Flexible Regression and Smoothing:

Using GAMLSS in R

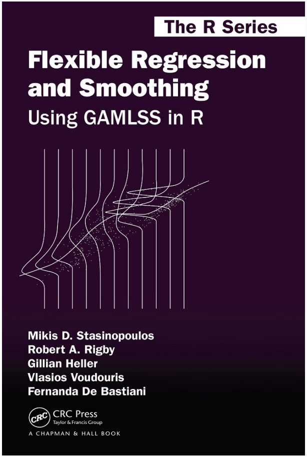
Mikis Stasinopoulos, Rob Rigby, Gilllian Heller, and Fernanda De Bastiani

BCAM – Basque Center for Applied Mathematics, Bilbao, Basque Country - Spain, October 2019

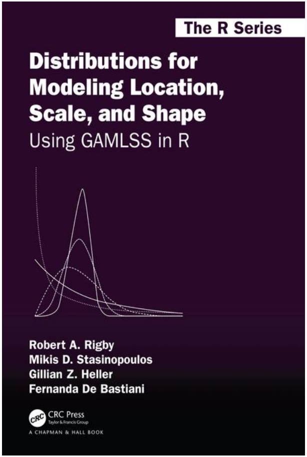
About the course

1. Day 1
   - Why GAMLSS?
   - Introduction to the R packages
2. Day 2
   - The `gamlss.family` distributions.
   - Continuous distributions in GAMLSS
3. Day 3
   - Discrete distributions in GAMLSS
   - Mixed distributions in GAMLSS
4. Day 4
   - Additive terms [linear, smoothing and random effects]
   - Model selection
5. Day 5
   - Case study
   - Centile estimation

Book 1: about GAMLSS and R packages



Book 2: about distributions in GAMLSS



Book 3: about different modes of estimating GAMLSS

in preparation:

Generalized Additive Models for Location, Scale and Shape
A Distributional Regression Approach

Authors: Mikis Stasinopoulos, Gillian Heller, Andreas Mayr, Thomas Kneib, Nadja Klein

More material about GAMLSS from

www.gamlss.com

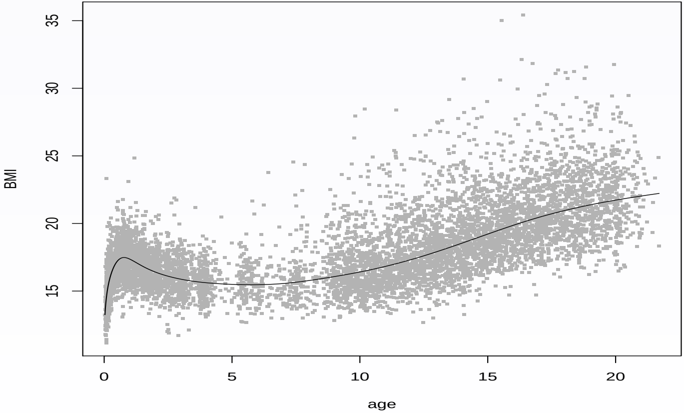
The Dutch boys data

`BMI` : the BMI of 7294 boys

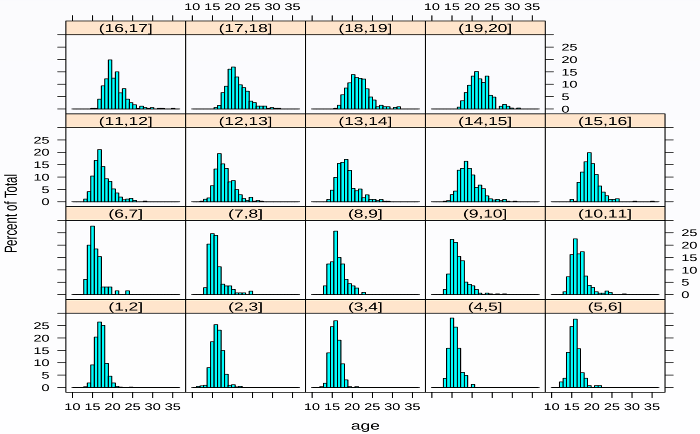
`age` : the age in years

Source: van Buuren and Fredriks (2001)

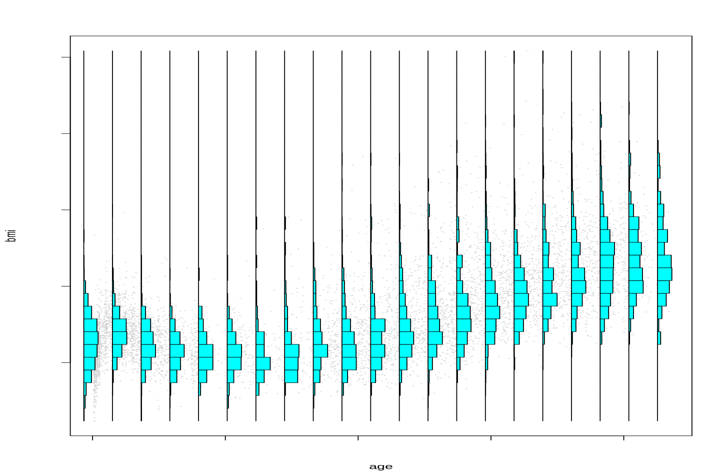
The Dutch boys data: statistical challenges



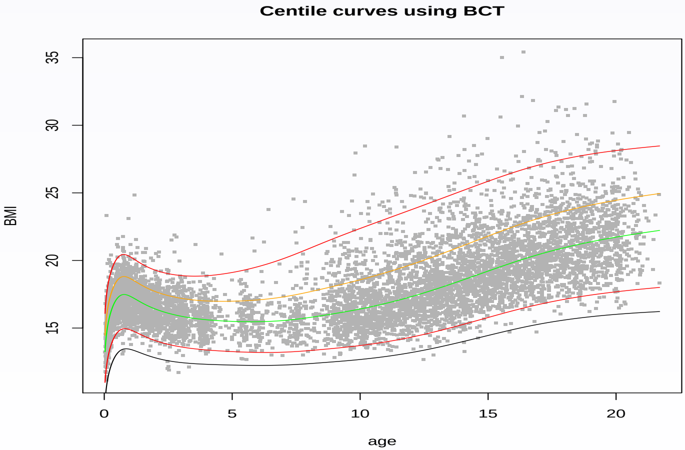
The Dutch boys data: Histograms by age



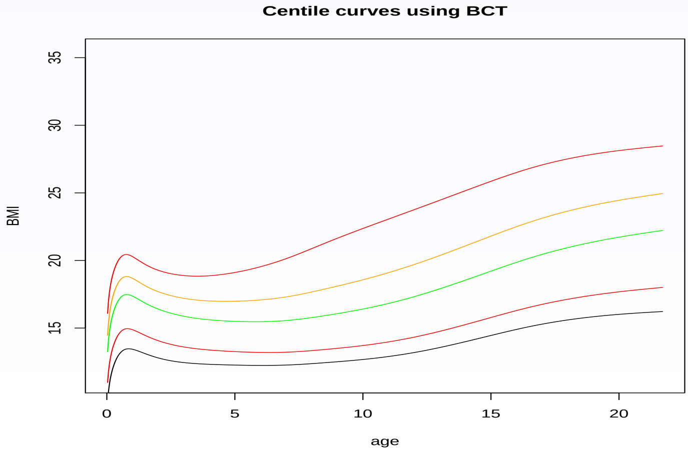
The Dutch boys data: Conditional histograms by age



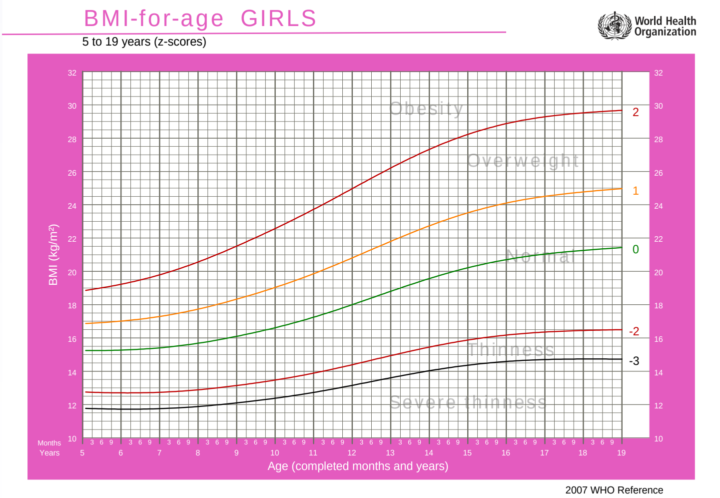
The Dutch boys data: centile estimation



The Dutch boys data: centiles

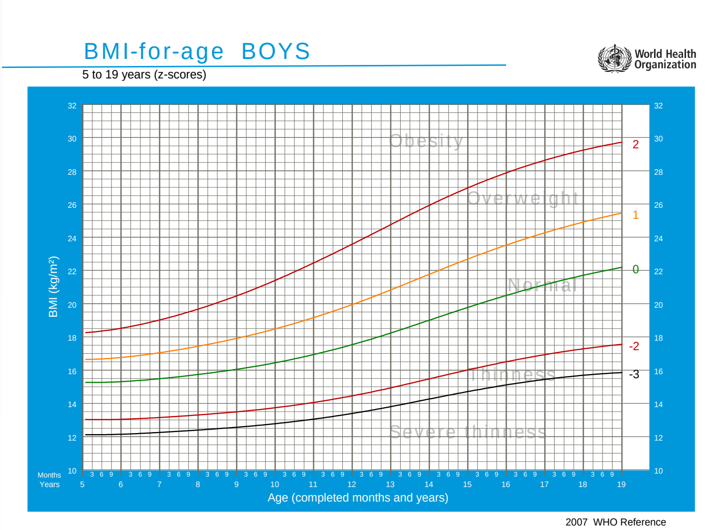


World Health Organisation Child Growth Standards: Girls



gamlss

World Health Organisation Child Growth Standards: Boys



gamlss

The film data

4031 film openings in US 1988-1999:

- `lborev1` : the response variable
  log of the (total revenues - the first week revenue), calculated in 1987 prices
- `lboopen` : the log the first week revenue, calculated in 1987 prices
- `lnosc` : the log of the number of screens
- `dist` : the type of distributor i.e. whether independent or major distributor

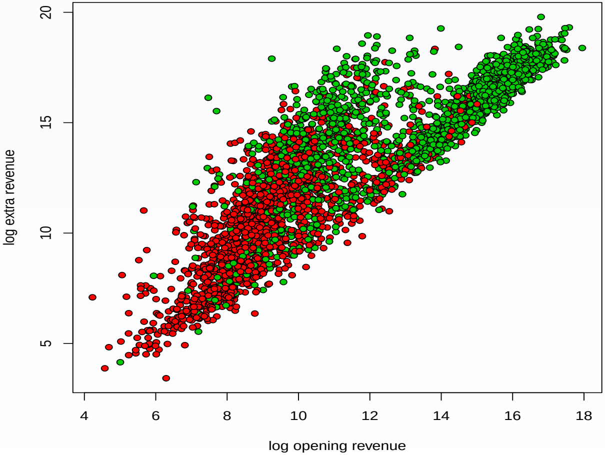
Source: Voudouris *et al.* (2010) Go to Quiz

gamlss

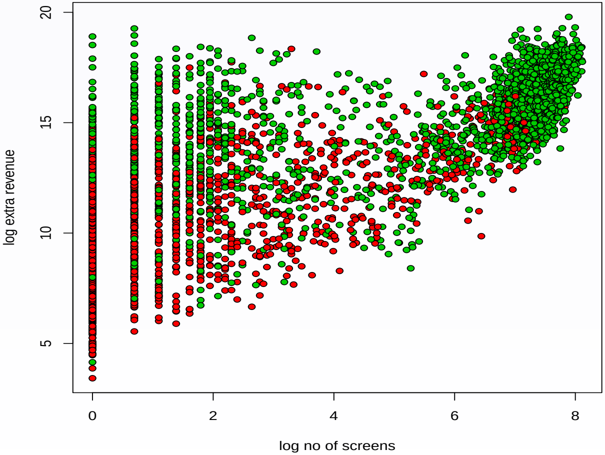
The film data: Question of interest

- Checking whether the "nobody knows anything" principle is valid
- Can we predict the extra revenue given the revenue at the first week?
- Which other explanatory variables are important?
- What is the conditional distribution of the response variable given the explanatory variables?

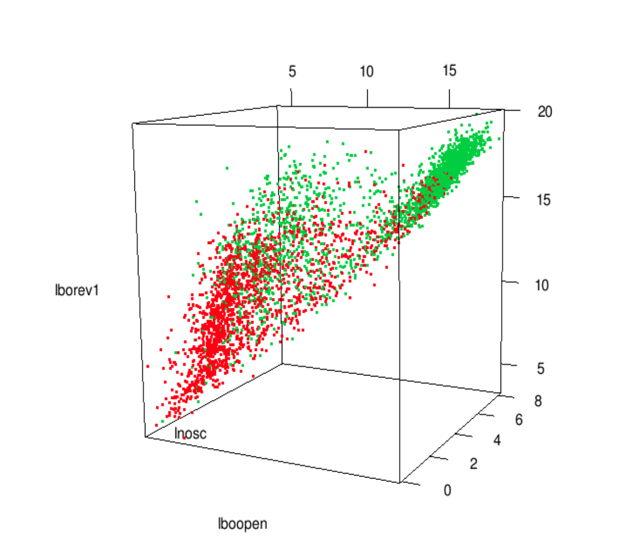
The film data: log extra revenue against log first week revenue



The film data: log extra revenue against log number of screens

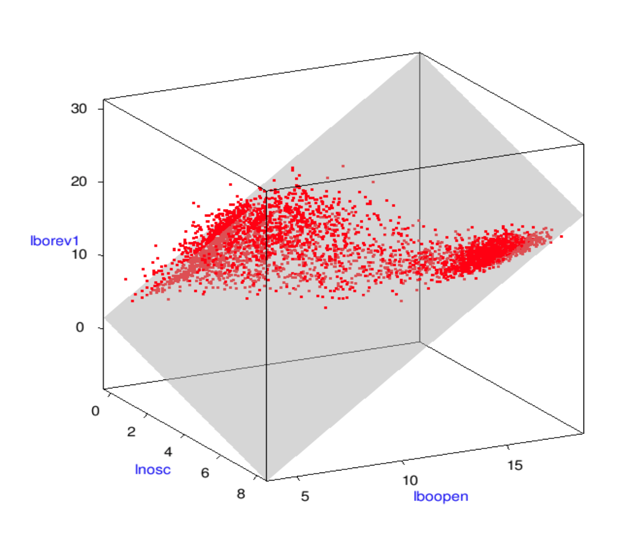


The film data: 3d plots

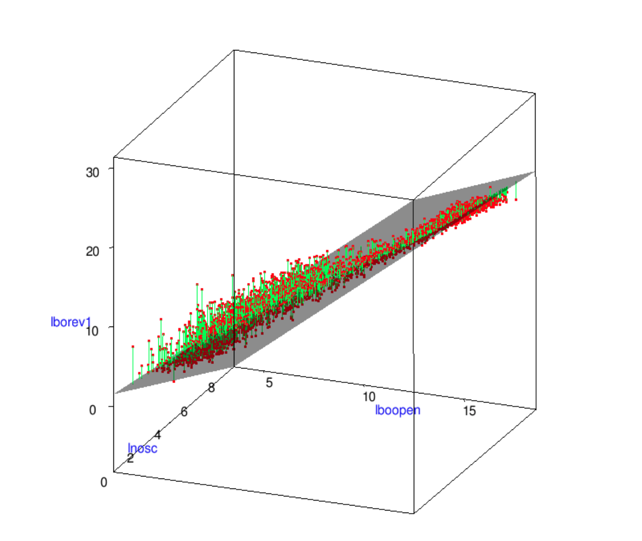


gamlss

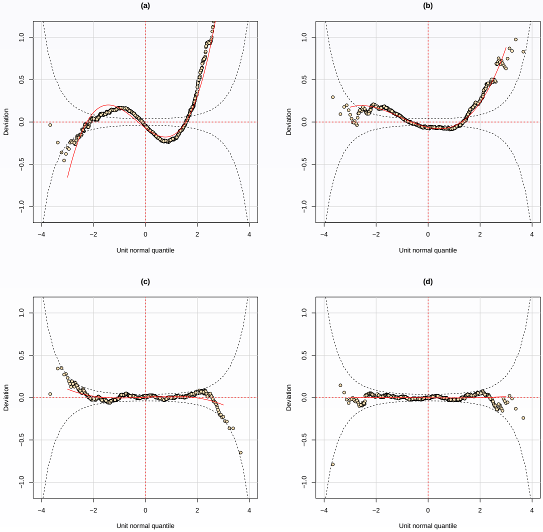
The film data: Least Square fit



The film data: the residuals from least squares fit



The film data: worm plot of the residuals



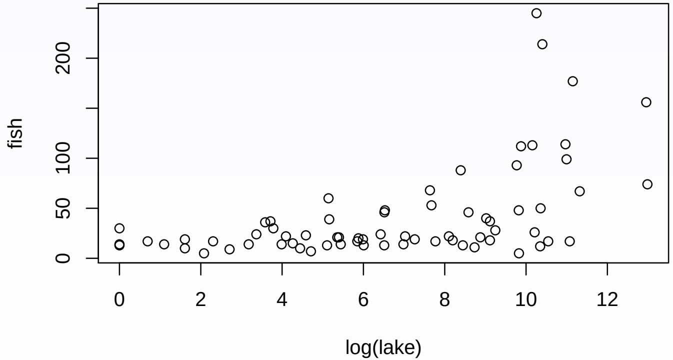
The fish species data

`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



gamlss

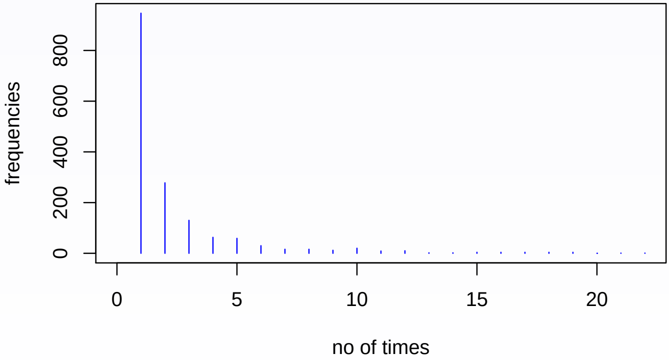
A stylometric application

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 statistical modelling philosophy

Statistical modelling is the art of building parsimonious statistical models for a better understanding of the phenomena of interest.

- Any model is a simplification of reality therefore no model is correct but some of them are useful

- Occam's Razor which states *'entities should not be multiplied beyond necessity'* or KISS (Keep It Simple Stupid)

The statistical modelling philosophy

- Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise. – John W. Tukey (1962), "The future of data analysis", Annals of Mathematical Statistics 33, 1-67 (page 13)

- *"no matter how beautiful your theory, no matter how clever you are or what your name is, if it disagrees with experiment, it's wrong"* (Richard Feynman)

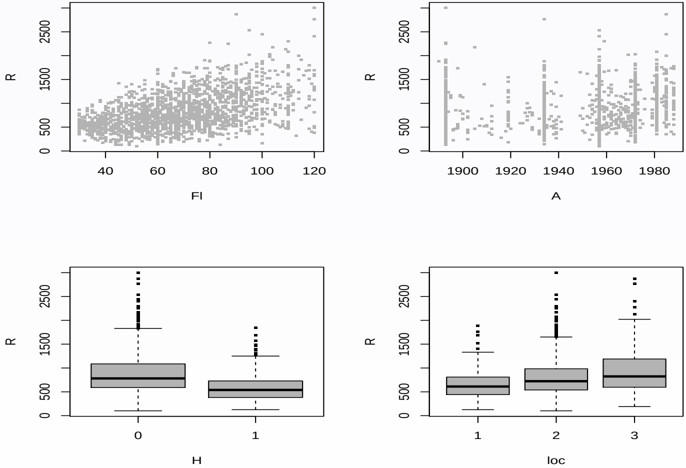
- As statisticians we should try different models and choose the one appropriate for the data

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 Munich rent data: Linear model

Model assumptions

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

Estimation

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

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

gamlss

The linear model assumptions

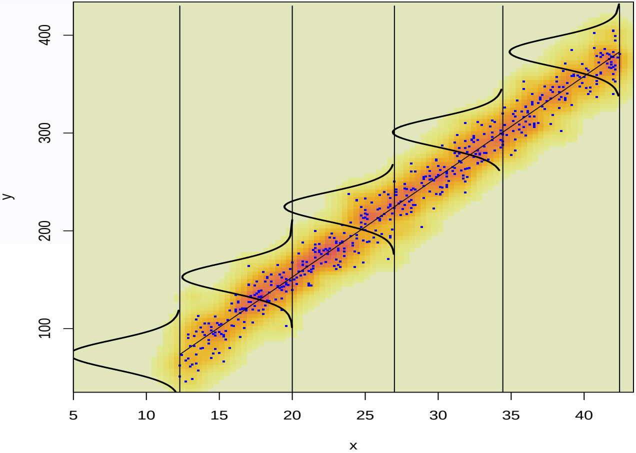
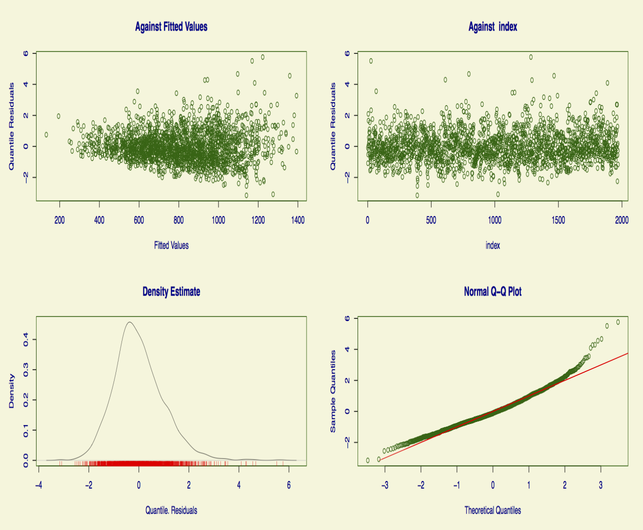


Figure: The linear regression model assumptions

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

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

The exponential family

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

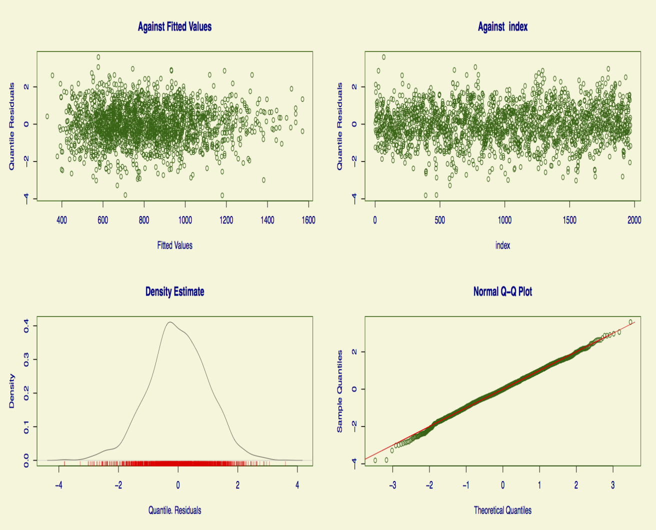
The Munich rent data: GLM fit

```
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



The Munich rent data: Generalised Additive Model

The model

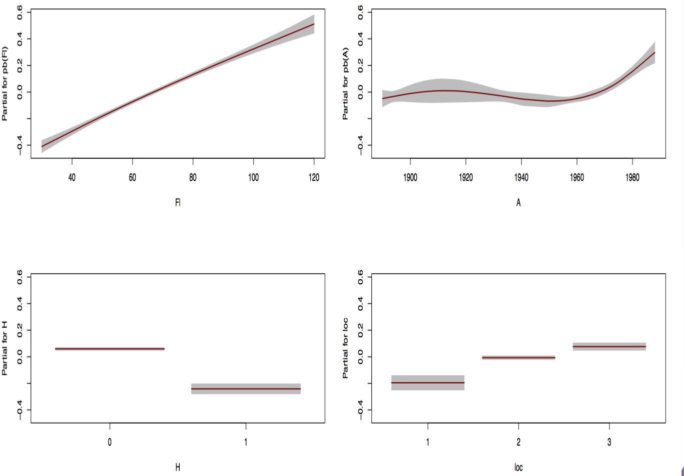
$$\begin{aligned} \mathbf{y} &\stackrel{\text{ind}}{\sim} \mathit{ExpFamily}(\boldsymbol{\mu}, \phi) \\ g(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta} + s_1(\mathbf{x}_1) + \ldots + s_J(\mathbf{x}_J) \end{aligned}$$

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
```

gamlss

The Munich rent data: GAM term plot



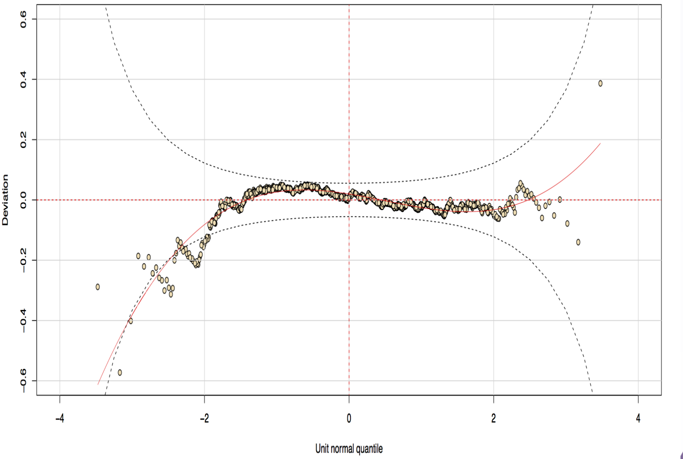
The Munich rent data: GAM summary

```
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
```

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

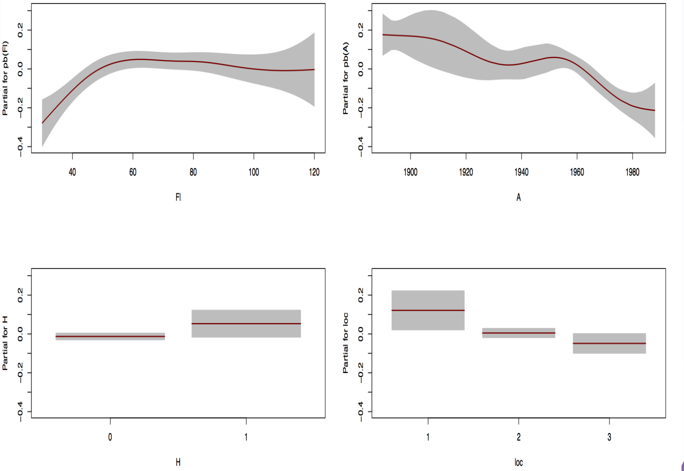
The model

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

The Munich rent data: MADAM commands

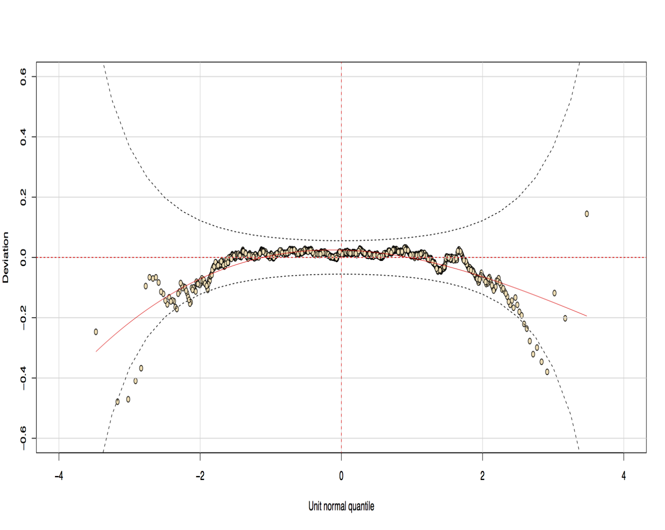
```
r4 <-  gamlss(R ~ pb(Fl)+pb(A)+H+loc, sigma.fo=~pb(Fl)+pb(A)+H+loc, family=GA,
              data=rent)

## 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
```

The Munich rent data: MADAM term plot for σ



gamlss

The Munich rent data: MADAM worm plot

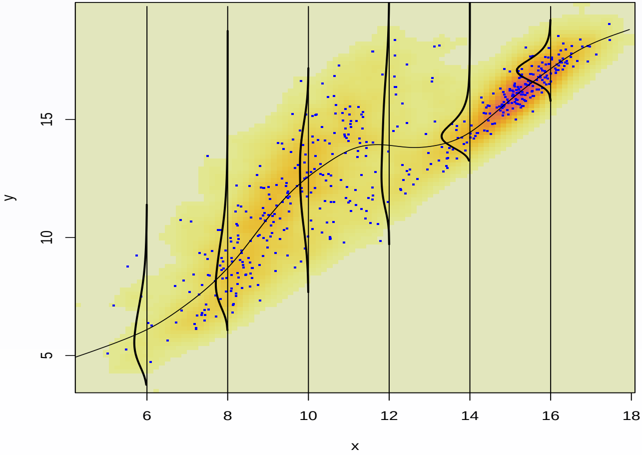


The Munich rent data: GAMLSS

The model

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

The Munich rent data: GAMLSS assumptions



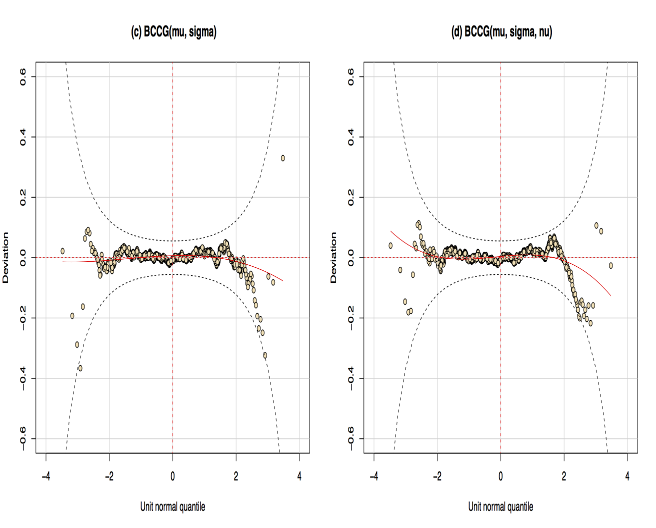
The Munich rent data: GAMLSS commands

```
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 distributional based semi-parametric regression type models.

- regression type: we have many explanatory variables $\mathbf{X}$ and one response variable $\mathbf{y}$ and we believe that $\mathbf{X} \rightarrow \mathbf{y}$
- distributional: a full parametric distribution assumption is made for the response variable
- semi-parametric: All the parameters of the distribution can be modelled as function of the explanatory variables. Those functions can be parametric or non-parametric smoothing functions
- GAMLSS philosophy: try different models

GAMLSS is a generalisation of GLM and GAM models.

The R packages

gamlss	the original package
gamlss.dist	all `gamlss.family` distributions
gamlss.data	different sets of data
gamlss.add	for extra additive terms
gamlss.cens	for censored (left, right or interval) response variables
gamlss.demo	demos for distributions and smoothing
gamlss.inf	models for inflated 0 to 1 or for adjusted 0 to ∞
gamlss.nl	non-linear term fitting
gamlss.tr	generating truncated distributions
gamlss.mx	finite mixtures distributions and random effects
gamlss.spatial	for spatial models [GMRF models]
gamlss.util	for extra utilities

The gamlss() function

```
gamlss( formula = ~1, sigma.formula= ~1,
      nu.formula = ~1, tau.formula = ~1,
      family = NO(),
      data = sys.parent(), weights = NULL,
      contrasts = NULL, method = RS(),  start.from = NULL,
      mu.start = NULL, sigma.start = NULL,
      nu.start = NULL, tau.start = NULL,
      mu.fix = FALSE, sigma.fix = FALSE, nu.fix = FALSE,
      tau.fix = FALSE, control = gamlss.control(...),
        i.control = glim.control(...), ...)
```

Arguments of the gamlss() function

$$
\begin{aligned}
\text{formula} &= y \sim x1 + x3 \\
\text{sigma.fo} &= \sim x1 \\
\text{nu.fo} &= \sim x2 \\
\text{tau.fo} &= \sim 1 \\
\text{data} &= \text{abdom} \\
\text{family} &= \text{LO} \\
\text{weights} &= \text{freq} \\
\text{method} &= \text{mixed(10,50)} \\
\text{control} &= \text{gamlss.control(trace=FALSE)}
\end{aligned}
$$

Starting values

Generally are not needed:

`start.mu= 2` or `start.mu=fitted(m1, "mu")`

The same applies for other parameters
`start.sigma` , `start.nu`, `start.tau`

Starting from a previous model

`start.from= model1`

The gamlss package

Available functions

- Fitting or Updating a Model
- Extracting Information from the Fitted Model
- Selecting a Model
- Plotting and Diagnostics
- Centile Estimation

Fitting or Updating a Model

- `gamlss()` for fitting and creating a `gamlss` object
- `refit()` to refit a `gamlss` object (i.e. continue iterations)
- `update()` to update a given `gamlss` model object
- `gamlssML()` fitting a parametric distribution to a single (response) variable
- `histDist()` to fit and plot a parametric distribution
- `chooseDist()` select a conditional parametric distribution for y for given x's
- `fitDist()` select a marginal parametric distribution for y from an appropriate list (with no explanatory variables)

Extracting Information from the Fitted Model

- `GAIC()` generalised Akaike information criterion (or AIC)
- `coef()` the linear coefficients
- `deviance()` the global deviance $-2 \log L$
- `fitted()` the fitted values for a distribution parameter
- `predict()` to predict from new data individual distribution parameter values
- `predictAll()` to predict from new data all the distribution parameter values
- `print()` : to print a `gamlss` object
- `residuals()` to extract the normalised (randomised) quantile residuals
- `summary()` to summarise the fit in a `gamlss` object
- `vcov()` to extract the variance-covariance matrix of the beta estimates.

Selecting a Model

`add1()` `drop1()` to add or drop a single term

`stepGAIC()` to select explanatory terms in one parameter

`stepGAICAll.A()` to select explanatory terms in all the parameters (strategy A)

`stepGAICAll.B()` to select explanatory terms in all the parameters (strategy B)

`gamlssCV()` to evaluate a model using cross validation

`add1TGD()` `drop1TGD()` to add or drop a single term using a validation or test data set

`stepTGD()` selecting variables using a test data set
the global deviance for new (test) data set given a fitted GAMLSS model.

Diagnostics

`plot()` a plot of four graphs for the normalized (randomized) quantile residuals

`pdf.plot()` for plotting the pdf functions for a given fitted `gamlss` object or a given `gamlss.family` distribution

`Q.stats()` for printing and plotting the Q statistics of Royston and Wright (2000).

`rqres.plot()` for plotting QQ-plots of different realisations of randomised residuals (for discrete distributions)

`wp()` worm plot of the residuals from a fitted `gamlss` object

`dtop()` detrended Own's plot of the residuals

Conclusions

- GAMLSS is a very flexible statistical model (but rather complex)
- It is a unified framework for univariate distributional regression models
- Allows any parametric 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

gamlss

Conclusions (continue)

- The fitted algorithm is modular, where different components can be added easily
- it can easily introduced to students since it relies on known concepts
- For continuous response variables deals properly with skewness and kurtosis (heavy tails)
- For discrete type response variables deals with overdispersion (or underdispersion), with excess (or lack) of zeros and with heavy tails

Conclusions (continue)

The model

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

Questions:

- What type of distribution is appropriate for the response variable?
- Which explanatory variable to select for each parameter?
- How those variable effect the distribution of the response?
- What link function shall we use?

gamlss

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

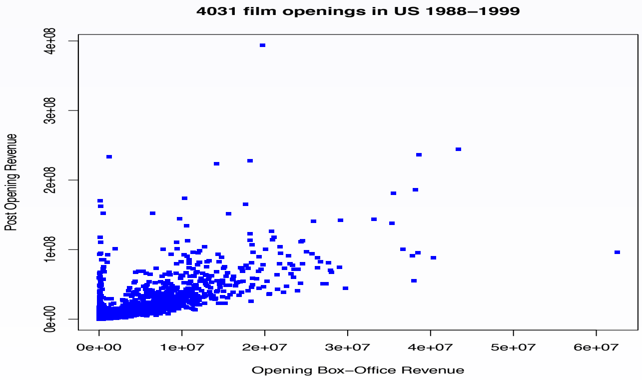
for more information see

www.gamlss.com

gamlss

Film Data

Go back



Film Data: the quiz

A small Quiz

Question 1 which film in the 90's has made the most money?

Question 2 which film in the 90's did badly in the first week but subsequently did very well?

Question 3 which film in the 90's did well in the first week but flopped after that?

Go back

gamlss

Film Data: Answers to the quiz

Go back