Semiparametric regression

Homework Assignment #4

Makowski Michał

10 may 2017

Contents

Introduction

This couple pages cover fourth homework for Semiparametric Regression, a course conducted by proffesor Jarosław Harężlak at University of Wrocław. On the following pages three exercises will be presented. They are focused on mixed models on grouped data. Examples of such data are obtain during medical studies in which patients are followed over time and measurements on them recorded repeatedly, educational studies in which students grouped into classrooms and schools are scored on examinations and sample surveys in which the respondents to questionnaires are grouped within geographical districts.

Mixed models are a good choice for the analysis of grouped data, with random effects used to account for within-group dependence.

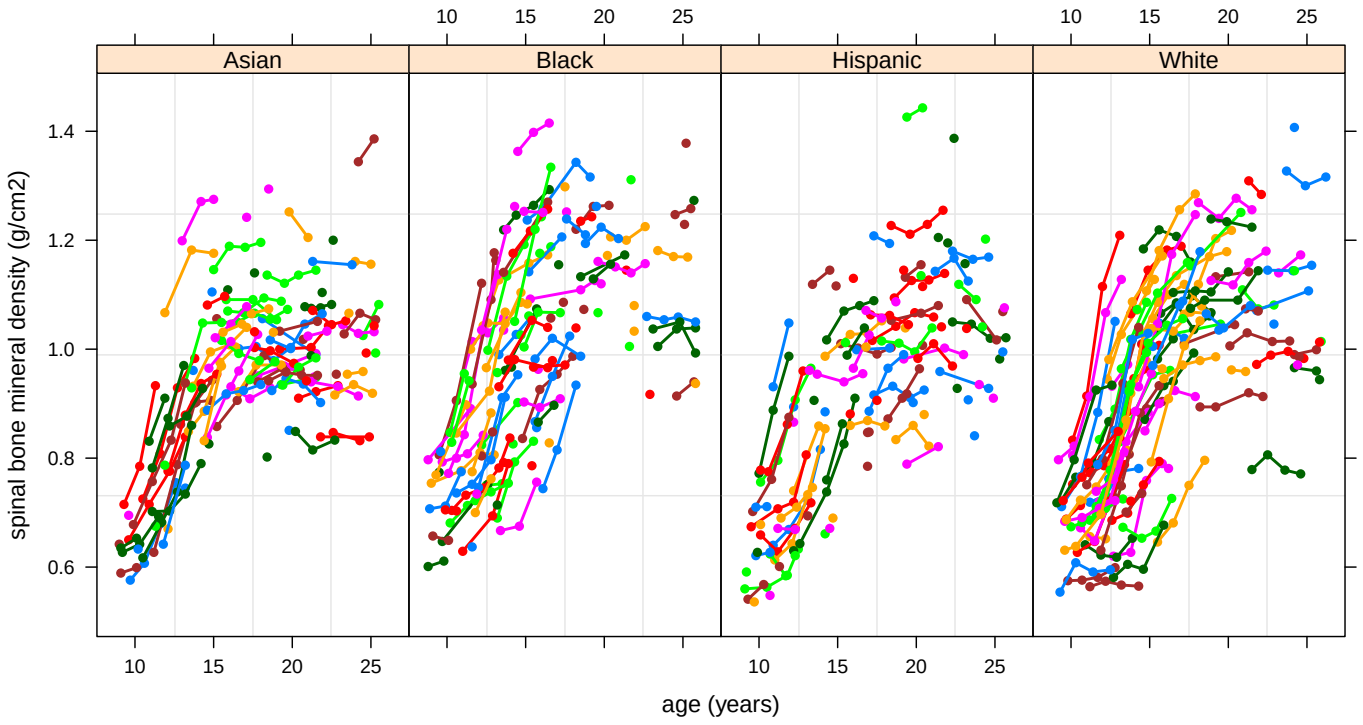
Packages **HRW** (data), **mgcv** (model fitting) and **lattice** (visualization) will be used.

Exercise II (Additive Mixed Models)

In exercise II we want to fit smooth line to the data about spinal bone mineral density.

At first, we plot the data divided by ethnic group.

```
xyplot(spnbmd~age|factor(ethnicity),
       group = idnum,
       data = femSBMD,
       xlab = "age (years)",
       ylab = "spinal bone mineral density (g/cm2)",
       panel = function(x,y,subscripts,groups)
       {
           panel.grid()
           panel.superpose(x,y,subscripts,groups,type = "b",pch = 16,lwd = 2)
       })
```



Above graph shows how the spinal bone mineral density chages during the life of a single represantative in each cohort. We can see some dependecies, which are different in each group and we would like to focus on them a little more.

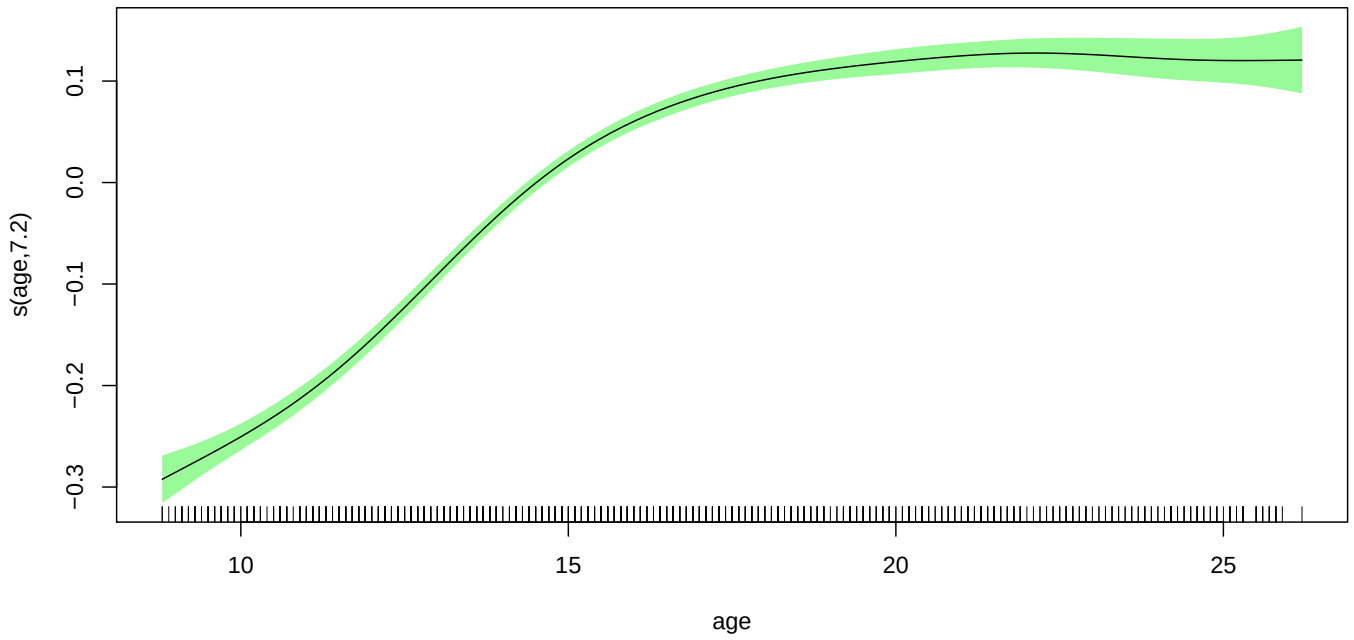
For such data we will use following model:

$$\text{spnbmd}_{i,j} = U_i + f(\text{age}_{i,j}) + \beta_1 \text{black}_i + \beta_2 \text{hispanic}_i + \beta_3 \text{white}_i + \epsilon_{i,j} 1 \leq j \leq n_i, \quad 1 \leq i \leq 230 U \sim N(0, \sigma_U^2), \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

where black_i, hispanic_i, white_i are of course indicators and $\text{spnbmd}_{i,j}$ are the mesurments of spinal bone density of ith subject at jth time.

We could interprate this model as following: Asian group is the "basis", the reference, the coefficients β_1, β_2 and β_3 are the modifications of the mean in other ethics groups.

```
fit = gamm(spnbmd~s(age) + black + hispanic + white, random = list(idnum = ~1), data = femSBMD)
plot(fit$gam, shade = TRUE, shade.col = "palegreen")
```



The shaded region corresponds to pointwise approximate 95% confidence in- tervals. Note that default plotting of the estimate of $f(\text{age})$ involves vertical centering about zero.

Let's have a look at the summary

```
summary(fit$gam)
```

```
Family: gaussian
Link function: identity

Formula:
spnbmd ~ s(age) + black + hispanic + white

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.92538    0.01243  74.444  < 2e-16 ***
black        0.08191    0.01718   4.769 2.13e-06 ***
hispanic    -0.01516    0.01754  -0.864    0.388
white        0.01503    0.01748   0.860    0.390
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
         edf Ref.df     F p-value
s(age) 7.201  7.201 225.6  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.519
  Scale est. = 0.0013551  n = 1003
```

Note that the fitted age effect involves 7.201 effective degrees of freedom. Let's look at the confidence intervals

```
intervals(fit$lme)
```

```
Approximate 95% confidence intervals

 Fixed effects:
                   lower        est.      upper
X(Intercept)  0.90101757  0.92538331 0.94974906
Xblack        0.04821097  0.08190524 0.11559951
Xhispanic    -0.04956727 -0.01515680 0.01925366
Xwhite       -0.01925714  0.01503146 0.04932006
Xs(age)Fx1    0.02267288  0.07541370 0.12815452
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: g
                lower       est.     upper
sd(Xr - 1) 0.01371071 0.04113408 0.1241885
  Level: idnum
                     lower      est.     upper
sd((Intercept)) 0.1138351 0.1222052 0.1311908

 Within-group standard error:
     lower       est.      upper
0.03474367 0.03681208 0.03900364
```

This output shows that an approximate 95% confidence interval for β_1 is $(0.0482, 0.116)$, which indicates a statistically significant difference between the Asian and Black females in terms of mean spinal bone mineral density. However, there is no significant difference between Hispanic or White females and Asian females. An approximate 95% confidence interval for σ_U is $(0.114, 0.131)$, which implies significant within-subject correlation. The 95% confidence interval for σ_ϵ is

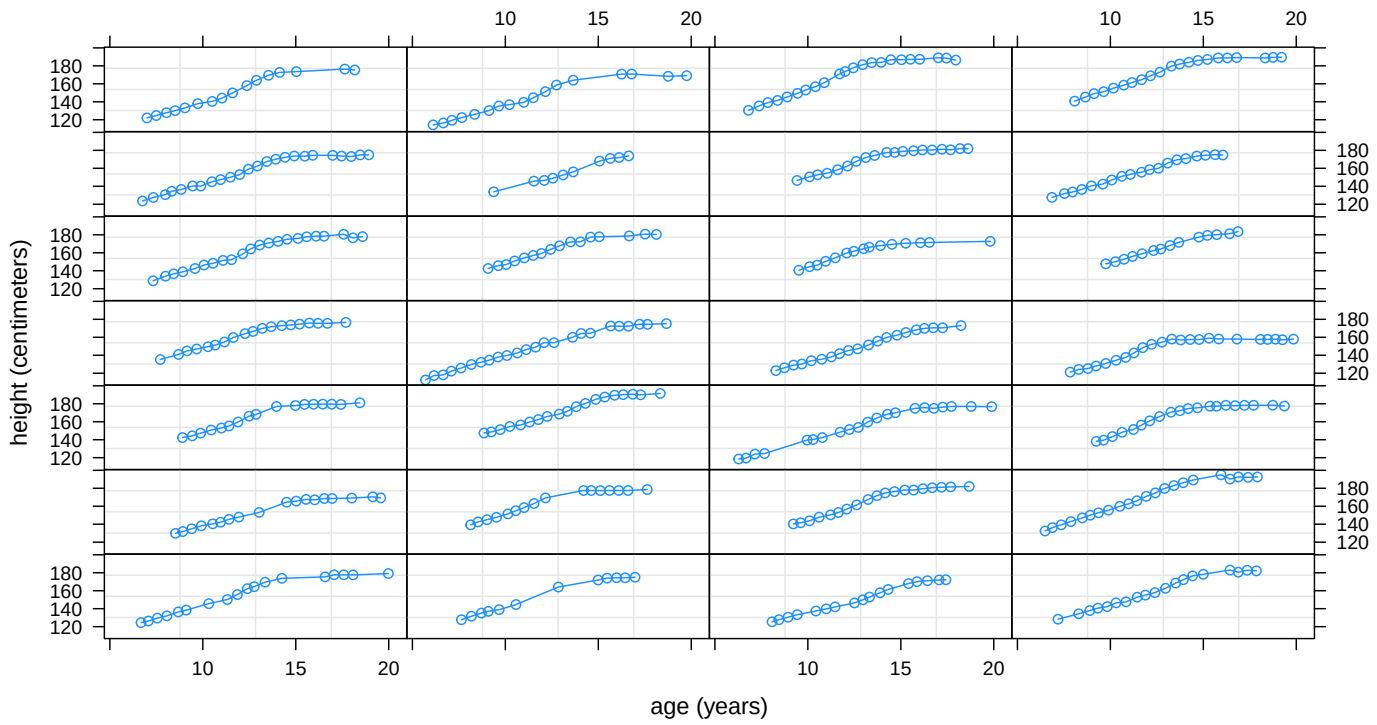
$(0.0347, 0.0390)$.

Exercise III (Models with Group-Specific Curves)

In exercise III we want to fit group-specific model to the data of adolescent somatic growth obtained from a study of the mechanisms of human hypertension development conducted at the Indiana University School of Medicine, Indianapolis, Indiana, USA. Pratt et al.(1989).

We restrict attention to the black males in the study. The following plot show the trend of each subject, there are 28 of them.

```
growthINblackMales = growthIndiana[(growthIndiana$male == 1) & (growthIndiana$black == 1),]

xyplot(height~age|idnum,
       group = idnum,
       data = growthINblackMales,
       layout = c(4,7),
       strip = F,
       xlab = "age (years)",
       ylab = "height (centimeters)",
       as.table = TRUE,
       panel = function(x,y,subscripts,groups)
       {
           panel.grid()
           panel.superpose(x,y,subscripts,groups,col = "dodgerblue",type = "b")
       })
```



Above graph shows the history for each, individual representative. We notice that he shapes of the curves for each adolescent differ quite markedly and the simple additive mixed models would not capture such behavior very well.

Instead, such data we will use following model:

$$\text{height}_{i,j} = f(\text{age}_{i,j}) + g_i(\text{age}_{i,j}) + \epsilon_{i,j} 1 \leq j \leq n_i, \quad 1 \leq i \leq 28 \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

where n_i is the number of measurments for ith subject, g_i is a function that represents that adolescent's departure from the overall mean function f. This is **semiparametric mixed model**.

Modeling of f and the g_i can proceed according to

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K_{gbl}} u_{gbl,k} z_{gbl,k}(x), \qquad u_{gbl,k} | \sigma_{gbl} \sim N(0, \sigma^2_{gbl})$$

$$g_i(x) = U_{0i} + U_{1i} x + \sum_{k=1}^{K_{grp}} u_{grp,ik} z_{grp,k}(x) \qquad \begin{bmatrix} U_{0i} \\ U_{1i} \end{bmatrix} \sim N(0, \Sigma), \qquad u_{grp,ik}|$$

where $z_{gbl,k}$ and $z_{grp,k}$ are suitable spline bases of sizes K_{gbl} and K_{grp} respectively. We will use canonical O'Sullivan splines. Typically, K_{grp} is smaller than K_{gbl} since fewer basis functions are needed to handle group-specific deviations.

Exercise IV (Fitting Group-Specific Curves Models)

To fit model introduced in exercise III we will use package **nlme**.

At first we wxtract the necessary variables from the **growthINblackMales** dataset

```
age = growthINblackMales$age
height = growthINblackMales$height
idnum = growthINblackMales$idnum
```

Then we create an array with identification numbers

```
uqID = unique(idnum)
uqID.tab = table(idnum)
uqID.len = length(uqID)
growthINblackMales$idnumBM = as.numeric(factor(rep(uqID, uqID.tab), labels= 1:uqID.len))
idnumBM = growthINblackMales$idnumBM
```

Next, we will set up the design matrices Z_{gbl}, containing the $Z_{gbl,k}$, and Z_{grp}, containing the $Z_{grp,k}$

```
numObs = length(height)
numGrp = uqID.len
numIntKnotsGbl = 20
intKnotsGbl = quantile(unique(age),
seq(0,1,length=numIntKnotsGbl+2))[-c(1,numIntKnotsGbl+2)]
range.age = c(5.5,20)
Zgbl = ZOSull(age,range.x=range.age,intKnots=intKnotsGbl)

numIntKnotsGrp = 10
intKnotsGrp = quantile(unique(age),
seq(0,1,length=numIntKnotsGrp+2))[-c(1,numIntKnotsGrp+2)]
Zgrp = ZOSull(age,range.x=range.age,intKnots=intKnotsGrp)
```

Then we set up the random effect structure for the call to lme()

```
dummyId = factor(rep(1,numObs))
Zblock = list(dummyId=pdIdent(~-1+Zgbl),idnumBM=pdSymm(~age),idnumBM=pdIdent(~-1+Zgrp))
```

Let have a brief look at operation done above:

- The dummy identification variable dummyID, an array of length numObs, the total number of observations, with all entries equal to one tricks lme() into accommodating the global penalized spline component.
- The list entry dummyId=pdIdent(-1+Zgbl) invokes the multiple of identity matrix structure $u_{gbl} \sim N(0, \sigma^2_{gbl} I)$ across the entire dataset regardless of within-subject grouping.
- The list item idnumBM=pdSymm(age) invokes the block-diagonal unstructured 2×2 covariance matrix form on the $[U_{0i}, U_{1i}]^T$, $1 \leq i \leq 23$, as required by the model,
- Similarly pdIdent(-1+Zgrp) accommodates $u_{grp,ik} | \sigma_{grp} \sim N(0, \sigma^2_{grp})$

We are now ready to call lme() with the random argument set to Zblock

```
blkMalGD = groupedData(height ~ age|rep(1,length = numObs),
                       data = data.frame(height,age,Zgbl,Zgrp,idnumBM))
fit = lme(height ~ age,data = blkMalGD,random = Zblock)
```

Now let's plot fitted curves, firtsly we wetup the fixed and random design matrices

```
ng = 201
ageg = seq(range.age[1],range.age[2],length = ng)
Xg = cbind(rep(1,ng),ageg)
Zgblg = ZOSull(ageg,range.x = range.age,
intKnots = intKnotsGbl)
Zgrpg = ZOSull(ageg,range.x = range.age,
intKnots = intKnotsGrp)
```

Then we extract the model coefficients

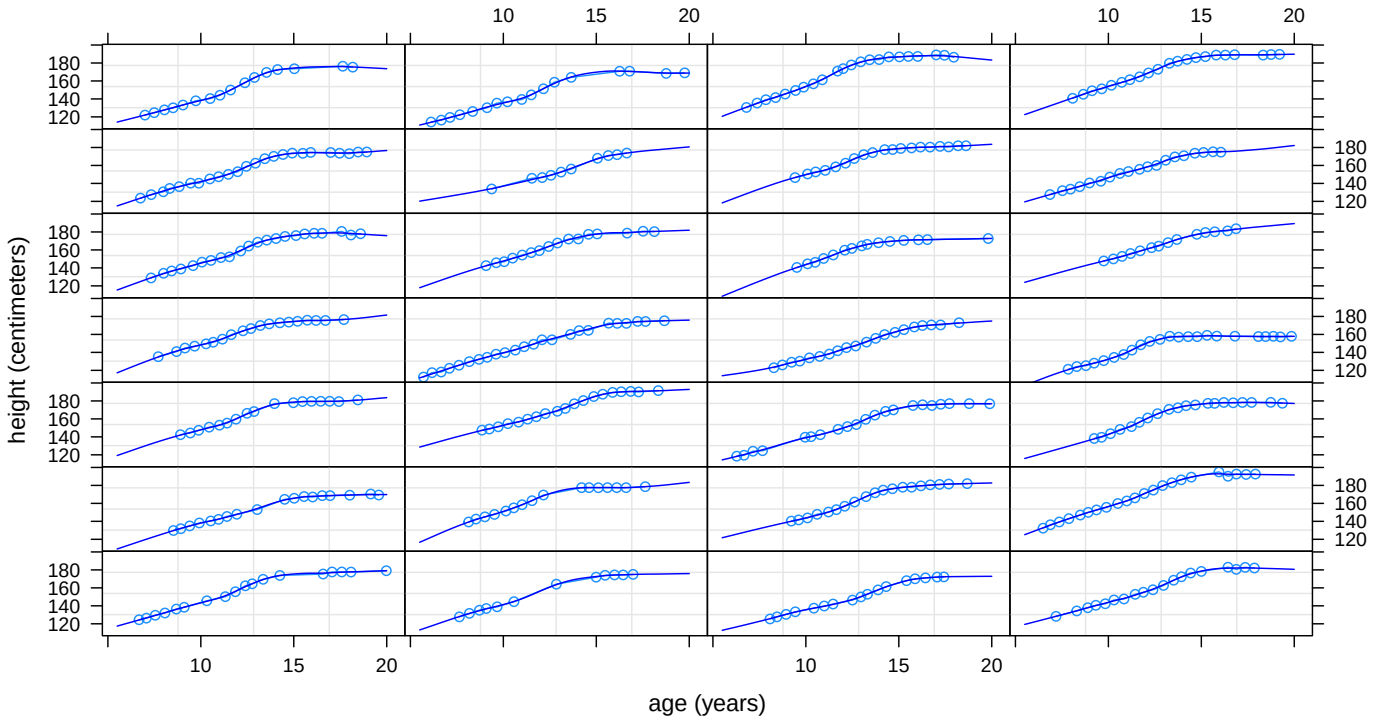
```
betaHat = as.vector(fit$coef$fixed)
uHat = as.vector(fit$coef$random[[1]])
fHatg = as.vector(Xg%*%betaHat + Zgblg%*%uHat)
```

In the end we estimate the subject-specific curves

```
curvEsts = vector("list",numGrp)
for (i in 1:numGrp)
{
    uLinHati = as.vector(fit$coef$random[[2]][i,])
    uSplHati = as.vector(fit$coef$random[[3]][i,])
    ghati = Xg%*%uLinHati + Zgrpg%*%uSplHati
    curvEsts[[i]] = fHatg + ghati
}
```

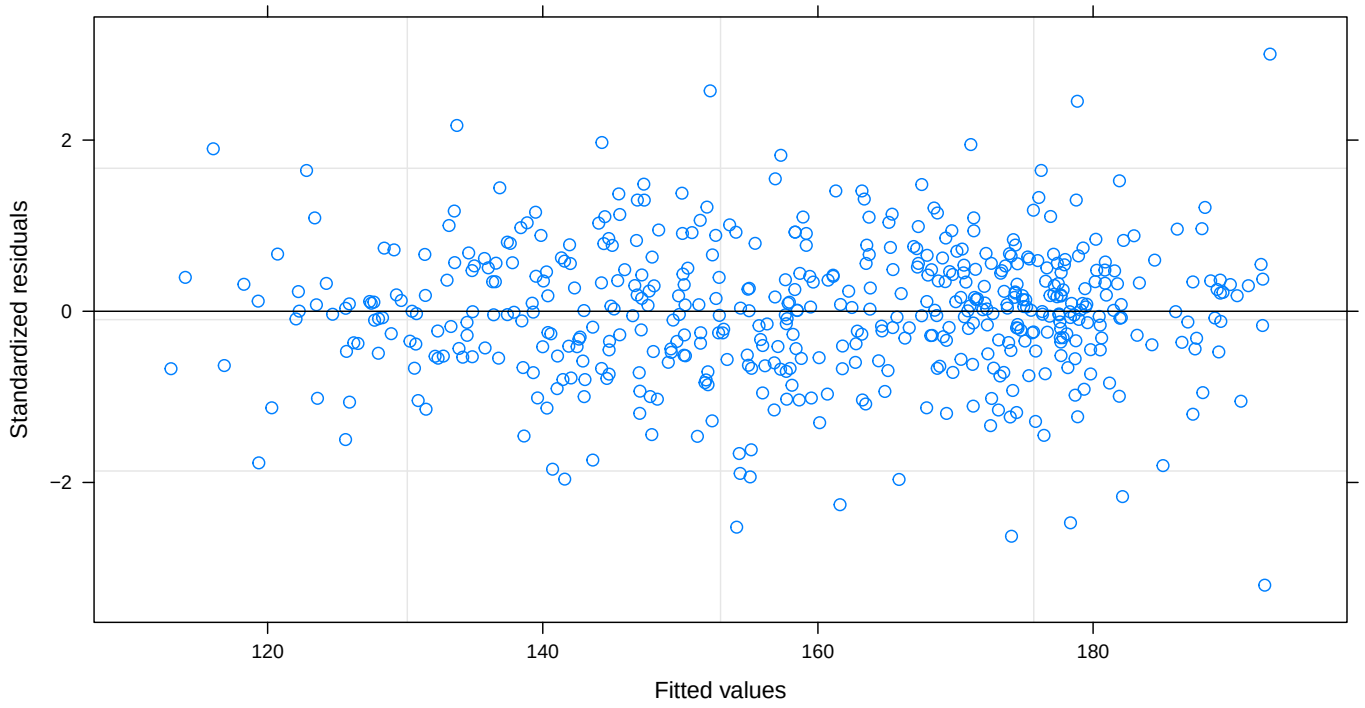
Finally, we are ready to plot the curves

```
xyplot(height ~ age|idnumBM,groups = idnumBM,
       data = growthINblackMales,
       strip = FALSE,
       xlab = "age (years)",
       ylab = "height (centimeters)",
       as.table = TRUE, layout = c(4,7),
       panel = function(x,y,subscripts,groups)
       {
           panel.grid()
           adolNum = idnumBM[subscripts][1]
           panel.superpose(x,y,subscripts,groups,col = "dodgerblue",type = "b")
           panel.xyplot(ageg,curvEsts[[adolNum]],col = "blue",type = "l")
       })
```



We have to remember to check the standardized residuals from the fit shown in the last plot. Let's see

```
plot(fit)
```



We connot see any distinct patterns or outliers.

Exercise V

Model presented in **Exercise II** for the female spinal bone mineral density data assumes that the mean functions for each ethnicity category differ only by vertical shifts. A more flexible model is

$$\text{spnbmd}_{i,j} = U_i + f_{ethnicity}(\text{age}_{i,j}) + \epsilon_{i,j} 1 \leq j \leq n_i, \quad 1 \leq i \leq 230 U \sim N(0, \sigma_U^2), \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

where

$$\text{etnicity} \in \{\text{asian}, \text{black}, \text{hispanic}, \text{white}\}, \quad 1 \leq i \leq 230.$$

Now we will create such model and look at it's summary

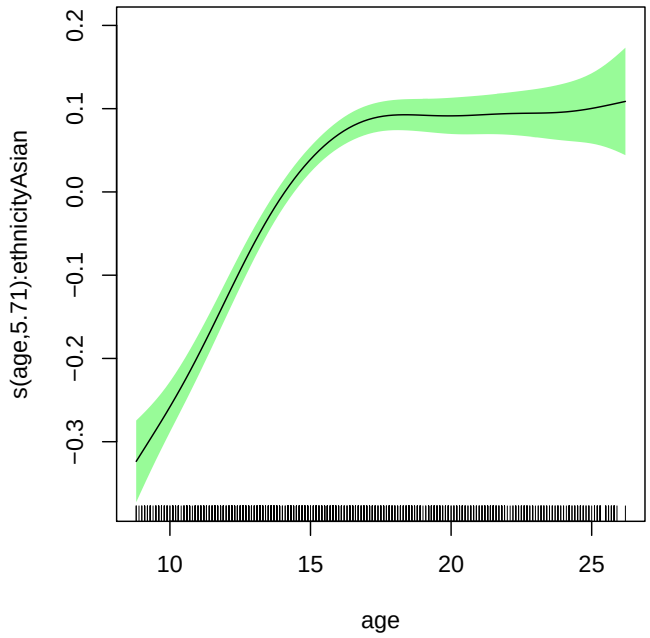
```
fitEthno = gamm(spnbmd~s(age, by=ethnicity), random = list(idnum = ~1), data = femSBMD)
summary(fitEthno$gam)
```

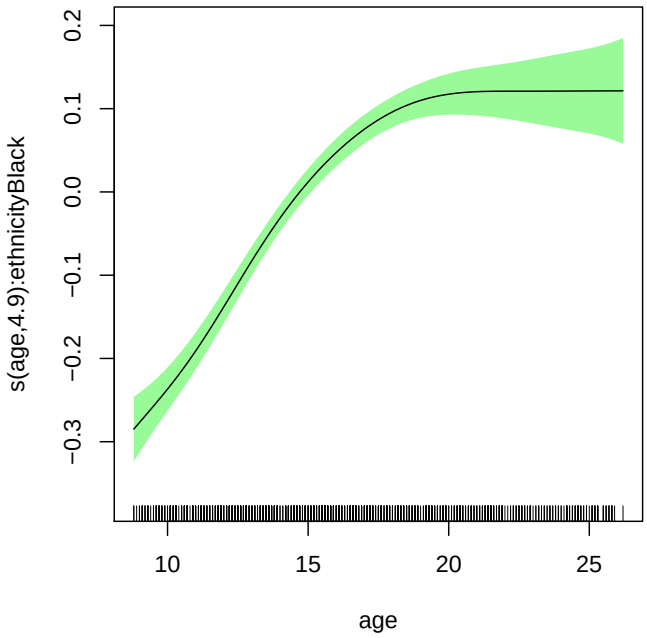
```
Family: gaussian
Link function: identity

Formula:
spnbmd ~ s(age, by = ethnicity)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.949441   0.006393   148.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                           edf Ref.df      F p-value
s(age):ethnicityAsian    5.709  5.709  53.42  <2e-16 ***
s(age):ethnicityBlack    4.899  4.899  72.46  <2e-16 ***
s(age):ethnicityHispanic 5.559  5.559  42.95  <2e-16 ***
s(age):ethnicityWhite    6.065  6.065 119.05  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.484
  Scale est. = 0.0013284  n = 1003
```

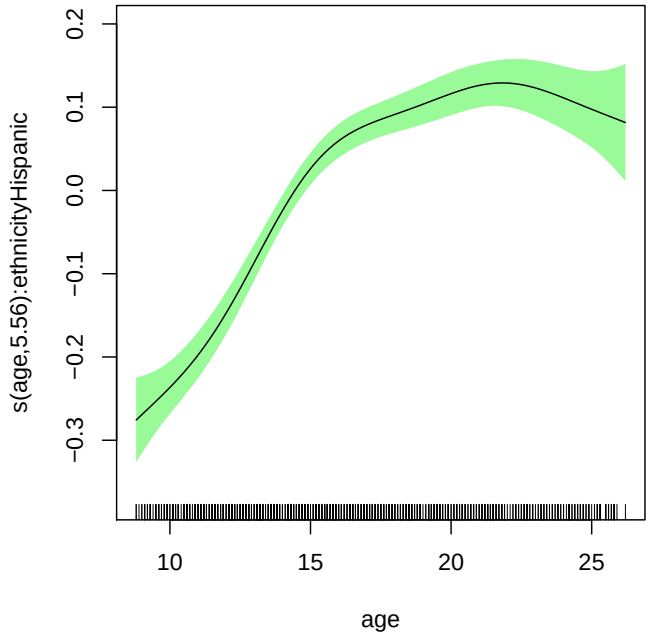
We can see that smoothing splines for Asian and Hispanic are quite similiar, while for other ethnics groups they differ. That correlate with our summory of **Exercise II**. Now we will plot the smooting splines for each group. For more analysis we could call summary(fitEthno$gam), it gives us more information about mixed effects.

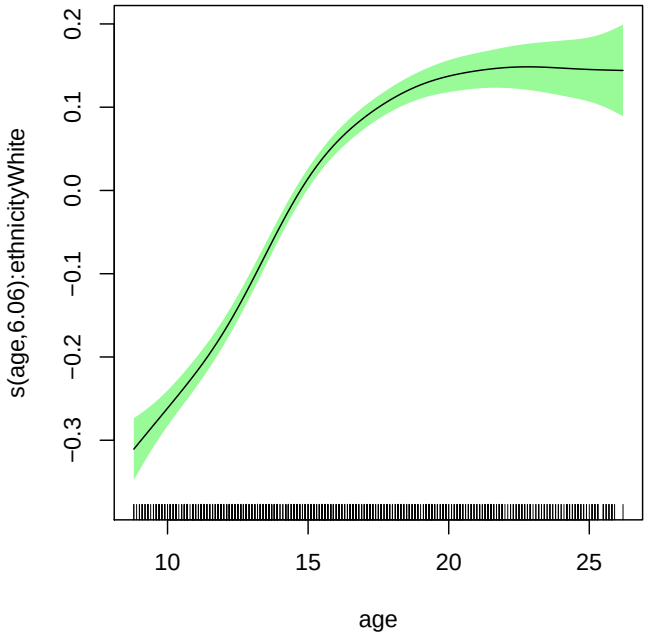
Let's compare fitted splines graphicaly

```
plot(fitEthno$gam, shade = TRUE, shade.col = "palegreen", pages = 2)
```









All plots slightly differ from each other. The axes are locked for every group, that let us to compare each other.