Flexible Regression and Smoothing:

Using GAMLSS in R

Mikis Stasinopoulos, Rob Rigby, Gilllian Heller,
Vlasios Voudouris and Fernanda De Bastiani

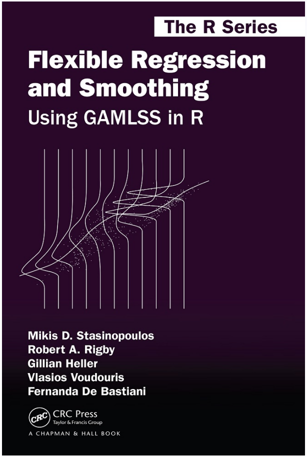
Graz University of Technology, Austria, November 2016

gamlss

About the course

1. Day 1 (Morning)
   - Why GAMLSS?
   - Introduction to the R packages, Diagnostics and Algorithms
   - Practical
2. Day 1 (Afternoon)
   - The `gamlss.family` distributions (Continuous, Discrete and Mixed distributions)
   - Practical
3. Day 2 (Morning)
   - Additive terms (linear, smoothing and random effects)
   - Practical
4. Day 2 (Afternoon)
   - Model selection
   - Centile estimation
   - Practical

Information



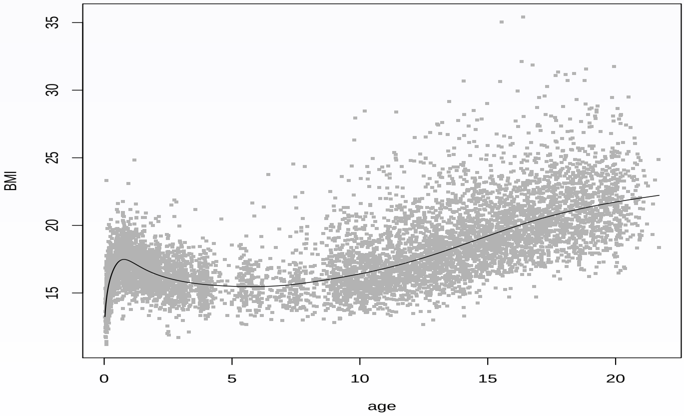
The Dutch boys data

`BMI` : the BMI of 7294 boys

`age` : the age in years

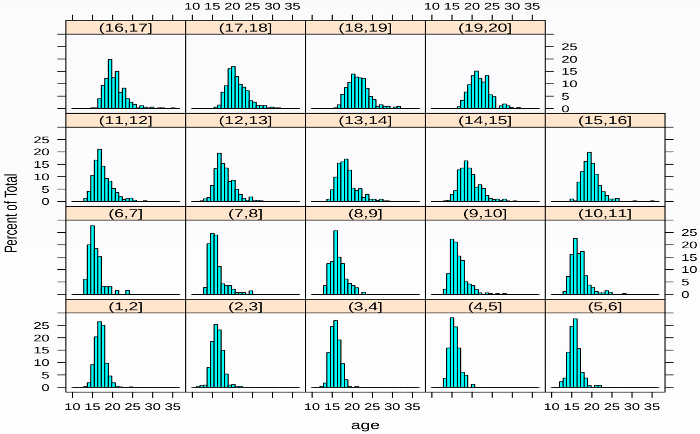
Source: van Buuren and Fredriks (2001)

The Dutch boys data: statistical challenges



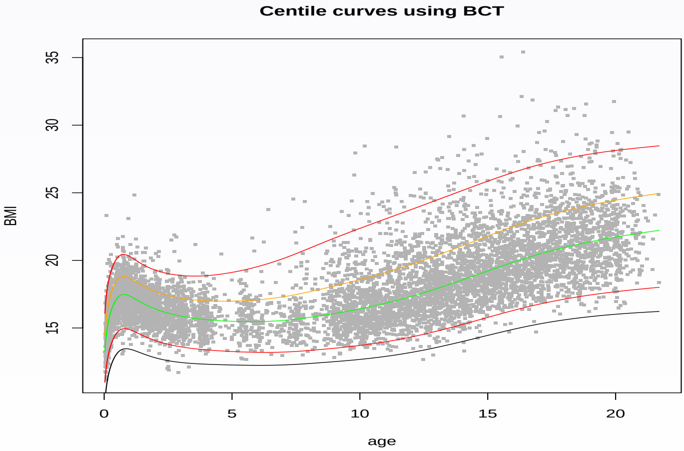
gamlss

The Dutch boys data: Histograms by age

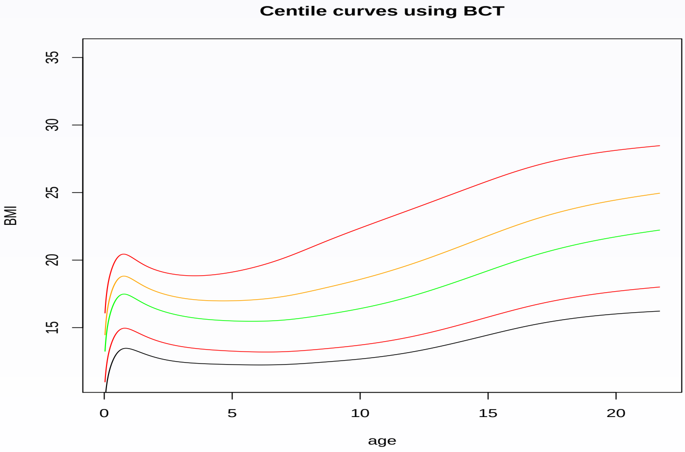


gamlss

The Dutch boys data: centile estimation



The Dutch boys data: centiles

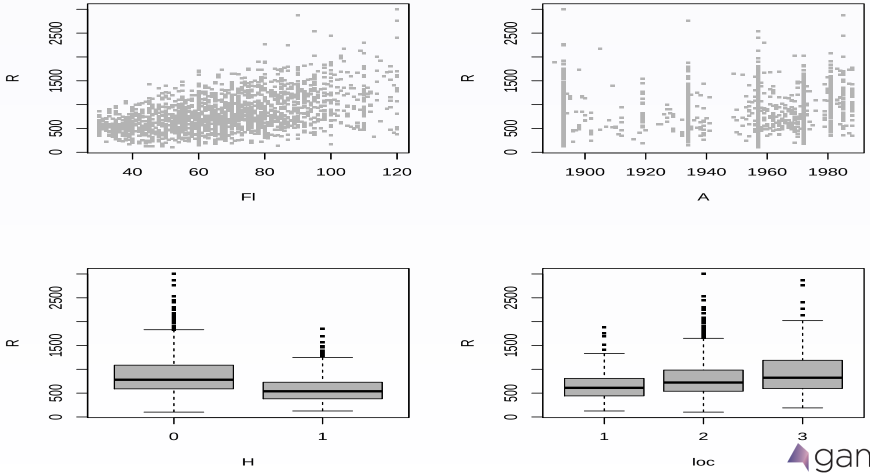


The Munich rent data

- `R` : the monthly net rent for flats in the city of Munich.
- `Fl` : the floor space area in square meters
- `A` : year of construction
- `loc` : whether the location is below, 1, average, 2, or above average 3
- `H` : two level factor indicating whether there is central heating, (0), or not, (1).

Source: Munich rental guide 1993

The Munich rent data



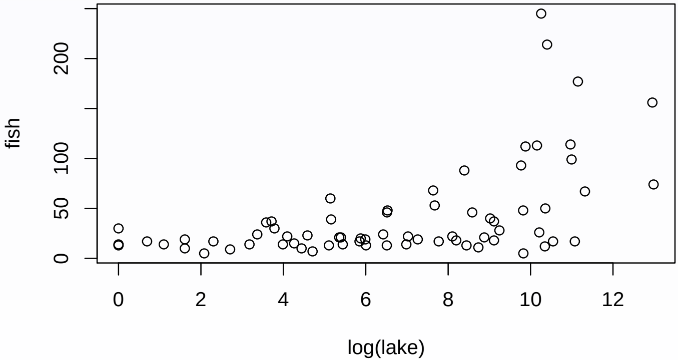
The fish species data

`fish` : the number of different species in 70 lakes in the world

`lake` : the lake area

Source: Stein and Juritz (1988)

The fish species data



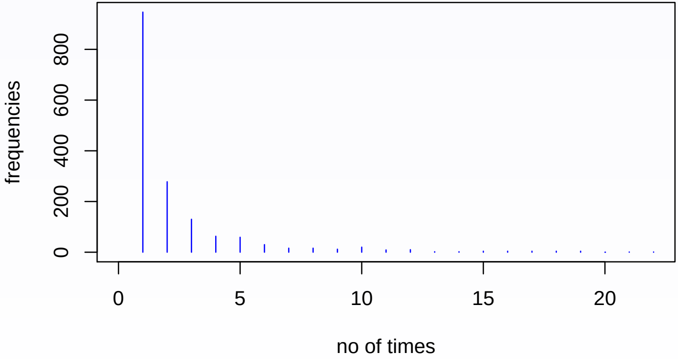
A stylometric application

64 observations

- `word` : is the number of times a word appears in a single text
- `freq` : the number of different words which occur exactly `word` times in the text

Source: Prof. Mario Cortina-Borja

The stylometric data

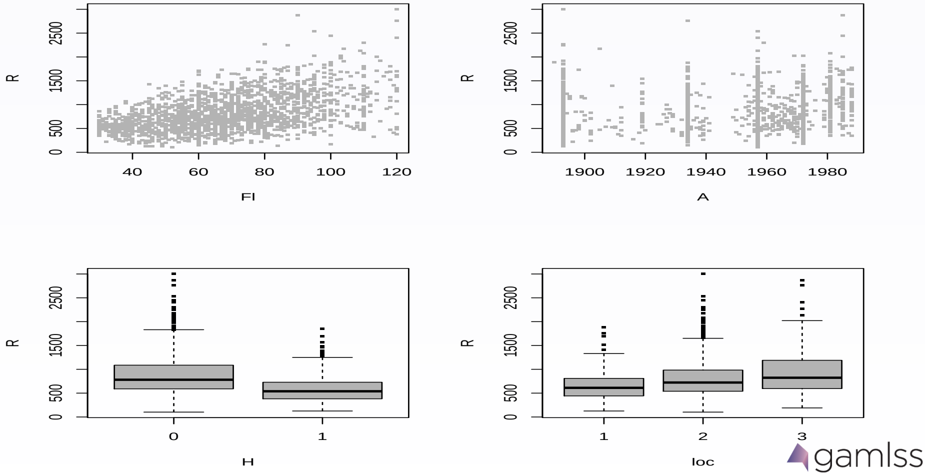


What we need for modelling the above data?

We need

- flexible distributions for the response variable
- to be able to deal with heterogeneity in the data
- to be able to model skewness and kurtosis
- to be able to model overdispersion in count data
- We need modelling all the parameters of the distributions
- flexible functions to model the relationship between the parameter of the distribution and the explanatory variables

The Munich rent data



gamlss

The Munich rent data: Linear model

Model assumptions

$$
\begin{aligned}
\mathbf{y} &\stackrel{\text{ind}}{\sim} N(\boldsymbol{\mu}, \sigma^2). \\
\boldsymbol{\mu} &= \mathbf{X}\boldsymbol{\beta}
\end{aligned}
$$

Estimation

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\epsilon}}^\top \hat{\boldsymbol{\epsilon}}}{n}$$

The linear model assumptions

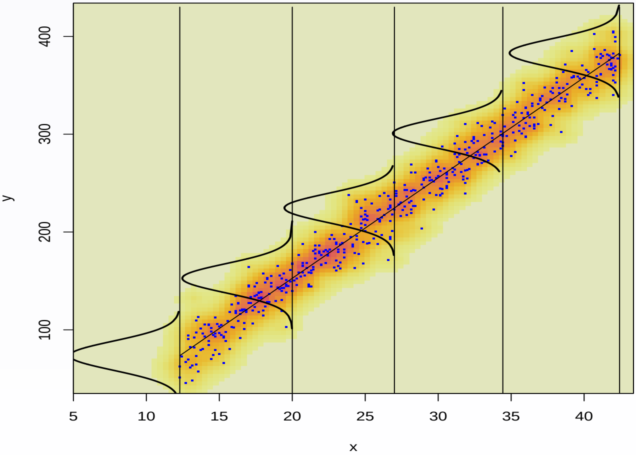
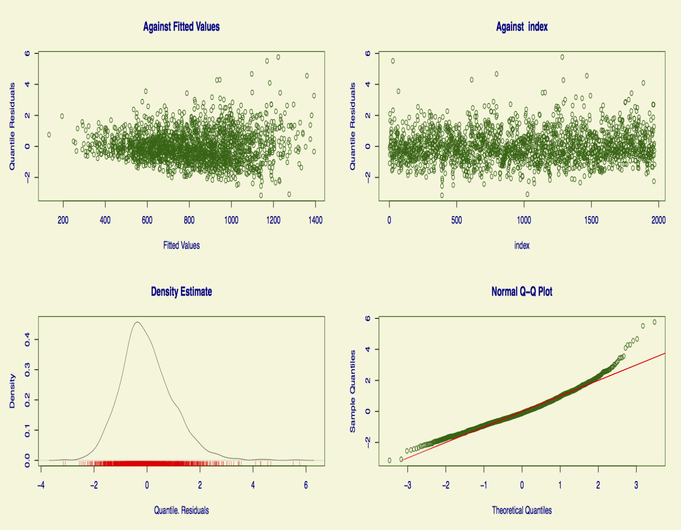


Figure: The linear regression model assumptions

The Munich rent data: Linear model

```
r1 <-  gamlss(R ~ Fl+A+H+loc, family=NO,  data=rent)

## GAMLSS-RS iteration 1: Global Deviance = 28159
## GAMLSS-RS iteration 2: Global Deviance = 28159

l1 <- lm(R ~ Fl+A+H+loc,data=rent)
coef(r1)

##  (Intercept)          Fl           A          H1        loc2
## -2775.038803    8.839445    1.480755 -204.759562  134.052349
##         loc3
##   209.581472

coef(l1)

##  (Intercept)          Fl           A          H1        loc2
## -2775.038803    8.839445    1.480755 -204.759562  134.052349
##         loc3
##   209.581472
```

The Munich rent data: LM residuals plot



The Munich rent data: Generalised Linear Model

The model

$$\begin{aligned} \mathbf{y} &\overset{\text{ind}}{\sim} \mathit{ExpFamily}(\boldsymbol{\mu}, \phi) \\ g(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

The exponential family

$$f_Y(y; \mu, \sigma) = \exp\left\{ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right\}$$

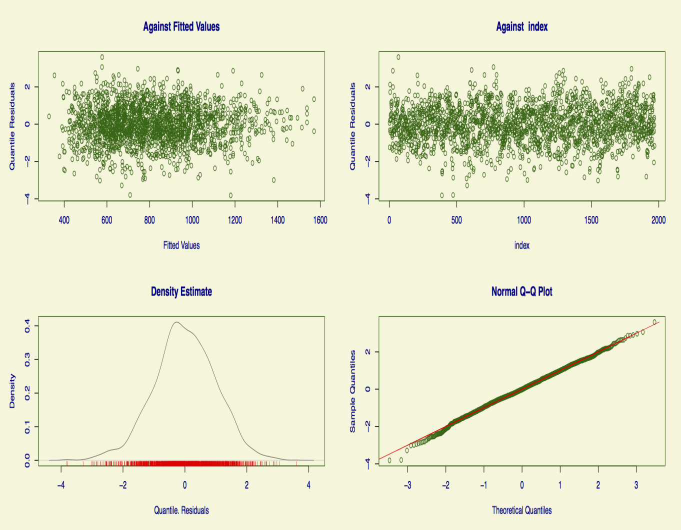
The Munich rent data: GLM fit

```
l2 <- glm(R ~ Fl+A+H+loc, family=Gamma(link="log"),  data=rent)
r2 <-  gamlss(R ~ Fl+A+H+loc, family=GA,  data=rent)

## GAMLSS-RS iteration 1: Global Deviance = 27764.59
## GAMLSS-RS iteration 2: Global Deviance = 27764.59
```

The Munich rent data: GLM and GAIC

```
r22 <-  gamlss(R ~ Fl+A+H+loc, family=IG,  data=rent)

## GAMLSS-RS iteration 1: Global Deviance = 27991.56
## GAMLSS-RS iteration 2: Global Deviance = 27991.56

GAIC(r1, r2, r22) # AIC

##     df      AIC
## r2   7 27778.59
## r22  7 28005.56
## r1   7 28173.00

GAIC(r1, r2, r22, k=log(length(rent$R))) # SBC or BIC

##     df      AIC
## r2   7 27817.69
## r22  7 28044.66
## r1   7 28212.10
```

The Munich rent data: GLM residuals plot



gamlss

The Munich rent data: Generalised Additive Model

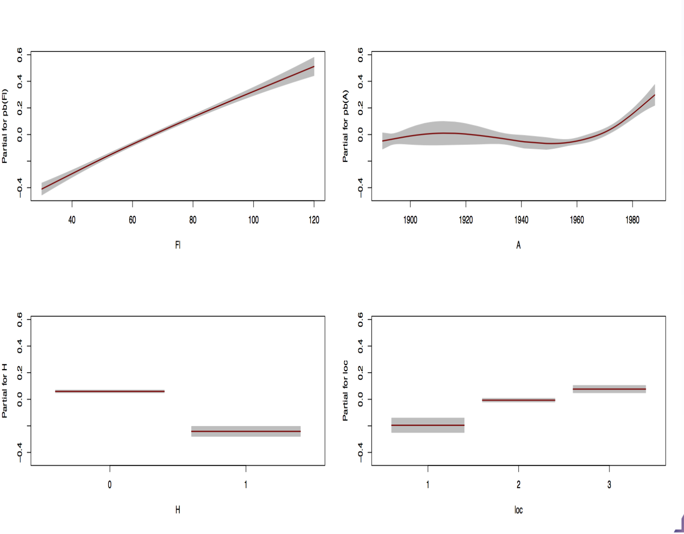
The model

$$
\begin{aligned}
\mathbf{y} &\overset{\text{ind}}{\sim} ExpFamily(\boldsymbol{\mu}, \phi) \\
g(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta} + s_1(\mathbf{x}_1) + \ldots + s_J(\mathbf{x}_J)
\end{aligned}
$$

The Munich rent data: GAM commands

```
r3 <-  gamlss(R ~ pb(Fl)+pb(A)+H+loc, family=GA,  data=rent)

## GAMLSS-RS iteration 1: Global Deviance = 27683.22
## GAMLSS-RS iteration 2: Global Deviance = 27683.22
## GAMLSS-RS iteration 3: Global Deviance = 27683.22

AIC(r2,r3)

##          df      AIC
## r3 11.21547 27705.65
## r2  7.00000 27778.59
```

The Munich rent data: GAM term plot



The Munich rent data: GAM summary

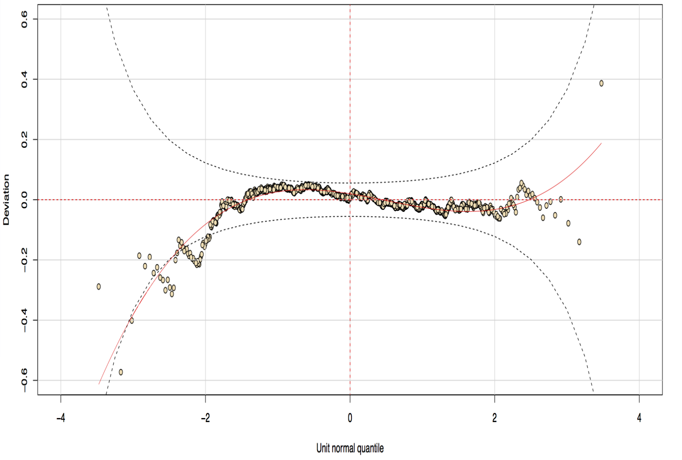
```
summary(r3)

## ******************************************************************
## Family:  c("GA", "Gamma")
##
## Call:
## gamlss(formula = R ~ pb(Fl) + pb(A) + H + loc, family = GA, data = rent)
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.0851197  0.5666171   5.445 5.84e-08 ***
## pb(Fl)       0.0103084  0.0004030  25.578  < 2e-16 ***
## pb(A)        0.0014062  0.0002879   4.884 1.12e-06 ***
## H1          -0.3008111  0.0225705 -13.328  < 2e-16 ***
## loc2         0.1886692  0.0299153   6.307 3.51e-10 ***
## loc3         0.2719856  0.0322699   8.428  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

gamlss

The Munich rent data: GAM drop one term

```
drop1(r3)

## Single term deletions for
## mu
##
## Model:
## R ~ pb(Fl) + pb(A) + H + loc
##             Df   AIC    LRT   Pr(Chi)
## <none>          27706
## pb(Fl) 1.4680 28261 558.59 < 2.2e-16 ***
## pb(A)  4.3149 27798 101.14 < 2.2e-16 ***
## H      1.8445 27862 160.39 < 2.2e-16 ***
## loc    2.0346 27770  68.02 1.825e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Munich rent data: GAM worm plot



gamlss

The Munich rent data: Mean and dispersion additive models

The model

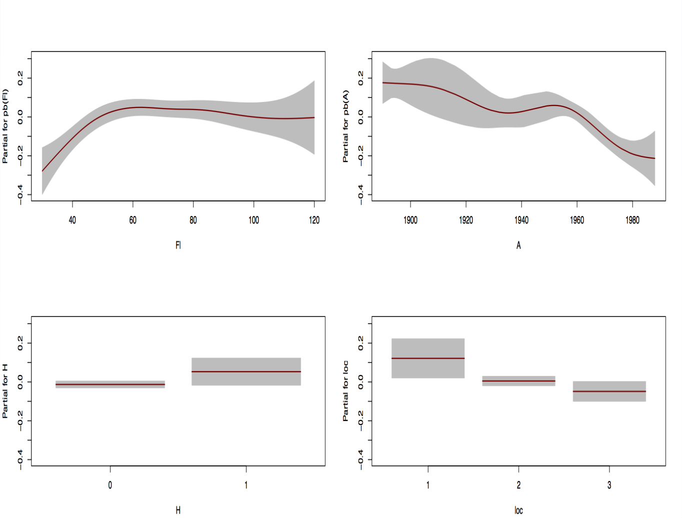
$$
\begin{aligned}
\mathbf{y} &\overset{\text{ind}}{\sim} D(\boldsymbol{\mu}, \boldsymbol{\sigma}) \\
g_1(\boldsymbol{\mu}) &= \mathbf{X}_1 \boldsymbol{\beta}_1 + s_{11}(\mathbf{x}_{11}) + \ldots + s_{1J_1}(\mathbf{x}_{1J_1}) \\
g_2(\boldsymbol{\sigma}) &= \mathbf{X}_2 \boldsymbol{\beta}_2 + s_{21}(\mathbf{x}_{21}) + \ldots + s_{2J_2}(\mathbf{x}_{2J_2})
\end{aligned}
$$

The Munich rent data: MADAM commands

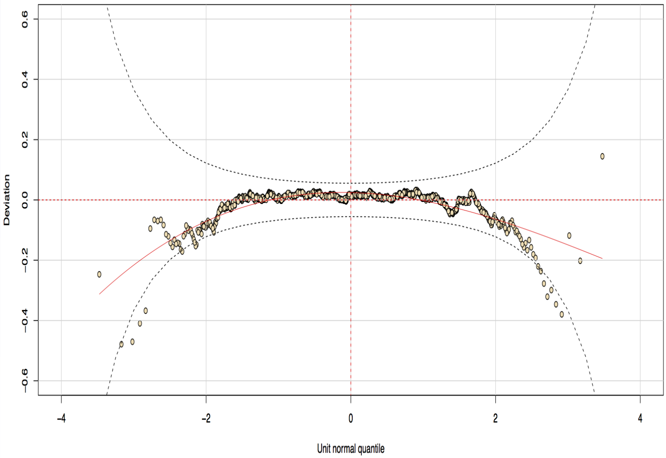
```
r4 <-  gamlss(R ~ pb(Fl)+pb(A)+H+loc, sigma.fo=~pb(Fl)+pb(A)+H+loc, family=GA,
              data=rent)

## GAMLSS-RS iteration 1: Global Deviance = 27572.14
## GAMLSS-RS iteration 2: Global Deviance = 27570.29
## GAMLSS-RS iteration 3: Global Deviance = 27570.28
## GAMLSS-RS iteration 4: Global Deviance = 27570.28

r5 <-  gamlss(R ~ pb(Fl)+pb(A)+H+loc, sigma.fo=~pb(Fl)+pb(A)+H+loc, family=IG,
              data=rent)

## GAMLSS-RS iteration 1: Global Deviance = 27675.74
## GAMLSS-RS iteration 2: Global Deviance = 27672.97
## GAMLSS-RS iteration 3: Global Deviance = 27673
## GAMLSS-RS iteration 4: Global Deviance = 27673.01
## GAMLSS-RS iteration 5: Global Deviance = 27673.01
## GAMLSS-RS iteration 6: Global Deviance = 27673.02

AIC(r3, r4, r5)

##          df      AIC
## r4 22.25035 27614.78
## r3 11.21547 27705.65
## r5 21.82318 27716.66
```

gamlss

The Munich rent data: MADAM term plot for σ



The Munich rent data: MADAM worm plot



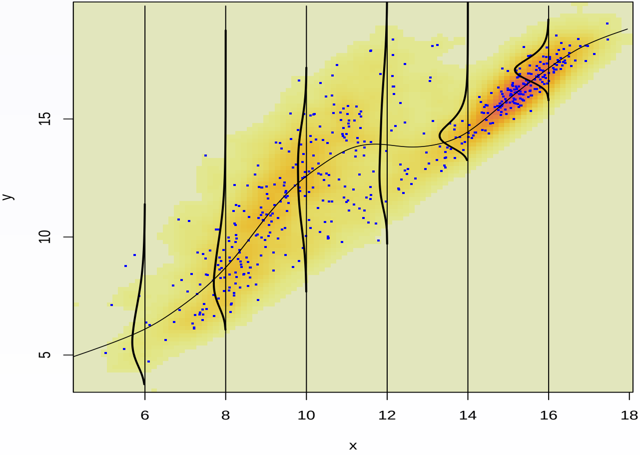
The Munich rent data: GAMLSS

The model

$$
\begin{aligned}
\mathbf{y} &\overset{\text{ind}}{\sim} D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\
g_1(\boldsymbol{\mu}) &= \mathbf{X}_1 \boldsymbol{\beta}_1 + s_{11}(\mathbf{x}_{11}) + \ldots + s_{1J_1}(\mathbf{x}_{1J_1}) \\
g_2(\boldsymbol{\sigma}) &= \mathbf{X}_2 \boldsymbol{\beta}_2 + s_{21}(\mathbf{x}_{21}) + \ldots + s_{2J_2}(\mathbf{x}_{2J_2}) \\
g_3(\boldsymbol{\nu}) &= \mathbf{X}_3 \boldsymbol{\beta}_3 + s_{31}(\mathbf{x}_{31}) + \ldots + s_{3J_3}(\mathbf{x}_{3J_3}) \\
g_4(\boldsymbol{\tau}) &= \mathbf{X}_4 \boldsymbol{\beta}_4 + s_{41}(\mathbf{x}_{41}) + \ldots + s_{4J_4}(\mathbf{x}_{4J_4})
\end{aligned}
$$

gamlss

The Munich rent data: GAMLSS assumptions



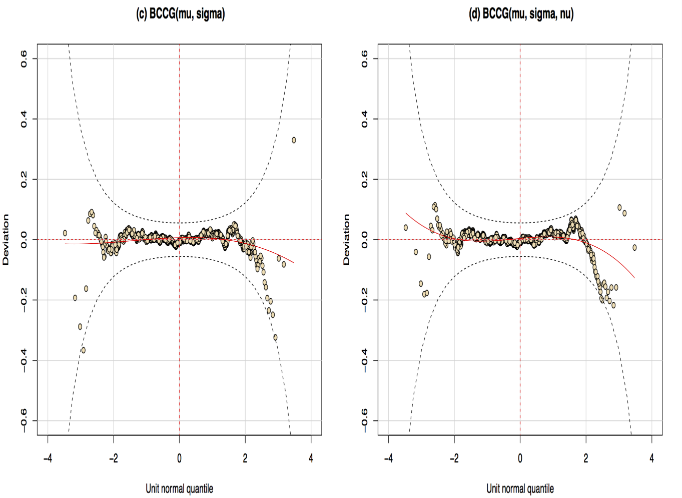
The Munich rent data: GAMLSS commands

```
r6 <-  gamlss(R ~ pb(Fl)+pb(A)+H+loc, sigma.fo=~pb(Fl)+pb(A)+H+loc,
              nu.fo=~1, family=BCCGo,  data=rent)
```

```
r7 <-  gamlss(R ~ pb(Fl)+pb(A)+H+loc,sigma.fo=~pb(Fl)+pb(A)+H+loc,
              nu.fo=~pb(Fl)+pb(A)+H+loc, family=BCCGo,  data=rent)
```

```
AIC(r4, r6, r7)

##          df      AIC
## r7 28.41391 27608.15
## r6 22.48092 27611.02
## r4 22.25035 27614.78
```

The Munich rent data: GAMLSS worm plot



What is GAMLSS?

GAMLSS: are semi-parametric regression type models.

- regression type: we have many explanatory variables $\mathbf{X}$ and one response variable $\mathbf{y}$ and we believe that $\mathbf{X} \rightarrow \mathbf{y}$
- parametric: a parametric distribution assumption for the response variable,
- semi: the parameters of the distribution, as functions of explanatory variables, may involve non-parametric smoothing functions
- GAMLSS philosophy: try different models

GAMLSS is a generalisation of GLM and GAM models.

Conclusions

- GAMLSS is a very flexible statistical model
- It is a unified framework for univariate regression type of models
- Allows any distribution for the response variable Y
- Models all the parameters of the distribution of Y
- Allows a variety of penalised additive terms in the models for the distribution parameters
- The fitted algorithm is modular, where different components can be added easily
- it can easily introduced to students since it relies on known concepts
- It deals with overdispersion, skewness and kurtosis

This is a collaborative work: A list of people who help with the R implementation

present	past
Vlasios Voudouris	Popi Akantziliotou
Paul Eilers	Nicoleta Mortan
Gillian Heller	Fiona McElduff
Marco Enea	Raydonal Ospina
Majid Djennad	Konstantinos Pateras
Fernanda De Bastiani	
Luiz Nakamura	
Daniil Kiose	
Andreas Mayr	
Thomas Kneib	
Nadja Klein	
Abu Hossain	

There are lot more to be done

the END

for more information see

www.gamlss.org