Semiparametric regression

Homework Assignment #3

Makowski Michał

10 may 2017

Contents

Introduction

This couple pages cover third homework for Semiparametric Regression, a course conducted by proffesor Jarosław Harężlak at University of Wrocław. On the following pages two exercises will be presented. First one compares two different methods of fitting splines on data about ozone concentration in Los Angeles. Second is focused on comparing diffrent curve fits on data about milk characteristc among different cows with different diets.

Exercise I

In this exercise we want to construct a model, which says us how ozone concentration depends on others variables. To do it we use Generalized Additive Models.

Description of data

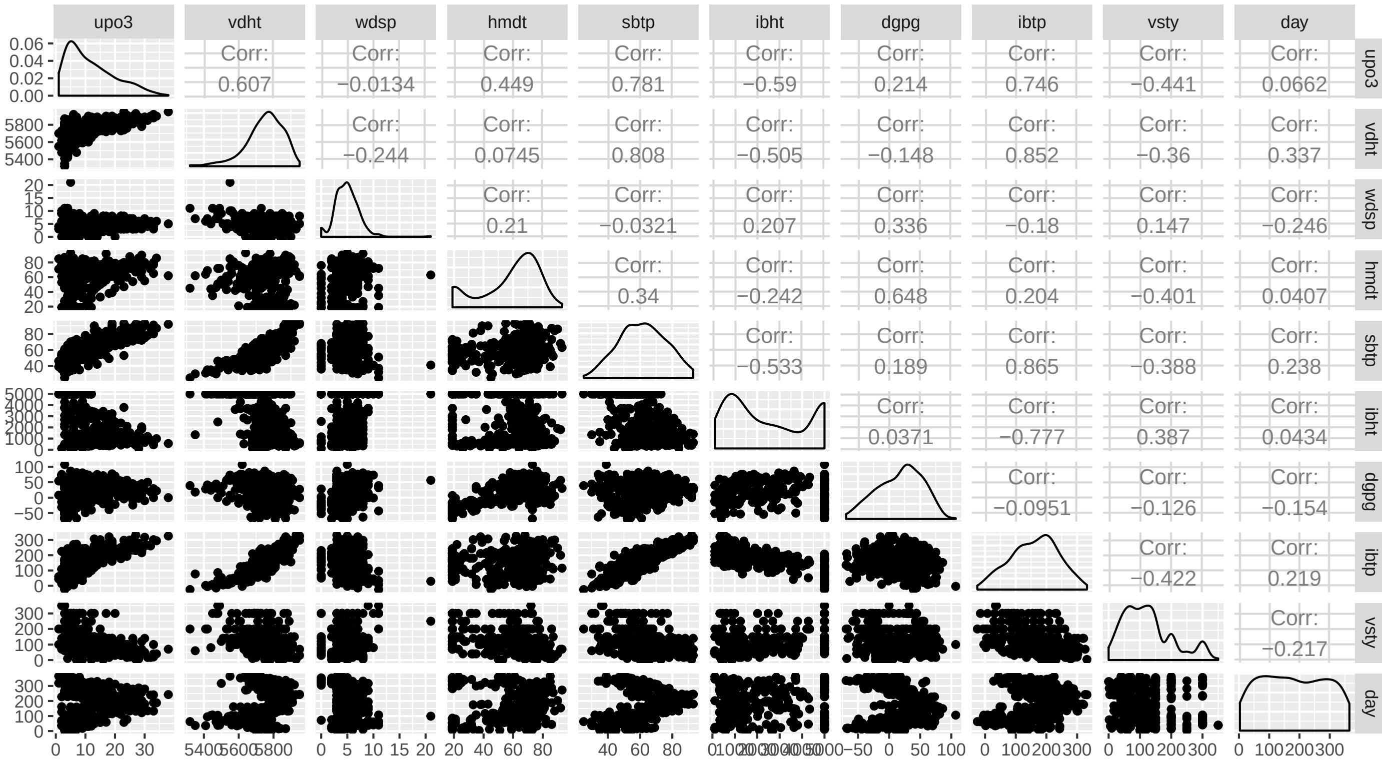
Daily measurements of ozone concentration and eight meteorological quantities in the Los Angeles basin for 330 days of 1976.

A data frame containing 330 observations on the following variables.

- **upo3-** Upland ozone concentration, in ppm.
- **vdht-** Vandenberg 500 millibar height, in meters.
- **wdsp-** Wind speed, in miles per hour.
- **hmdt-** Humidity.
- **sbtp-** Sandburg Air Base temperature, in Celsius.
- **ibht-** Inversion base height, in foot.
- **dgpg-** Dagget pressure gradient, in mmHg.
- **ibtp-** Inversion base temperature, in Fahrenheit.
- **vsty-** Visibility, in miles.
- **day-** Calendar day, between 1 and 366.

Relations beetween data

Let's invistigate data with `ggpairs(ozone)` command.



We could see that relations are mostly non linear.

Model selection

We build full model and then, using *step.Gam()* function, we to select predictors and their diffrent forms among Gaussian GAMs with `upo3` as the response variable. We will choose beetween linear and nonlinear forms.

```
## Start:  upo3 ~ vdht + wdsp + hmdt + sbtp + ibht + dgpg + ibtp + vsty +      day; AIC= 1932.326
## Step:1 upo3 ~ vdht + wdsp + hmdt + s(sbtp, 2) + ibht + dgpg + ibtp +      vsty + day ; AIC= 1903.884
## Step:2 upo3 ~ vdht + wdsp + hmdt + s(sbtp, 2) + ibht + dgpg + ibtp +      vsty + s(day, 2) ; AIC= 1880.303
## Step:3 upo3 ~ vdht + wdsp + hmdt + s(sbtp, 2) + ibht + s(dgpg, 2) +      ibtp + vsty + s(day, 2) ; AIC= 18
## Step:4 upo3 ~ vdht + wdsp + hmdt + s(sbtp, 2) + ibht + s(dgpg, 2) +      ibtp + s(vsty, 2) + s(day, 2) ; A
## Step:5 upo3 ~ vdht + wdsp + hmdt + s(sbtp, 2) + ibht + s(dgpg, 2) +      s(ibtp, 2) + s(vsty, 2) + s(day, 
## Step:6 upo3 ~ vdht + wdsp + hmdt + s(sbtp, 2) + s(ibht, 2) + s(dgpg,      2) + s(ibtp, 2) + s(vsty, 2) + s
```

For stepwise search AIC is the smallest for model with predictors presented below.

```
names(stepFitupo3$"model")[-1]
```

```
## [1] "vdht"       "wdsp"       "hmdt"       "s(sbtp, 2)" "s(ibht, 2)" "s(dgpg, 2)" "s(ibtp, 2)"
## [8] "s(vsty, 2)" "s(day, 2)"
```

Model with GCV-based smoothing parameter selection

Now we fit model with predictors, which was chosen in previous part.

```
## upo3 vdht wdsp hmdt sbtp ibht dgpg ibtp vsty  day
##   35   53   12   65   63  196  128  193   24  330
```

Model summary is presented below.

```
fitStepOzone <- gam(upo3 ~ vdht + wdsp + hmdt +
                        s(sbtp, k=2) + s(ibht, k=2) + s(dgpg, k=2) +
                        s(ibtp, k=2) + s(vsty, k=2) + s(day, k=2), data = ozone)
```

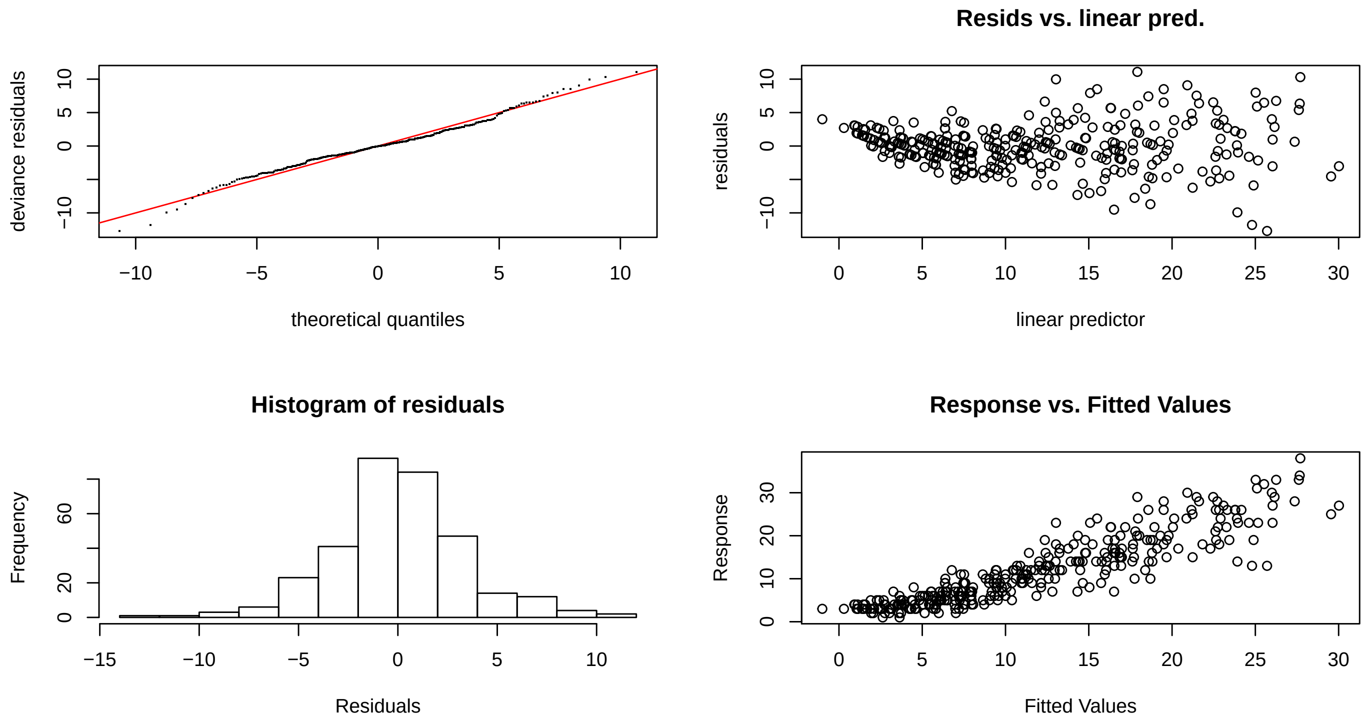
```
fitStepOzone
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## upo3 ~ vdht + wdsp + hmdt + s(sbtp, k = 2) + s(ibht, k = 2) +
##     s(dgpg, k = 2) + s(ibtp, k = 2) + s(vsty, k = 2) + s(day,
##     k = 2)
##
## Estimated degrees of freedom:
## 1.84 1.76 1.96 1.54 1.88 1.99  total = 14.95
##
## GCV score: 15.23404
```

Model modyfication

In this part we want to choose model with sufficient number of basis functions. To do this we loop over a number of basis function level until we receive satisfactory results. We have to modify k's manually - chosen formula is presented below. There should be more diagnostics done in this part to develop best model and fully check it.

```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 8 iterations.
## The RMS GCV score gradient at convergence was 0.00001515444 .
## The Hessian was positive definite.
## Model rank =  70 / 70
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##            k'   edf k-index p-value
## s(sbtp)  4.00  2.51    1.05    0.77
## s(ibht)  6.00  2.71    0.96    0.17
## s(dgpg) 14.00  3.37    1.06    0.86
## s(ibtp) 24.00  9.82    0.96    0.26
## s(vsty)  4.00  2.03    1.05    0.81
## s(day)  14.00  4.34    0.95    0.13
```



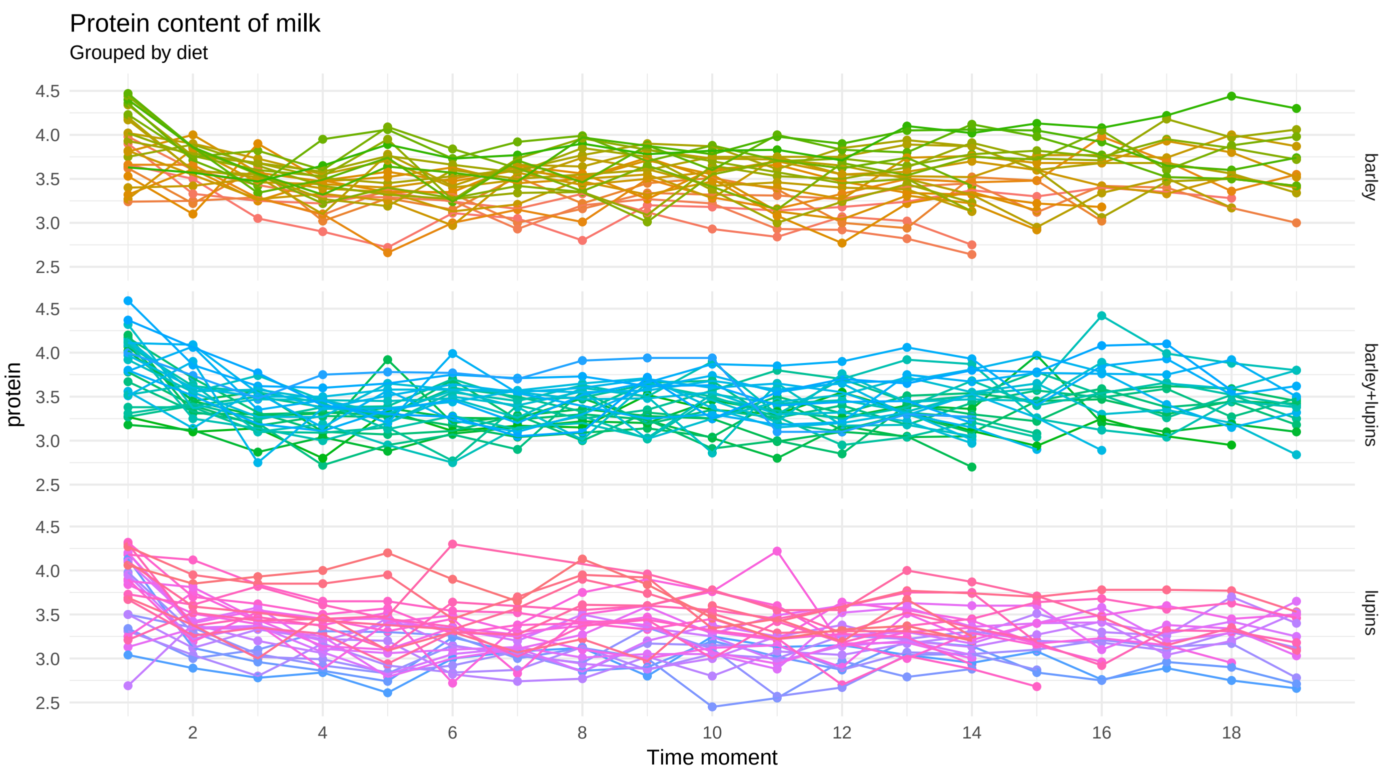
Probability distribution of deviance residuals is almost normal, but the residuals variance is not constant. There is a very quadratic dependence between residuals and linear predictors, Unfortunetelly, we were unable to find model, which looks MUCH better than this presented above.

Exercise II

In this task we have data contaning measurements of the milk protein level over time since calving. Data contains information for of 79 cows. There were three types of diet. In first part we illustrate the fitting of additive mixed models for all cows. Next we show group-specific curves semiparametric mixed model for each diet group.

First we, we will dive into the data more.

```
##     protein           Time              Cow                   Diet          barley      
##  Min.   :2.450   Min.   : 1.000   B02    :  19   barley       :425   Min.   :0.0000  
##  1st Qu.:3.200   1st Qu.: 5.000   B17    :  19   barley+lupins:459   1st Qu.:0.0000  
##  Median :3.410   Median : 9.000   B24    :  19   lupins       :453   Median :0.0000  
##  Mean   :3.422   Mean   : 9.185   B09    :  19                       Mean   :0.3179  
##  3rd Qu.:3.630   3rd Qu.:13.000   B11    :  19                       3rd Qu.:1.0000  
##  Max.   :4.590   Max.   :19.000   B05    :  19                       Max.   :1.0000  
##                                   (Other):1223                                       
##      lupins      
##  Min.   :0.0000  
##  1st Qu.:0.0000  
##  Median :0.0000  
##  Mean   :0.3388  
##  3rd Qu.:1.0000  
##  Max.   :1.0000  
## 
```



GAMM

Now we build the first model. We have 79 cows and three groups: **barley, lupins** and **barley+lupins**.

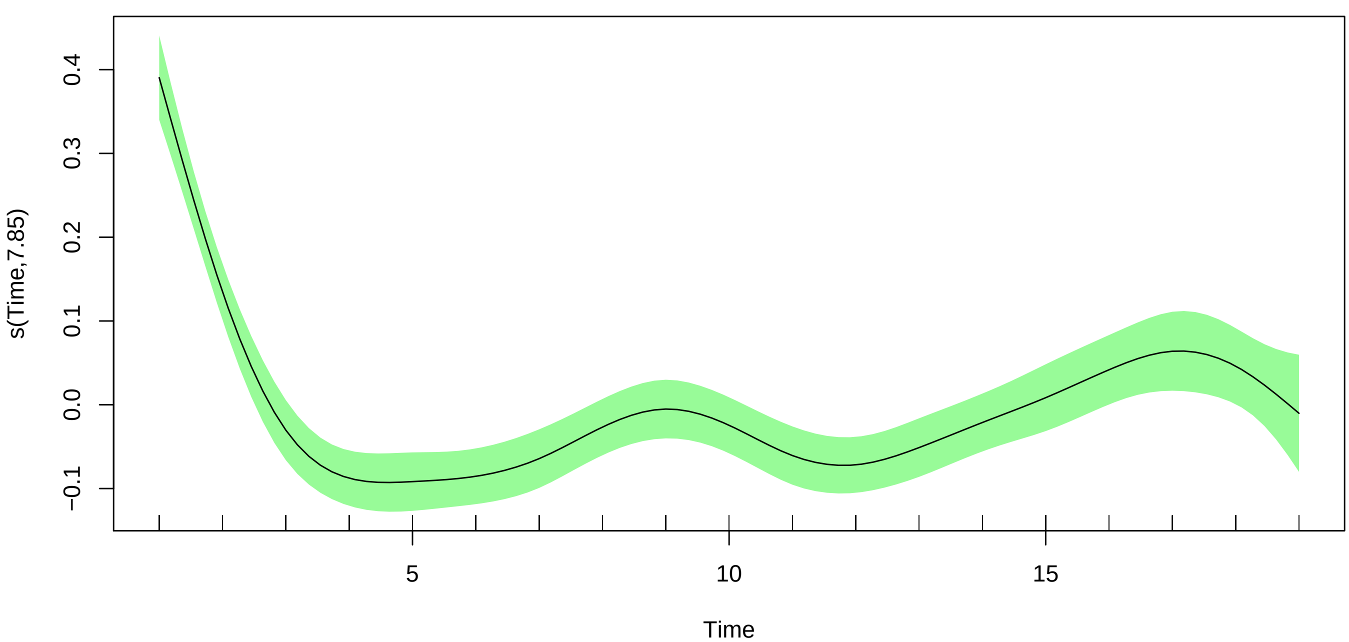
$$protein_{ij} = U_i + f(time_{ij}) + \beta_1 barley_i + \beta_2 lupins + \epsilon_{ij}$$

$$U_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \qquad \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$$

We fit GAMM with fixed **time**, **diet group** and random subject **cow number.

Now we plot and sum up centralized penelized spline.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## protein ~ s(Time) + lupins + barley
##
## Parametric coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  3.43083    0.03412 100.563 <0.0000000000000002 ***
## lupins      -0.10768    0.04827  -2.231              0.0259 *
## barley       0.09592    0.04920   1.949              0.0515 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##          edf Ref.df     F             p-value
## s(Time) 7.85   7.85 35.72 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.188
##   Scale est. = 0.061831  n = 1337
```

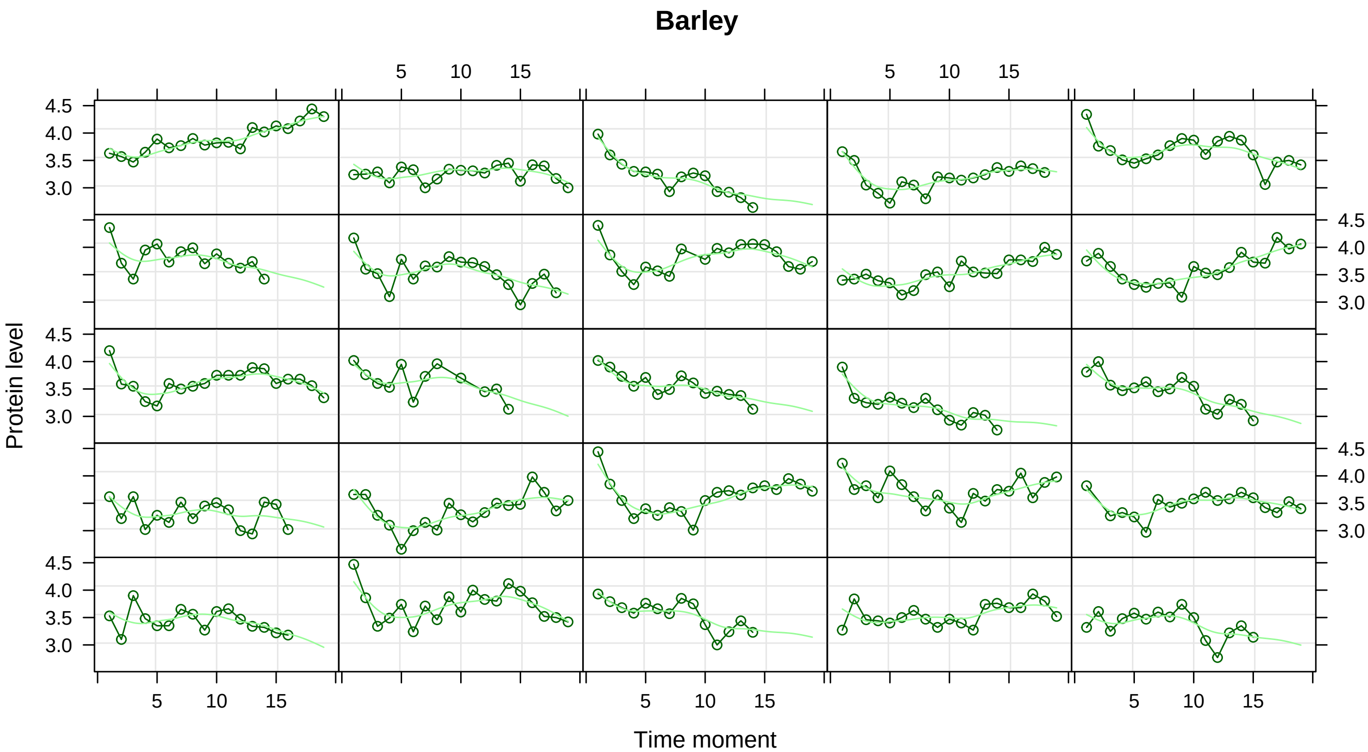


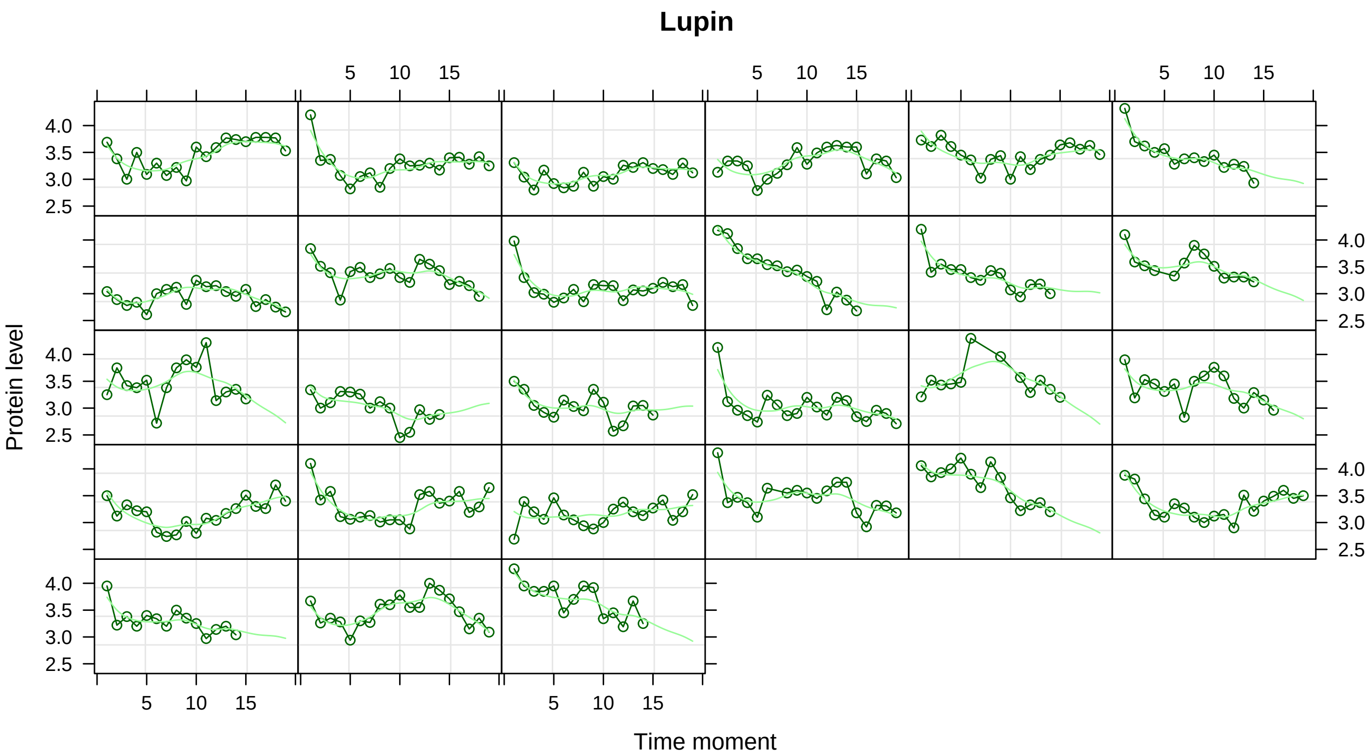
We could observe low R^2. All predictors are useful regarding t-test

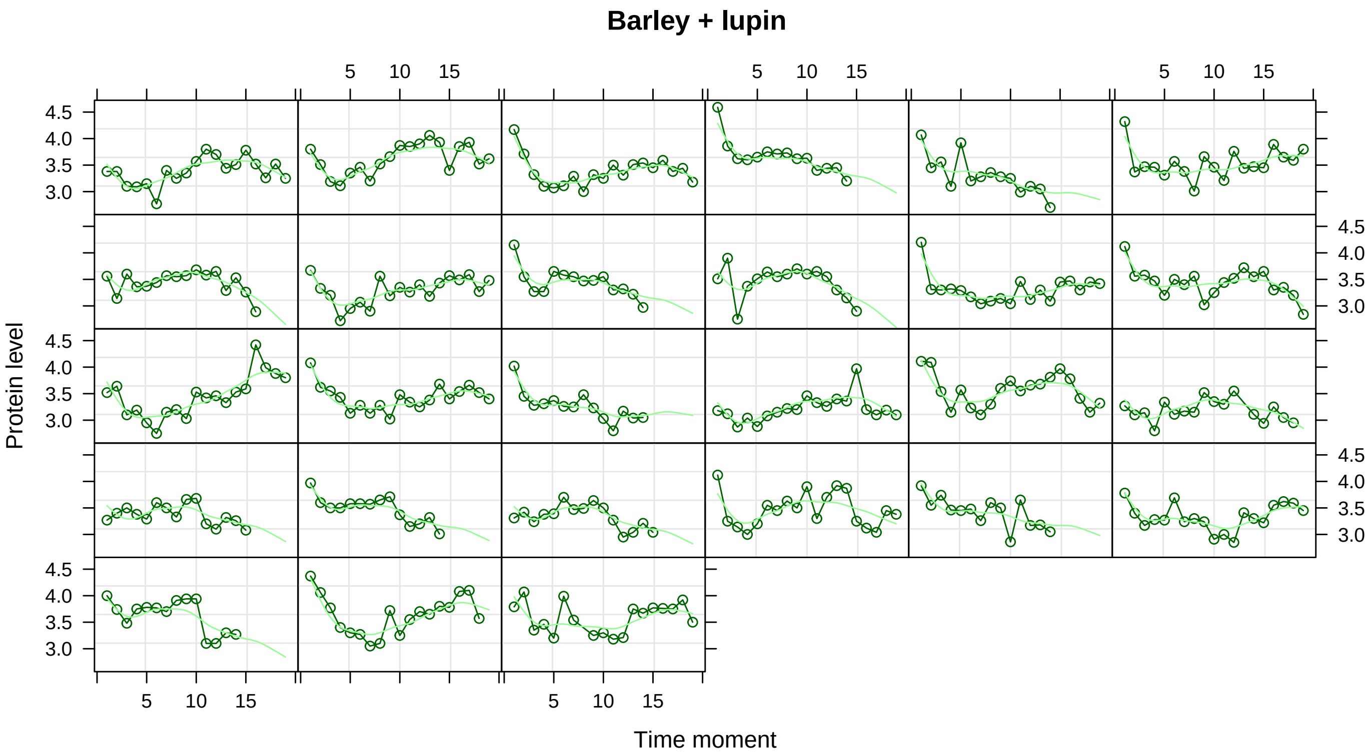
Group specific splines

Here we focus on group-specific curvs

As we could see below models seems to be rather well fitted, but there are some outlier visible.


Barley
Protein level
Time moment


Lupin
Protein level
Time moment


Barley + lupin
Protein level
Time moment
4.5
4.0
3.5
3.0
5
10
15