedly increased during recent decades in many countries. This increase is hard to explain only by hereditary factors. Epidemiological studies have confirmed that a longer education and higher occupational status, for example, are connected with an increased prevalence of myopia [1]. Many studies have shown that the younger the age of onset the faster is myopia progression [2]. Based on a small sample, Thorn et al. [3] estimated that myopia beginning in childhood "stabilizes" in 80% of cases by about 19-years of age. There has been an increased interest in long term follow-up studies to model the progression of myopia. The estimated models with possible covariates then often provide individual prediction curves as well. One may also wish to test, for example, whether the progression of myopia levels off at a certain age.

The dynamic progression of myopia has not been widely studied in the epidemiological or biostatistical literature. The studies mainly consider the individual progression curves from retrospective material or by means of progression in different materials [4]. Möttönen at al. [5] suggested to model myopia using a polynomial (quadratic or cubic) random coefficient model. The model is not satisfactory, however, as even cubic myopia progression curves are not flexible enough to explain strong changes between 10 and 20 years. In this paper we will show that mixed-effects regression analysis using quadratic or cubic splines is a flexible tool to model the progression of myopia that easily provides individual prediction curves. Furthermore, we show that different important questions concerning the progression of myopia may be answered simply by changing the spline basis functions. For a general theory for mixed-effects regression splines, see e.g. [6,7]. One of the aims of this paper is to collect and present the theory in an easy and coherent way for the users. Mixed-effects regression spline models have recently been applied to CD4 counts [8] and to the growth of cattle [9]. However, these models have never before been applied to myopia progression data.

The paper is structured as follows. In Section 2 we discuss the use of the random effect regression model in the case that subject i, i=1, ..., n, has p_i measurements at time points t_{i1} , ..., t_{ip_i} . Note that the number of measurements as well as the time points vary individually as is always the case for myopia follow-up data. Each subject is then supposed to have his/her own myopia progression curve as a function of time. This curve is observed only through the measurements at the p_i time points (with measurement errors coming from N $(0, \tau^2)$). The individual curves are linear combinations of k basis curves, and the k coefficients are assumed to come from a k-variate normal distribution with an unknown mean vector depending on parameter θ and unknown covariance matrix Σ. Parameters θ and Σ (providing the mean curve and variation of the curves) are the population quantities we are interested in. We discuss the estimation of parameters θ, Σ, and τ^2 and show how the estimated parameters may be used to find prediction curves with corresponding confidence and tolerance intervals. In Section 3 we discuss alternative choices of the basis functions; a special interest is in the set of principal component functions that can be used for a careful analysis of the variation of the curves in the population. In Section 4 we discuss the use of spline functions, truncated polynomial spline functions (2 types) and B-spline functions, as basis functions. In Section 5 the theory is then illustrated with a real data set from a Finnish longitudinal myopia study. The paper ends with some final comments in Section 6.

***Corresponding author:** Nordhausen K, Department of Mathematics and Statistics, University of Turku 20014 Turku, Finland, Tel: +358 2 333 5441; E-mail: klaus.nordhausen@utu.fi

Received June 01, 2015; **Accepted** July 13, 2015; **Published** July 20, 2015

Citation: Nordhausen K, Oja H, Pärssinen O (2015) Mixed-Effects Regression Splines to Model Myopia Data. J Biom Biostat 6: 239. doi:10.4172/2155-6180.1000239

Random Coefficient Regression Model

Random coefficient regression model using k basis functions

Let $f_1(t),...,f_k(t)$ be the selected k basis functions for myopia regression curve. We write

$$f(t) = (f_1(t),...,f_k(t))'$$

for the corresponding vector valued function. We assume that individual i has p_i observations

$$y_i = (y_{i1},...,y_{ip_i})'$$

at time points t_{i1} , ..., t_{ip_i}, i=1, ..., n. We then write

$$F_i = \begin{pmatrix} f_1(t_{i1}) & f_2(t_{i1}) & ... & f_k(t_{i1}) \\ f_1(t_{i2}) & f_2(t_{i2}) & ... & f_k(t_{i2}) \\ ... & ... & ... & ... \\ f_1(t_{ip_i}) & f_2(t_{ip_i}) & ... & f_k(t_{ip_i}) \end{pmatrix}$$

for the $p_i \times$ k time design matrix for individual i, i=1, ..., n. The r-variate covariate vector $x_i=(x_{i1}, ..., x_{ir})'$ is used to explain the myopia progression curve of the ith individual, i=1, ..., n. Finally, write $N = \sum_{i=1}^{n} p_i$ for the total number of measurements.

We then make the following model assumption.

Assumption 1:

(I) The myopia progression curve for individual i is

$$b_i' f(t) = \sum_{j=1}^{k} b_{ij} f_j(t)$$

where

$$b_i = \Theta x_i + \xi_i,$$

Θ is a k × r-variate matrix of regression coefficients, and $\xi_i \sim N_k(0, \Sigma)$, i=1, ..., n.

(II) For individual i, the observed refraction values are

$$y_i = F_i b_i + \epsilon_i,$$

where

$\epsilon_i \sim N_{p_i}(0, \tau^2 I_{p_i})$, i=1, ..., n.

(III) The random variables $\xi_1,...,\xi_n$ and $\epsilon_1,...,\epsilon_n$ are all mutually independent.

The parameter Θ stands for the connection between x_i and the myopia progression curve for the ith individual, Σ shows the variation of the curves (in the population of the progression curves), and τ^2 tells about the random variation of the measurements around the individuals curves. Note that, if we collect the assumptions together, the model can be also written as a mixed model

$$y_i = (X_i' \otimes F_i) vec(\Theta) + F_i \xi_i + \epsilon_i, \quad i = 1,...,n,$$

where ⊗ denotes the Kronecker product. Then $X_i' \otimes F_i$ is the matrix of fixed effects and F_i is the matrix of random effects for individual i, i=1, ..., n. If we further write $X_i = X_i' \otimes F_i$, $\theta = vec(\Theta)$, and $V_i = F_i \Sigma F_i' + \tau^2 I_{p_i}$ then

$$y_i \sim N_{p_i}(X_i\theta, V_i)$$

and, if Σ and τ were known, then the maximum likelihood (ML) estimate of θ, namely

$$\hat{\theta} = \hat{\theta}(\Sigma, \tau) = \left\{ \sum_{i=1}^{n} X_i' V_i^{-1} X_i \right\}^{-1} \sum_{i=1}^{n} X_i' V_i^{-1} y_i$$

has a multivariate normal distribution

$$N_{kr}\left(\theta, \left\{ \sum_{i=1}^{n} X_i' V_i^{-1} X_i \right\}^{-1} \right).$$

The maximum likelihood (ML) estimates of the parameters θ, Σ, and τ^2 can be found by minimizing the -2 times the log-likelihood function

$$-2\log L = \log(2\pi)N + \sum_{i=1}^{n} \log|V_i| + \sum_{i=1}^{n} (y_i - X_i\theta)' V_i^{-1} (y_i - X_i\theta).$$

The minimization can be done using the expectation maximization (EM) algorithm or other optimization routines. For more details about basis functions and their use in mixed models [6,7].

Prediction curves based on the model

Throughout this section we assume that we have a model with estimated parameters $\hat{\theta}$, $\hat{\Sigma}$ and $\hat{\tau}^2$ their estimated variances and covariances. Consider first the prediction curve for a new individual from the same population with a covariate vector x, throughout this section we are interested in the curves on some time interval $[t_0, t_1]$ only. The mean progression curve of the new individual is then

$$t \rightarrow (\Theta x)' f(t) = (x \otimes f(t))'\theta .$$

Consider next the estimate of the mean progression curve with its pointwise 100(1-α)% confidence interval as well as its 100(1-α)% confidence band. One also often wishes to estimate the 100(1-α)% point wise tolerance interval and the 100(1-α)% tolerance band for the progression curve. These are found as follows. Notice that the confidence intervals and bands in 2 and 3 are based on the joint limiting normal distribution of the estimates and therefore only approximate.

The mean progression curve $(x \otimes f(t))'\theta$ is estimated by $(x \otimes f(t))'\hat{\theta}$.

An approximate 100(1-α)% pointwise confidence interval for the mean progression curve $(x \otimes f(t))'\theta$ at time t is given by

$$(x \otimes f(t))'\hat{\theta} \pm c_{1-\alpha} \sqrt{(x \otimes f(t))' \mathrm{Cov}(\hat{\theta})(x \otimes f(t))}.$$

1. Notice that pointwise confidence intervals do not give a confidence band for the (whole) mean progressive curve over all possible values of t. The confidence band is considered next.

An approaximate 100(1-α)% confidence band for the mean progression curve $(x \otimes f(t))'\theta$ is given by

$$(x \otimes f(t))'\hat{\theta} \pm c_{1-\alpha} \sqrt{(x \otimes f(t))' \mathrm{Cov}(\hat{\theta})(x \otimes f(t))}.$$

where $c_{1-\alpha}$ is given by

$$P\left\{ \sup_{t \in [t_0, t_1]} \frac{((x \otimes f(t))'\tilde{\theta})^2}{(x \otimes f(t))' \mathrm{Cov}(\hat{\theta})(x \otimes f(t))} \le c_{1-\alpha}^2 \right\} = 1 - \alpha$$

where $\tilde{\theta} \sim N_{kr}(0, \mathrm{Cov}(\hat{\theta}))$. Note that $c_{1-\alpha}$ depends on limits t_0 and t_1 but can be easily found by simulation. Note also that $c_{1-\alpha}^2 \le \chi_{kr,1-\alpha}^2$.

An estimate for 100(1-α)% pointwise tolerance interval for

the individual value of the refraction curve at time t, that is, $b'f(t) = (\Theta x + \xi)' f(t)$ (in the subpopulation of individuals having the same covariate vector value x) is given by

$$(x \otimes f(t))'\hat{\theta} \pm \chi_{1,1-\alpha} \sqrt{f(t)' \hat{\Sigma} f(t)}.$$

Finally an estimate for 100(1-α)% tolerance band for $b'f(t) = (\Theta x + \xi)' f(t)$ (in the subpopulation of individuals having the same covariate vector value x) is given by

$$(x \otimes f(t))'\hat{\theta} \pm c_{1-\alpha} \sqrt{f(t)' \hat{\Sigma} f(t)}$$

where $c_{1-\alpha}$ is now determined by

$$P\left\{ \sup_{t \in [t_0, t_1]} \frac{(f'(t)\xi)^2}{f(t))' \hat{\Sigma} f(t)} \le c_{1-\alpha}^2 \right\} = 1 - \alpha$$

where $\xi \sim N_k(0, \hat{\Sigma})$. Now $c_{1-\alpha}^2 \le \chi_{k,1-\alpha}^2$.

Assume next that the parameters in the model are known, and we predict the mean progression curve of individual i with covariate vector x_i, that is, the curve

$$t \rightarrow f(t)' b_i = (x_i \otimes f(t))'\theta + f(t)' \xi_i.$$

The prediction is based on observation vector $y_i = (y_{i1}, \ldots, y_{ip_i})'$. Then

$$\begin{pmatrix} b_i \\ y_i \end{pmatrix} \sim N_{k+p_i} \left(\begin{pmatrix} (X_i' \otimes I_k)\theta \\ X_i \theta \end{pmatrix}, \begin{pmatrix} \Sigma & \Sigma F_i' \\ F_i \Sigma & V_i \end{pmatrix} \right)$$

and

$$b_i \mid y_i \sim N_k \left((X_i' \otimes I_k)\theta + \Sigma F_i' V_i^{-1} (y_i - X_i \theta), \Sigma - \Sigma F_i' V_i^{-1} F_i \Sigma \right).$$

Then we have the following estimates and predictions.

1. The predicted progression curve for $b_i' f(t)$ is $\hat{b}_i' f(t)$ where

$$\hat{b}_i = (X_i' \otimes I_k)\theta + \Sigma F_i' V_i^{-1} (y_i - X_i \theta)$$

2. The pointwise 100 (1-α)% tolerance interval for $b_i' f(t)$ is

$$f(t)' \hat{b}_i \pm \chi_{1,1-\alpha} \sqrt{f(t)' \mathrm{Cov}(b_i \mid y_i) f(t)}$$

3. The 100(1-α)% tolerance band for $b_i' f(t)$ is given by

$$f(t)' \hat{b}_i \pm c_{1-\alpha} \sqrt{f(t)' \mathrm{Cov}(b_i \mid y_i) f(t)}$$

where $c_{1-\alpha}$ is determined by

$$P\left\{ \sup_{t \in [t_0, t_1]} \frac{(f'(t)\tilde{b})^2}{f(t)' \mathrm{Cov}(b_i \mid y_i) f(t)} \le c_{1-\alpha}^2 \right\} = 1 - \alpha$$

where $\tilde{b} \sim N_k(0, \mathrm{Cov}(b_i \mid y_i))$. Now $c_{1-\alpha}^2 \le \chi_{k,1-\alpha}^2$.

The Choice of the Basis Curves

Alternative choices of the basis curves

Let us assume that the model fitting is performed using k basis functions $f_1, \ldots, f_k$ but, for the interpretation purposes, we wish to present the results using another set of basis functions $g_1, \ldots, g_k$. Assume also that these two sets of basis functions span the same set of progression functions, that is,

$$\left\{ \sum_{j=1}^k \beta_j f_j(t) : \beta_1, \ldots, \beta_k \in \mathbb{R} \right\} = \left\{ \sum_{j=1}^k \beta_j g_j(t) : \beta_1, \ldots, \beta_k \in \mathbb{R} \right\}.$$

Then, if we write

$$f(t) = (f_1(t), \ldots, f_k(t))' \quad \text{and} \quad g(t) = (g_1(t), \ldots, g_k(t))',$$

there exists a full-rank k × k matrix A such that

f (t)=Ag(t), for all t.

Then

$$G_i = \begin{pmatrix} g_1(t_{i1}) & g_2(t_{i1}) & \ldots & g_k(t_{i1}) \\ g_1(t_{i2}) & g_2(t_{i2}) & \ldots & g_k(t_{i2}) \\ \ldots & \ldots & \ldots & \ldots \\ g_1(t_{ip_i}) & g_2(t_{ip_i}) & \ldots & g_k(t_{ip_i}) \end{pmatrix} = F_i (A')^{-1}$$

and the model is transformed

$$y_i = F_i b_i + \epsilon_i \rightarrow y_i = G_i b_i^* + \epsilon_i$$

where

$$b_i^* = A' b_i \sim N_k(A'\Theta x_i, A'\Sigma A).$$

The parameters are then transformed in the following way

$$\Theta \rightarrow A'\Theta, \theta \rightarrow (I_r \otimes A')\theta, \text{ and } \Sigma \rightarrow A'\Sigma A$$

Recall that, in the ML estimation, the parameter estimates are transformed in the same way.

Principal component analysis for the variation of the curves

One popular set of basis functions are principal component functions which often can be interpreted and are also ordered according to the amount of variation they explain.

Consider the random functions

$$h(t) = f(t)' b$$

where, as before, $f(t) = (f_1(t), \ldots, f_k(t))'$ is a vector valued function and $b \sim N_k(\mu, \Sigma)$. Then we define the principal component functions $\tilde{f}_1(t), \ldots, \tilde{f}_k(t)$ and corresponding scores $z_1, \ldots, z_k$ as follows.

1. Write Σ_f for the positive definite k × k matrix with the elements

$$(\Sigma_f)_{ij} = \int f_i(t) f_j(t) dt, \; i, j = 1, \ldots, k.$$

2. Find the eigenvector eigenvalue decomposition

$$\Sigma_f^{1/2} \Sigma \Sigma_f^{1/2} = U \Lambda U',$$

where Λ is a diagonal matrix containing the eigenvalues of $\Sigma_f^{1/2} \Sigma \Sigma_f^{1/2}$ in decreasing order and U is an orthogonal matrix having has i^{th} column the eigenvector corresponding to the ith eigenvalue in Λ.

3. Write

$$z = U' \Sigma_f^{1/2} (b - \mu) \sim N_k(0, \Lambda).$$

4. Write

$$\tilde{f}(t) = U' \Sigma_f^{-1/2} f(t).$$

5. Then

$$h(t) - E(h(t)) = \tilde{f}(t)' z = \tilde{f}_1(t) z_1 + \ldots + \tilde{f}_k(t) z_k$$

where

$$\int \tilde{f}_i(t)\tilde{f}_j(t)dt = \delta_{ij},\ \ i,j = 1,...,k,$$

and therefore

$$\int (h(t) - E(h(t)))^2 dt = z_1^2 \int \tilde{f}_1(t)^2 dt + ... + z_k^2 \int \tilde{f}_k(t)^2 dt$$
$$= z_1^2 + ... + z_k^2.$$

The magnitude of an eigenvalue related to a principal component function reflects its relevance and the associated score of it how strong expressed the feature represented by this function is, similar as in traditional PCA. For further ideas about the decomposition of the random effects covariance matrix, see for example [10].

Mixed-effects regression splines

In this section we consider spline functions which provide a flexible basis for smooth myopia progression modeling [6,11,12].

Truncated polynomial spline functions

We first consider the splines based on truncated power functions. For the definition, let $t_0 < t_1 < ... < t_m < t_{m+1}$ be the m+2 time points (knots) and assume that all the measurements are on the interval $[t_0, t_{m+1}]$. Write

$$(t)_+ = \max\{t, 0\} \text{ and } (t)_- = \min\{t, 0\}.$$

The truncated linear spline function is based on m+2 basis functions

$f_{0,1}(t) \equiv 1$, $f_{1,1}(t)=t$, $f_{2,1}(t)=(t - t_1)_+$,... , $f_{m+1,1}(t)=(t - t_m)_+$.

The truncated quadratic spline functions has a basis of m+3 functions

$$f_{0,2}(t) \equiv 1,\ f_{1,2}(t) = t,\ f_{2,2,}(t) = t^2, f_{3,2}(t) = (t - t_1)_+^2, ..., f_{m+2,2}(t) = (t - t_m)_+^2.$$

Finally the splines based on truncated polynomials of degree p are given by

$f_{k,p} = t^k$, $k = 0,..., p$, and $f_{p+k,p}(t) = (t - t_k)_+^p$, k=1, ..., m.

The function space spanned by these functions

$$\mathcal{F}_p = \left\{ \sum_{k=0}^{m+p} \beta_k f_{k,p}(t) : \beta_0, ..., \beta_{m+p} \in \mathbb{R} \right\}$$

is the set of piecewise p-polynomials on the interval $[t_0, t_{m+1}]$ with continuous p-1 derivatives at the knot points. Note that the same set of functions is obtained if one uses the basis functions

$g_{k,p} = t^k$, $k = 0,..., p$, and $g_{p+k,p}(t) = (t - t_k)_-^p$, k=1, ..., m.

If now

$$\mathcal{G}_p = \left\{ \sum_{k=0}^{m+p} \beta_k g_{k,p}(t) : \beta_0, ..., \beta_{m+p} \in \mathbb{R} \right\}$$

then $\mathcal{G}_p = \mathcal{F}_p$, and the sum of the first p+1 terms in $\sum_{k=0}^{m+p} \beta_k f_{k,p}(t)$ and $\sum_{k=0}^{m+p} \beta_k g_{k,p}(t)$ give the corresponding p-polynomials on the first and last intervals, correspondingly. Then, for example, the null hypothesis

H_0: β_1=...=β_p=0

for the set of functions $\mathcal{G}_p$ says that the mean curve is constant after the last knot point t_k, and this hypothesis can be tested using $(\hat{\beta}_1, ..., \hat{\beta}_p)$. More generally, any linear hypothesis

H_0: Aβ=b

for a chosen r × p matrix A having rank(A)=p and for a chosen r-vector b can be tested using $A\hat{\beta} - b$ that is, under the null hypothesis, approximate r-variate normal with zero mean vector and covariance matrix

$$A\left\{ \sum_{i=1}^{n} X_i' V_i^{-1} X_i \right\}^{-1} A' \text{ with } V_i = F_i \Sigma F_i' + \tau^2 I_{p_i}.$$

Under the null hypothesis, the limiting distribution of the squared form test statistic

$$(A\hat{\beta} - b)' \left[A\left\{ \sum_{i=1}^{n} X_i' \hat{V}_i^{-1} X_i \right\}^{-1} A' \right]^{-1} (A\hat{\beta} - b)$$

is then a chi squared distribution with r degrees of freedom.

B-spline functions

Yet another alternative is to use the basis of B-spline functions [13,14]. B-spline functions are constructed recursively using the original and additional knot points

$$... < t_{-1} < t_0 < ... < t_{m+1} < t_{m+2} < ...$$

as follows. First, basis functions of degree p=0 are given by piecewise constant

$$B_{k,0}(t) = I(t_k \le t < t_{k+1}),$$

k=0, ±1, ±2, ... For p=1, 2, ..., then

$$B_{k,p}(t) = \frac{t - t_k}{t_{k+p} - t_k} B_{k,p-1}(t) + \frac{t_{k+p+1} - t}{t_{k+p+1} - t_{k+1}} B_{k+1,p-1}(t),$$

k=0, ±1, ±2, ... The choice of the additional knot points outside the interval $[t_0, t_{m+1}]$ has naturally an effect on some of the functions $B_{k,p}(t)$ on the interval $[t_0 , t_{m+1}]$ but the function space

$$\mathcal{B}_p = \left\{ \sum_{k=-p}^{m} \beta_k B_{k,p}(t) : \beta_{-p}, \beta_{-p+1}, ..., \beta_m \in \mathbb{R} \right\}$$

spanned by $B_{-p,p}(t),...,B_{m,p}(t)$ does not depend on the choice of the outside knots. Note also that, at interval $[t_k,t_{k+1}]$, $\sum_{k=-p}^{m} \beta_k B_{k,p}(t)$ is a linear combination of p +1 functions (polynomials) $B_{k-p,p}(t),...,B_{k,p}(t)$ only. It is then remarkable that $\mathcal{B}_p = \mathcal{F}_p = \mathcal{G}_p$.

Remark 1: Note that, naturally, all three parametric function sets $\mathcal{B}_p$, $\mathcal{F}_p$, and $\mathcal{G}_p$ provide the same fit (see for example Section 3.7.1 of [6]). Estimation of parameters of $\mathcal{B}_p$ has best numerical properties (see for example Chapter 5 of [11]) but the interpretation of the regression parameters in $\mathcal{F}_p$, and $\mathcal{G}_p$ is often easier. The first p+1 parameters in $\mathcal{G}_p$, for example, give the p-polynomial on the last time interval $[t_m, t_{m+1}]$. The null hypothesis H_0: β_1=...=β_p=0 then says that the mean value is constant (does not depend on time) on the last interval.

A Real Data Example

The data set

We illustrate the theory with a data set from a Finnish longitudinal myopia study. 240 children in central Finland were recruited for this study in 1983-1984. For details about the design and inclusion criteria, see [15]. In this analysis, the measurements are refraction values on the right eyes of 118 girls and 118 boys; 4 children were excluded for

various reasons. In the beginning of the study the ages of the children were between 8.8 and 12.8 years old (mean age 10.9 years). In the first three years the children were measured yearly (they actually were also randomized into three different treatment arms which is ignored here; there were no statistical differences between the arms [16]. As a part of the follow-up, the subjects were supposed to see the same ophthalmologist for measurements 15 and 25 years after the start of the study. Available measurements based on glass prescriptions and files from different ophthalmologists and opticians were used to obtain additional observations between these time points. During the follow-up, 2 subjects died and 18 subjects dropped out because of refractive surgery. The measurements were not equally spaced and often sparse with the mean number of measurements per subject 8.1 (range 2-15). The total number of recordings is 1908 with the oldest age 39.0 years. Table 1 lists the numbers of measurements for different time intervals.

In the original study, measurements of several covariates were collected but, for our demonstration purpose, we consider only sex. The observed individual curves are then shown separately for boys and girls in Figure 1. These are the raw data for our modeling, and the aim is to build prediction curves (with confidence and tolerance bands) for a new subject based on these data. A minor question was to consider the age when the progression of myopia levels off. The statistical tools described in the previous sections are used for the analysis. The analysis was done using R 2.15.0 [17], especially the packages splines and lme4 [18].

Estimates of the parameters for B-spline basis functions

B-spline quadratic basis functions were used for the estimation with the knots at the ages 12, 16, 22. These choices were based on the consultations with specialists. If no prior knowledge were available, the number and locations of the knots could be determined using cross-validation, model se- lection criteria (AIC or BIC), likelihood ratio tests or by using a roughness penalization approach, etc (see for example [7]). In our model the number of basis functions is then k=6. With one dichotomous explaining variable (sex), we have 12 parameters for θ (mean curves) and 21 parameters for Σ, and the residual variance τ^2.

	Age Interval (Years)						
	[8.8, 10)	[10, 15)	[15, 20)	[20, 25)	[25, 30)	[30, 35)	[35, 40)
N	78	890	247	289	122	202	80

Table 1: Number of measurements for different age intervals.

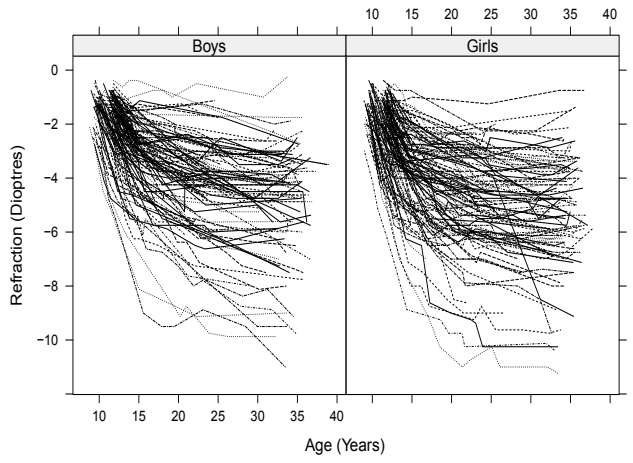

Figure 1: Observed progression curves for boys and girls.

Function	Estimate	Std-Error	Function	Estimate	Std-Error
f_1 (age)	-0.7643	0.1275	f_1 (age)*girl	0.5111	0.1756
f_2 (age)	-1.2067	0.1093	f_2 (age)*girl girl	-0.1676	0.1544
f_3 (age)	-3.1533	0.1417	f_3 (age)*girl	-0.4388	0.2000
f_4 (age)	-4.1287	0.1796	f_4 (age)*girl	-0.3573	0.2503
f_5 (age)	-4.9552	0.2077	f_5 (age)*girl	-0.6107	0.2878
f_6 (age)	-4.8680	0.2558	f_6 (age)*girl	-0.1322	0.3534

Table 2: Estimates of fixed effects and their standard errors.

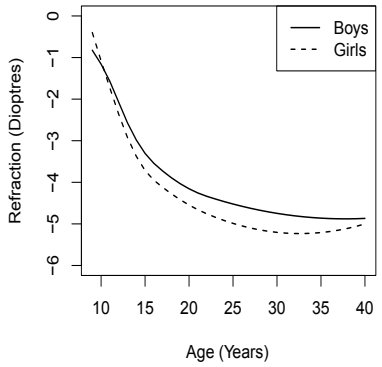

Figure 2: Estimated mean curves for boys and girls.

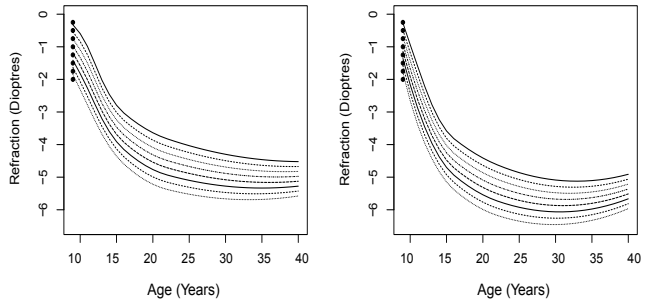

Figure 3: Charts of predicted progression of myopia at age 9. The left panel shows the charts for boys and the right for girls. Each curve describes then the predicted progression curve of myopia based on a single measurement at age 9.

Using the R-function lmer [18], we obtain the fixed effects estimates of θ that are presented in Table 2. The covariance matrix of the random effects, the estimate of Σ, is

$$\hat{\Sigma} = \begin{pmatrix} 0.4058 & 0.4402 & 0.3369 & 0.3760 & 0.2798 & 0.1873 \\ 0.4402 & 1.2050 & 1.0489 & 0.9415 & 1.0892 & 1.0110 \\ 0.3369 & 1.0489 & 2.2398 & 2.3378 & 2.2350 & 2.5036 \\ 0.3760 & 0.9415 & 2.3378 & 3.2986 & 2.8482 & 3.6086 \\ 0.2798 & 1.0892 & 2.2350 & 2.8482 & 3.7131 & 3.6068 \\ 0.1873 & 1.0110 & 2.5036 & 3.6086 & 3.6068 & 5.3909 \end{pmatrix}$$

Finally, the error variance estimate is $\hat{\tau}^2 = 0.0633$

Prediction

The estimated mean curves for boys and girls based on $\hat{\theta}$ are shown in Figure 2. The girls seem to have a faster development of myopia but the differences between boys and girls seem to vanish with the time. After the last knot at age 22, the boys' mean curve seems still to be almost linearly descending while the girls' curve seems to level off.

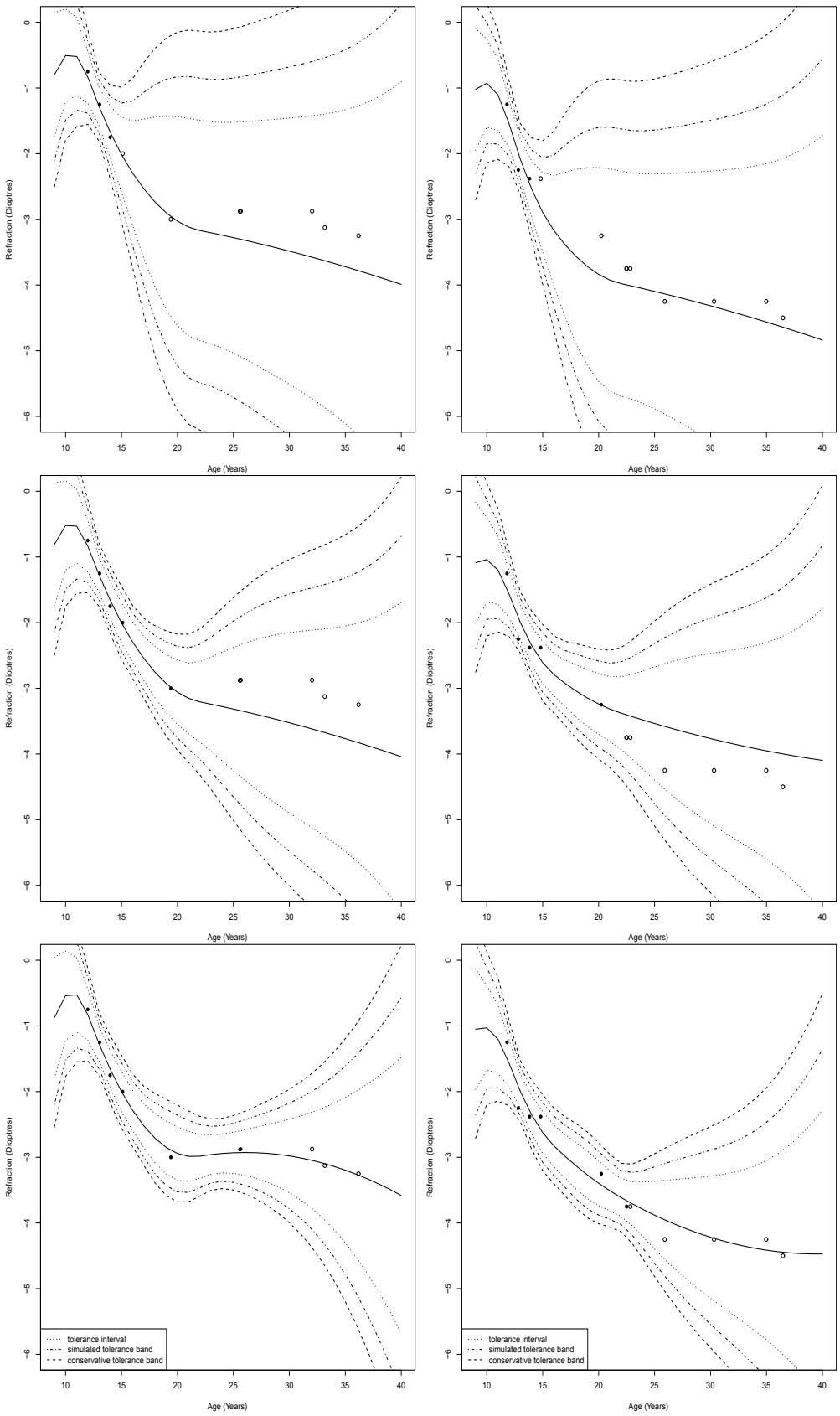

Figure 4: Prediction curves with point wise tolerance intervals and tolerance bands for a randomly selected girl (right column) and boy (left column). In each case, the full points are used for prediction. Intervals and bands are at the 95% level.

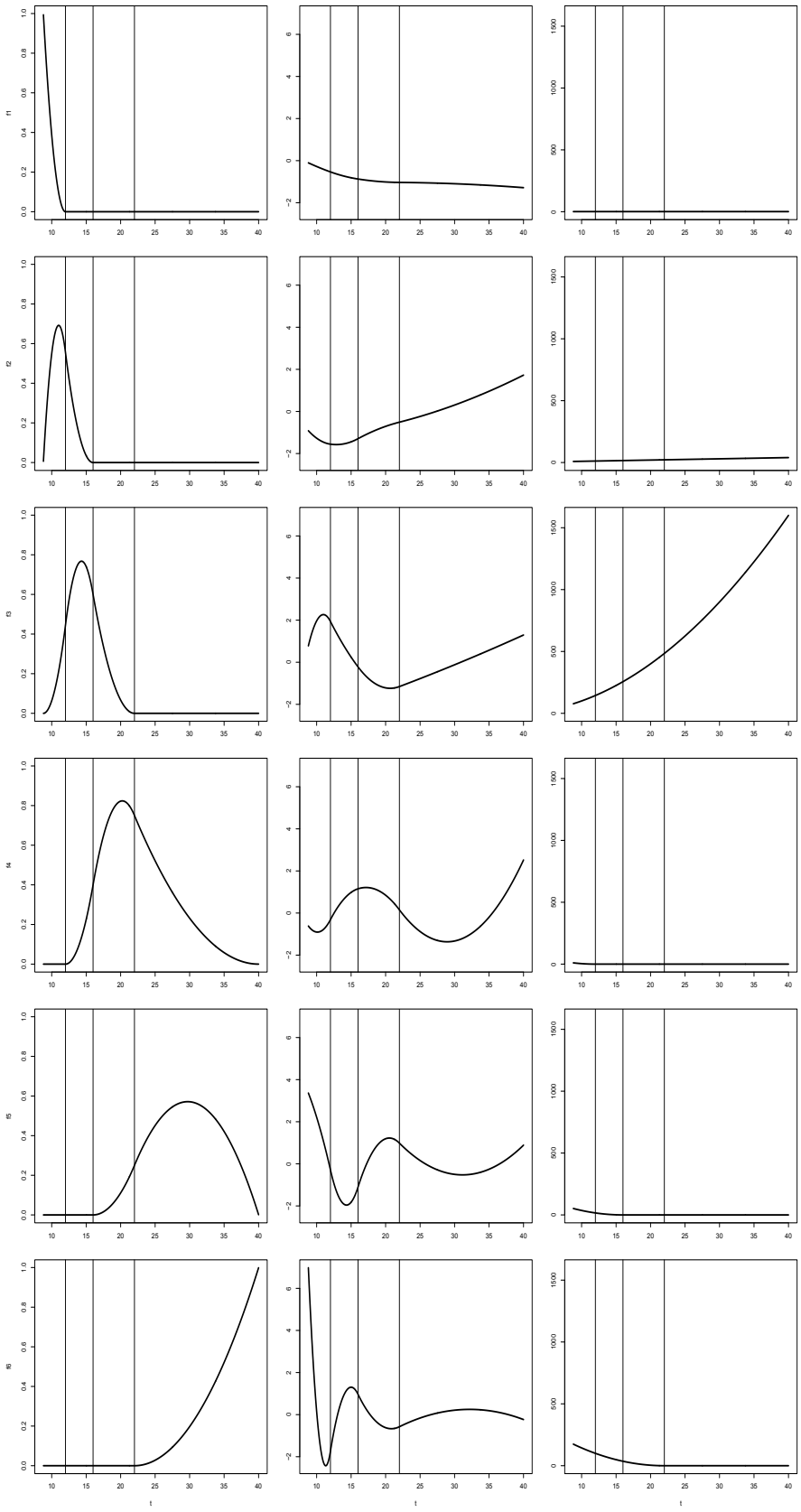

Figure 5: Spline basis functions used in this analysis. The left column gives B-spline basis functions, the middle column principal basis functions, and the right column truncated power spline functions. Vertical lines indicate the knot locations at the different time points.

From a medical point of view, it is interesting to predict the progression of myopia just using the measurement at the first visit. As an example, we provide in Figure 3 the progression chart for the first visit at the age of 9 years.

As explained before, the prediction curves may be based on several observations as well. To illustrate this, we randomly selected one boy and one girl with at least eight measurements. For both subjects, we found the prediction curves with pointwise 95% tolerance intervals and 95% tolerance bands based on the first three, first five, and first seven observations. As explained in Section 2.2, tolerance bands may be here based on $\chi^2_{6,0.95} = 12.5916$ ("conservative") or on the constants $c^2_{0.95}$ ("simulated") that can be estimated through simulations. Our simulated values are based upon 2000 replications. The predictions, tolerance intervals, and tolerance bands are given in Figure 4.

For the girl, the prediction curves based on three or seven observations seem to work well. If five points were used, the (strange) later upward trend at later age is missed. The results for the boy look similar. The conservative tolerance band that is easy to compute is too conservative to be helpful here.

Alternative sets of basis functions

B-spline basis functions are mathematically and numerically attractive but the interpretation of the parameters in Table 2 is difficult. Consider first the presentation of the model using the principal component functions as presented in Section 3.2. First, the matrix Σ_f can be computed using numerical integration, for example. The six eigenvalues of $\Sigma_f^{1/2}\Sigma\Sigma_f^{1/2}$ then are 2.8463, 0.1375, 0.0477, 0.0315, 0.0120, 0.0017, and the cumulative proportions of the variation explained by the eigenvectors are 0.9251, 0.9698, 0.9853, 0.9955, 0.9994, 1.0000, respectively. The B-spline basis functions and the principal component functions are shown in Figure 5 (two first columns). It is remarkable that the first three PCA basis functions explain together over 98% of the variation. The first principal function is simply for the general progression level (with a typical progression). The second component seems to be an indicator for strong early progression as a subject with a large score will exhibit much faster progression than one with a small score. The third function roughly describes the contrast between the first and third interval reflected by the two opposing peaks of the curve.

Although the principal component functions may be used to find the main type of variations between the individual curves, they are not practical if one wishes to have parameters to tell what is happening in the beginning or in the end of the follow-up period. These studies can be performed with the truncated power splines. One may be interested, for example, whether the progression of myopia levels off after the last knot value. Figure 2 shows that on the average the stabilization does happen neither for boys nor for girls. Separate analyses were made for boys and girls, however, to confirm that. For both cases, we fitted the model using B-splines and then changed to the truncated power spline functions shown in Figure 5 (the rightmost column) as explained in Section 3.1. (The matrix A may be simply found by calculating the values of all 12 functions at six suitable time points).

To demonstrate the conversion Tables 3 and 4 gives the estimates for the fixed effects using B-spline functions and truncated power spline functions for boys and girls, respectively. The null hypothesis that the myopia progression levels off at 22, then says that the second and third coefficients for truncated power spline functions are zero. Both p-values are less than 0.0001 and the null hypotheses should therefore be rejected, see Section 4.1.

Function	Estimate	Std-Error	Function	Estimate	Std-Error
f_1(age)	-0.8406	0.1067	g_1(age)	-1.7838	0.8234
f_2(age)	-1.2020	0.0912	g_2(age)	-0.1625	0.0580
f_3(age)	-3.1655	0.1333	g_3(age)	0.0021	0.0010
f_4(age)	-4.1235	0.1751	g_4(age)	-0.0939	0.0164
f_5(age)	-4.9521	0.1928	g_5(age)	0.0339	0.0065
f_6(age)	-4.8842	0.2479	g_6(age)	0.0081	0.0035

Table 3: Estimates of fixed effects and their standard errors for different splines for the boys only data.

Function	Estimate	Std-Error	Function	Estimate	Std-Error
f_1(age)	-0.2390	0.1379	g_1 (age)	-0.7419	0.8342
f_2(age)	-1.3712	0.1258	g_2 (age)	-0.2761	0.0592
f_3(age)	-3.5872	0.1492	g_3 (age)	0.0042	0.0010
f_4(age)	-4.4930	0.1782	g_4 (age)	-0.0400	0.0825
f_5(age)	-5.5646	0.2091	g_5 (age)	0.0465	0.0239
f_6(age)	-4.9913	0.2489	g_6 (age)	0.0034	0.0036

Table 4: Estimates of fixed effects and their standard errors for different splines for the girls only data.

Discussion

In this paper we showed that the random regression analysis using polynomial spline functions is a flexible way to model myopia progression. The estimated model then also provides easy prediction charts for myopia progression depending on previous measurements. In the estimation of the parameters in the model, the B-spline functions are preferable but the results can be easily transformed to any set of basis functions spanning the same function space. The model also easily allows the use of covariates.

In our example, the theory was illustrated by testing the hypothesis that on the average the myopia progression levels off at the age of 22. The hypothesis was rejected but we believe that this may have happened partly due to the bias coming from the data collection procedure. The numbers of measurements p_i and data points $t_{i1}, \ldots, t_{ip_i}$ are informative in the sense that fast myopia progression is naturally connected to a high number of measurements. The subjects with early stabilized level have sparse late measurements, and the estimates of the parameters describing the progression in the end of the follow-up are then mainly based on the measurements from subjects with late progression. This causes bias in the estimation of the model. This is similar to the analysis of missing data problems (with informative missingness) or to the analysis of clustered data with informative cluster size [19]. To correct the bias is an interesting topic for a future work. Furthermore, robust estimation procedures should be developed here as the data contain clear outliers, that is, the individual curves with atypical or even impossible behaviour. The outliers naturally have a strong effect on the fit of the data.

Acknowledgment

We thank the referee for careful reading of the paper and helpful comments. The work of Klaus Nordhausen and Hannu Oja was supported by the Academy of Finland (grants 131929, 218327 and 268703). The work of Olavi Pärssinen was supported by grants from the Finnish Eye Foundation and The Society of Finnish Ophthalmologists.

References

1. Pärssinen O (1987) Relation between refraction, education, occupation, and age among 26- and 46-year-old finns. American Journal of Optometry and Physiological Optics 64: 136-143.

2. Mäntyjärvi MI (1985) Predicting of myopia progression in school children. Journal Pediatric Ophthalmology and Strabismus 22: 71-76.

3. Thorn F, Gwiazda J, Held R (2005) Myopia progression is specified by a double exponential growth function. Optometry and Visual Science 82: 286-297.
4. Curtin BJ (1985) The myopias. Basic science and clinical management. Harper & Row, Philadelphia.
5. Möttönen J, Oja H, Krause U, Rantakallio P (1995) Application of random coefficient regression model to myopia data. Biometrical Journal 37: 657-672.
6. Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, Cambridge.
7. Wu H, Zhang JT (2006) Nonparametric regression methods for longitudinal data analysis. Wiley, Hobocken.
8. Shi M, Weiss RE, Taylor JMG (1996) An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. Journal of the Royal Statistical Society Series C 45: 151-163.
9. Mayer K (2005) Random regression analyses using B-splines to model growth of Australian Angus cattle. Genet Sel Evol 37: 473-500.
10. Rice JA, Wu CO (2001) Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 57: 253-259.
11. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference and prediction. (2nd ed), Springer, New York.
12. Ramsay JO, and Silverman BW (2005) Functional Data Analysis (2nd ed) Springer, New York.
13. Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalities. Statistical Science 11: 89-102.
14. De Boer C (2001) A practical guide to splines. Springer, New York.
15. Hemminki E, Pärssinen O (1987) Prevention of myopic progress by glasses. study design and the first-year results of a randomized trial among school children. Am J Optom Physiol Opt 64: 611-616.
16. Pärssinen O, Lyyra AL (1993) Myopia and myopic progression among school children: a three-year follow-up study. Invest Ophthalmol Vis Sci 34: 2794-2802.
17. R Development Core Team (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
18. Bates D, Maechler M, Bolker B (2011) lme4: linear mixed-effects models using S4 classes. R package version 0.999375-42.
19. Nevalainen J, Datta S, Oja H (2014) Inference on the marginal distribution with clustered data and informative cluster size. Statistical Papers 55: 71-92.