# Generalised Additive Models for Location Scale and Shape

## Past, Present and Future

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

Palermo, June 2018

Statistical Modelling Society

gamlss

# The statistical modelling philosophy

*Statistical modelling* is the art of using statistical reasoning to build a parsimonious models for a better understanding of the phenomena of interest.

- get data
- build a model
- interpretate/predict

gamlss

# The statistical modelling principals

- 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)
- 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
- *"no matter how beautiful your theory/model, no matter how clever you are or what your name is, if it disagrees with experiment"/data, "it's wrong"* (Richard Feynman)
- Test all the time your assumptions (there is no free meal)
- Try different models and choose the most appropriate for the data (have a data scientist attitude).

gamlss

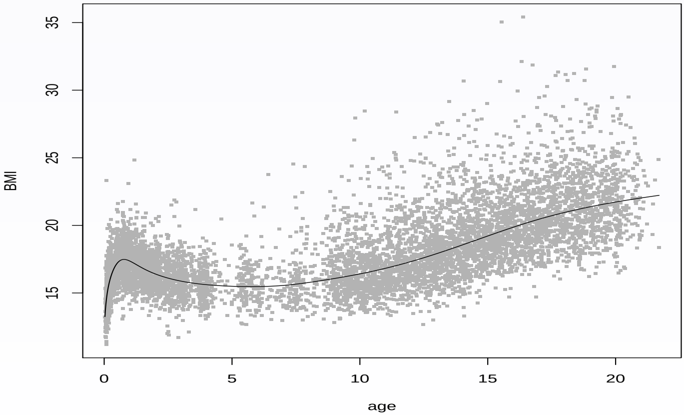
# The Dutch boys data

`BMI` : the BMI of 7294 boys

`age` : the age in years

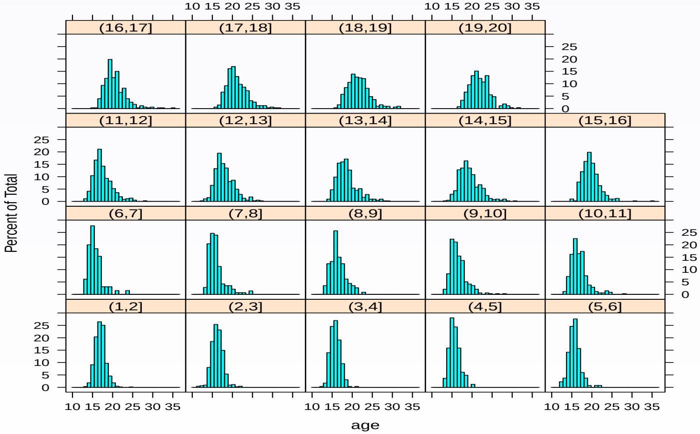
Source: van Buuren and Fredriks (2001)

# The Dutch boys data: statistical challenges

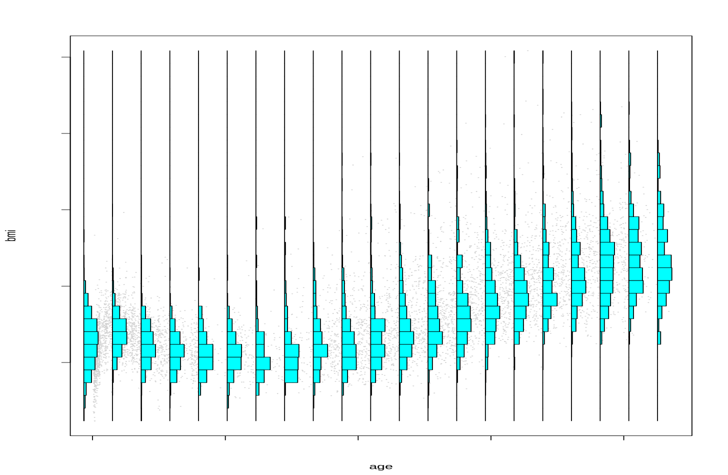



gamlss

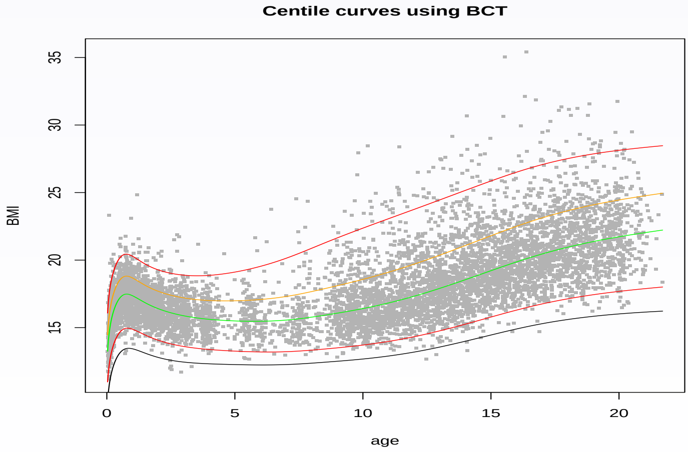
# The Dutch boys data: Histograms by age



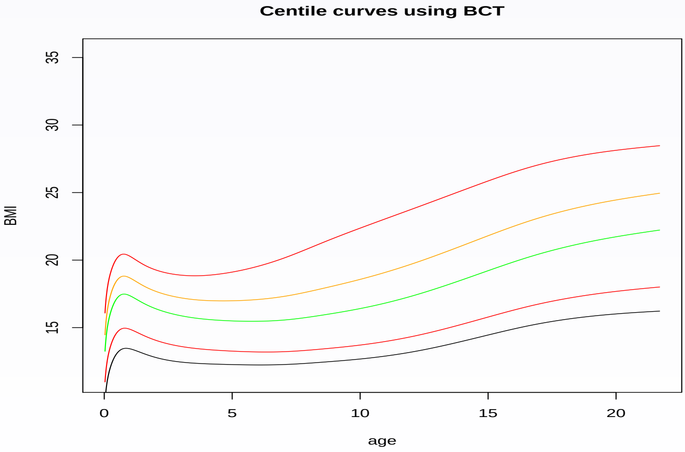
# The Dutch boys data: Conditional histograms by age



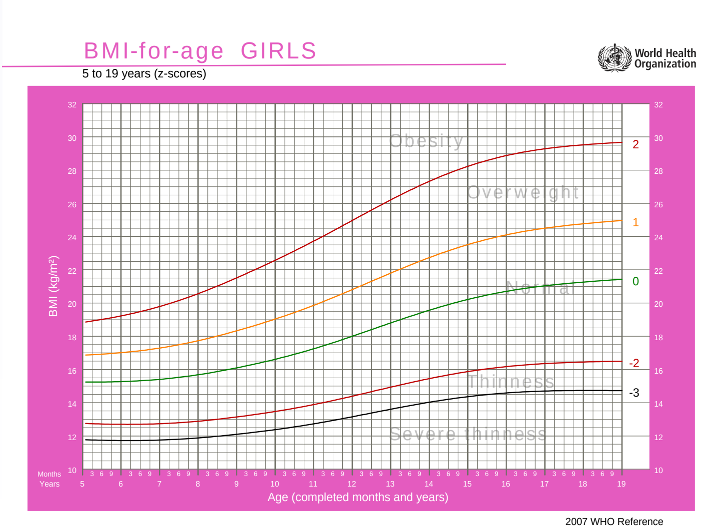
# The Dutch boys data: centile estimation



# The Dutch boys data: centiles

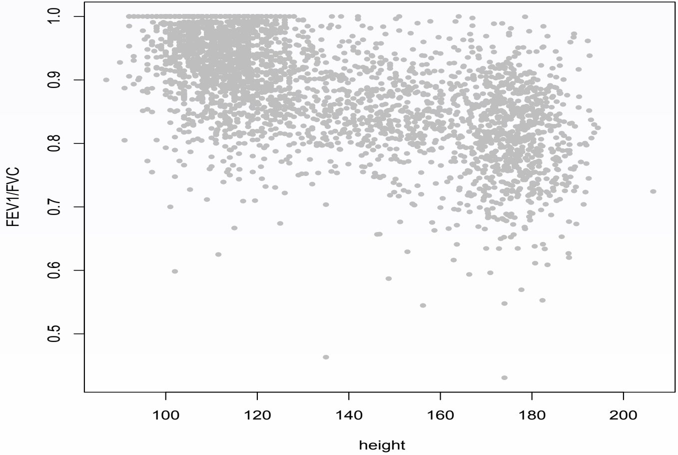


# World Health Organisation Child Growth Standards: Girls




gamlss

# 3164 male observations of lung function data




gamlss

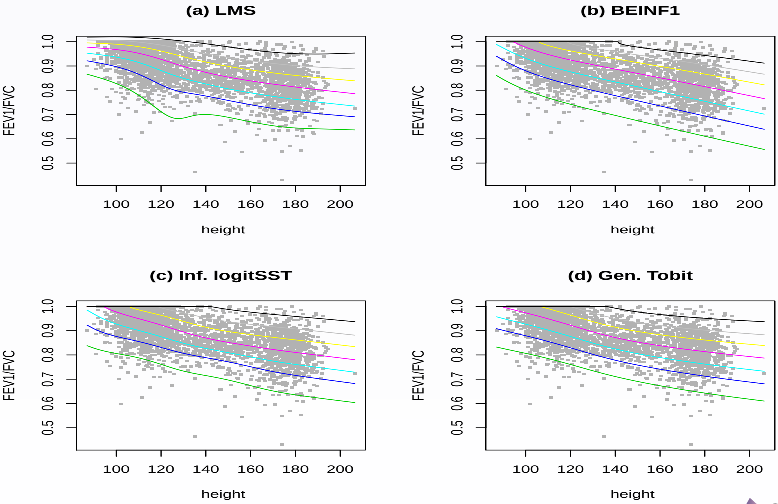
# The lung function data

$Y = FEV_1/FVC$ : the Spirometric lung function an established index for diagnosing airway obstruction (3164 male)

`height` : the height in cm

Source: Stanojevic et al. 2009

# The lung function data: fitted centile curves



# A stylometric application

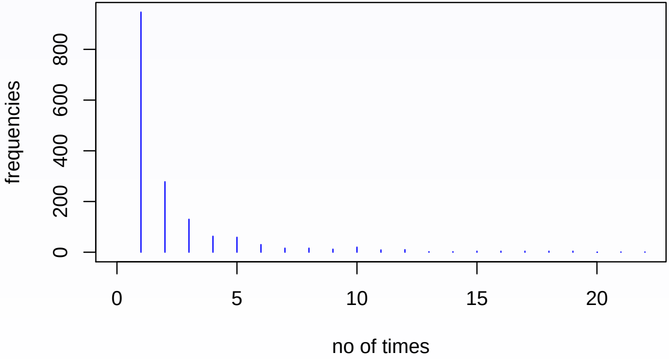
64 observations

`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

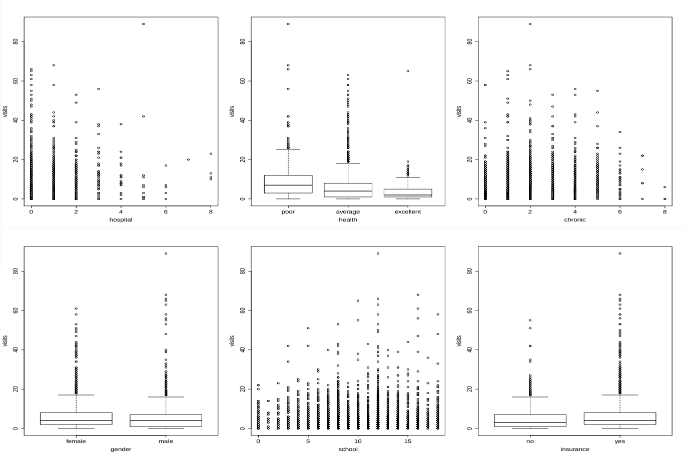


# The number of physician office visit

- `visits`: number of physician office visits,
- `hospital`: number of hospital stays,
- `health`: health status: a factor indicating whether self-perceived health is poor, average (reference category) or excellent,
- `chronic`: number of chronic conditions,
- `gender`: a factor indicating gender,
- `school`: number of years of education,
- `insurance`: a factor indicating whether the individual is covered by private insurance.

Data in AER package in R

# The number of physician office visit



# 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, excess of zeros and long tails 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

# Historical development

Important events in the creation of the GAMLSS models

- Linear model (Gauss, 1809) Go to LM
- 1972 Generalised Linear Models (Nelder and Wedderburn) Go to GLM
- 1990 Generalised Additive Models (Hastie and Tibshirani) Go to GAM
- 2005 Generalised Additive Models for Location Scale and Shape (GAMLSS) (Rigby and Stasinopoulos).

# Generalised Additive Model for Location Scale and Shape

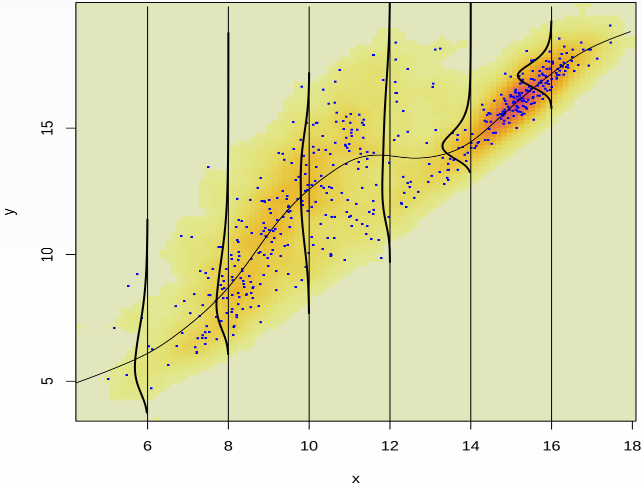
Generalised Additive Model for Location Scale and Shape Rigby and Stasinopoulos (2005)

$$\mathbf{y} \sim D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$$

$$\begin{aligned}
g_\mu(\boldsymbol{\mu}) &= \mathbf{X}_\mu \boldsymbol{\beta}_\mu + h_{1,\mu}(\mathbf{x}_{1,\mu}) + \ldots + h_{k,\mu}(\mathbf{x}_{k,\mu}) \\
g_\sigma(\boldsymbol{\sigma}) &= \mathbf{X}_\sigma \boldsymbol{\beta}_\sigma + h_{1,\sigma}(\mathbf{x}_{1,\sigma}) + \ldots + h_{k,\sigma}(\mathbf{x}_{k,\sigma}) \\
g_\nu(\boldsymbol{\nu}) &= \mathbf{X}_\nu \boldsymbol{\beta}_\nu + h_{1,\nu}(\mathbf{x}_{1,\nu}) + \ldots + h_{k,\nu}(\mathbf{x}_{k,\nu}) \\
g_\tau(\boldsymbol{\tau}) &= \mathbf{X}_\tau \boldsymbol{\beta}_\tau + h_{1,\tau}(\mathbf{x}_{1,\tau}) + \ldots + h_{k,\tau}(\mathbf{x}_{k,\tau})
\end{aligned}$$

where $D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$ can be any distribution and where $h_j(\mathbf{x}_j)$ are smooth functions of the $X$'s.

# GAMLSS assumptions



# 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.

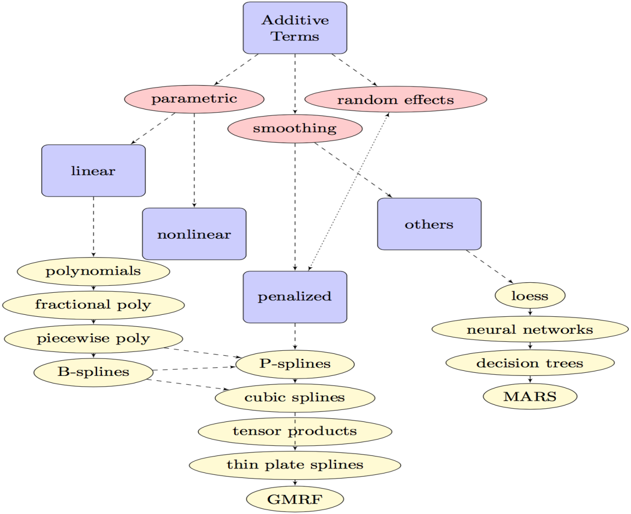
gamlss

# GAMLSS: Distributions

There are more than 100 explicit discrete discrete, continuous continuous, and mixed distributions, mixed, implemented as gamlss.family in the R including highly skew and kurtotic distributions Shapes,

- creating a new distribution is relatively easy
- truncating truncated an existing distribution
- using a censored version of an existing distribution
- mixing mixture different distributions to create a new finite mixture distribution.
- discretise discretise continuous distributions
- log or logit any continuous distribution in $(-\infty, \infty)$
- any distribution in $(0, \infty)$ can be zero adjusted to $[0, \infty)$
- any distribution in $(0, 1)$ can be inflated to $[0, 1]$

# Additive Terms



# GAMLSS: R implementation

GAMLSS is implemented in series of packages in R

| | |
|---|---|
| gamlss | the original package |
| gamlss.dist | for distributions |
| gamlss.data | for distributions |
| gamlss.demo | for demos |
| gamlss.nl | for non-linear terms |
| gamlss.tr | for truncated distributions |
| gamlss.cens | for censored (left, right or interval) response variables |
| gamlss.mx | for finite mixtures and random effects |
| gamlss.spatial | for Gaussian Markov Random Fields |
| gamlss.inf | for zero adjusted and inflated mixed distributions |

The GAMLSS packages can be downloaded from CRAN, the R library at http://www.r-project.org/

# GAMLSS components

Let $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \boldsymbol{\lambda}\}$ represent the GAMLSS model

- $\mathcal{D}$: distribution
- $\mathcal{G}$: the link function for distributional parameters
- $\mathcal{T}$: predictor terms for ($\boldsymbol{\eta}$' s) i.e. $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \sum_j h_j(\mathbf{x}_j)$
- $\boldsymbol{\lambda}$ : the hyper-parameters

# Problems, solutions and future research

- which distribution $\mathcal{D}$?
  - a book on distribution is prepared
  - a new function `chooseDist()` Go to chooseDist
  - robustify distributions
    - before fitting
    - after fitting
- which additive term for $\mu$, $\sigma$, $\nu$ and $\tau$?
  - all step-GAIC's are now parallel
  - possible connection of `ChooseDist()` and `stepGAIC()`
  - Machine learning techniques
    - GAMLSS boosting is well developed
    - connection to `glmnet`
- choosing the smoothing hyper parameters for terms
  - connection to **caret** package

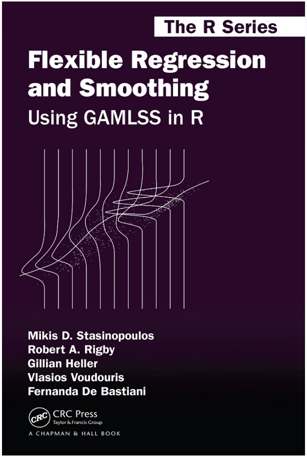
# Problems, solutions and future research

- selection between different (GAMLSS or not) models
  - GAIC and diagnostics exist but more work is needed to see where the benefits of using GAMLSS are coming from
  - influential observations
- Which inferential procedure?
  - penalised likelihood
  - Bayesian, see package **BAMLSS**
  - boosting, see package **gamboostLSS**
- Forcasting
  - developing time series modelling within GAMLSS
  - distributional forecast
  - automazation

# The Books

- Flexible Regression and Smoothing: Using GAMLSS in R (out in April 2017)
- Distributions for Location Scale and Shape: Using GAMLSS in R (expected in six to eight months)
- Generalized Additive Models for Location Scale and Shape: A Distributional Regression Approach. (starts in September)

# The 1st Book (out in April 2017)



# 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

gamlss

# This is a collaborative work

| co-authors | current collaborators |
|---|---|
| Vlasios Voudouris | Paul Eleirs |
| Gillian Heller | Marco Enea |
| Andreas Mayr | Daniil Kiose |
| Fernanda De Bastiani | Majid Djennad |
| Thomas Kneib | Luiz Nakamura |
| Nadja Klein | Abu Hossain |
| | past collaborators |
| | Popi Akantziliotou |
| | Fiona McElduff |
| | Raydonal Ospina |
| | Konstantinos Pateras |
| | Nicoleta Mortan |

# For more GAMLSS

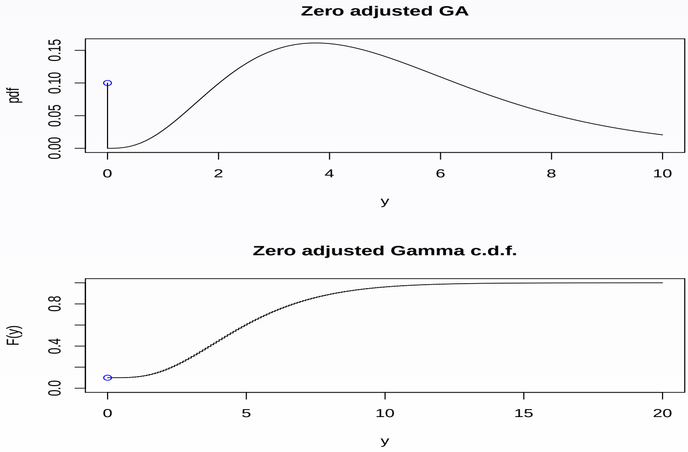
## the END

for more information see

www.gamlss.org

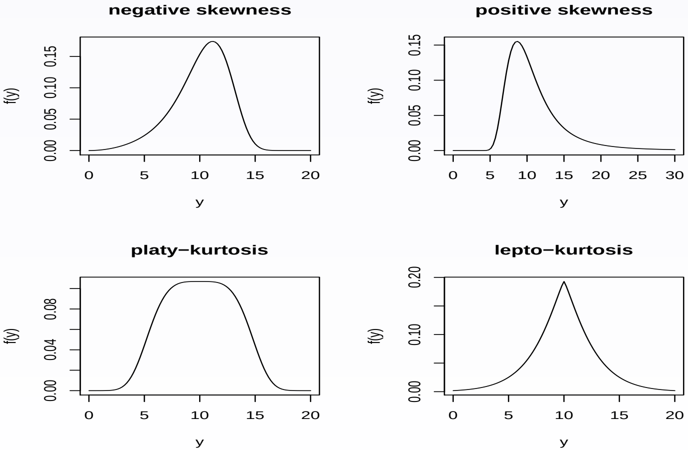
# Example of mixed distribution distributions

Go back Distributions



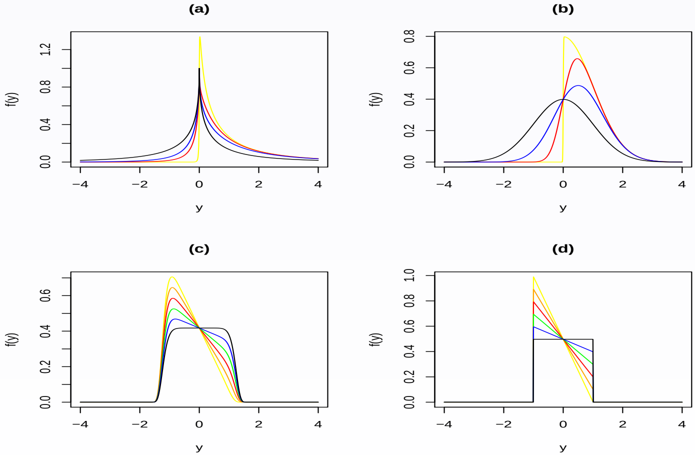
# Continuous distributions: different shapes

Go back Distributions



# Continuous distributions: different types

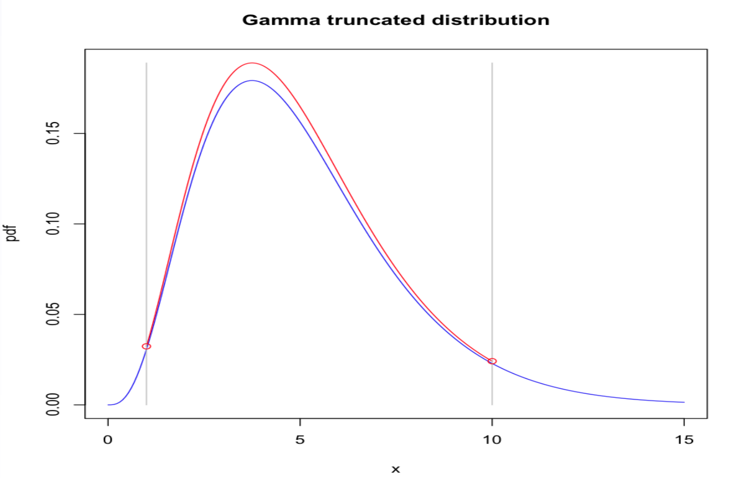
Go back Distributions




gamlss

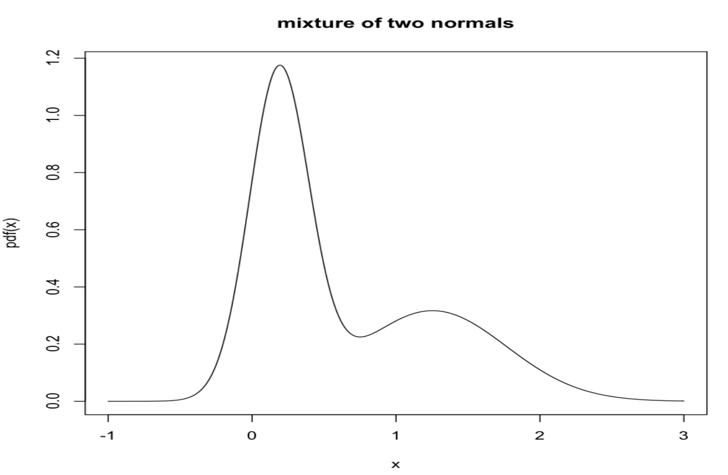
# Continuous distributions: different types

Go back Distributions



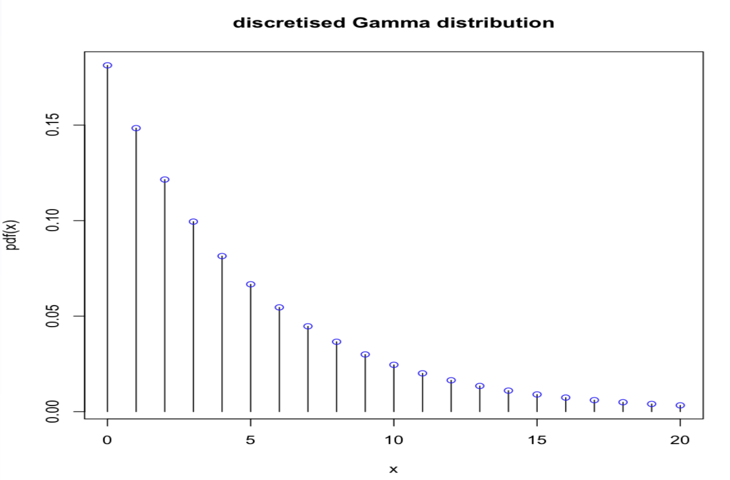
# Continuous distributions: different types

Go back Distributions



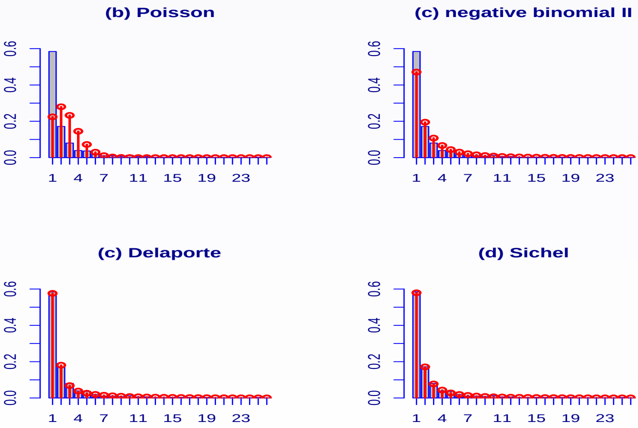
# Continuous distributions: different types

Go back Distributions




gamlss

# The stylometric data, Go back to distributions




gamlss

## Choose Distribution

Go back

```
mf <- chooseDist(m1, type="count")
```

| | 2 | 3.84 | 8.39 |
|---|---|---|---|
| PO | 35959.23 | 35973.95 | 36010.35 |
| GEOM | 24402.77 | 24417.49 | 24453.89 |
| GEOMo | 24402.77 | 24417.49 | 24453.89 |
| ... | ... | ... | ... |
| ZASICHEL | 24469.07 | 24502.19 | 24584.09 |
| ZINBF.1 | 24207.21 | 24240.33 | 24322.23 |
| ZIBNB | 24107.94 | 24141.06 | 24222.96 |
| ZISICHEL | 24196.25 | 24229.37 | 24311.27 |

```
getOrder(mf)
```

# The linear model

Linear Model, Gauss

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim NO(\mathbf{0}, \sigma^2\mathbf{I})$

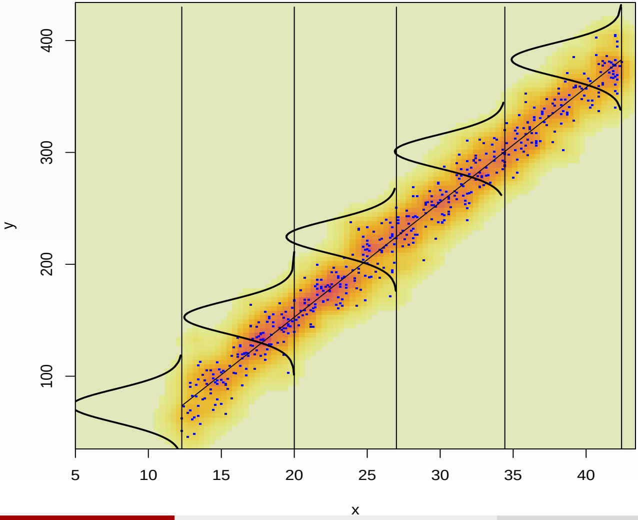
The model can be also written as:

$\mathbf{y} \sim NO(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$

Go back 70-80

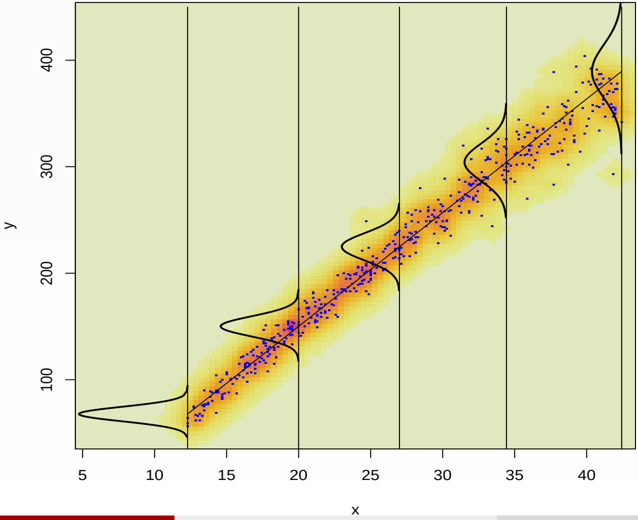
# The linear model assumptions

Go back 70-80 Next page



# The weighted linear model assumptions

Go back 70-80 Next page



# The linear model: comments

- estimation is achieved by Least Squares or Weighted Least Squares (WLS)
- the normal distribution is important for inference
- we only modelling the mean as linear function of the explanatory variables
- One of the top ten reasons to become statistician (according to Friedman, Friedman & Amoo, 2002, Journal of Statistics Education):

"Statisticians are mean lovers".

Go back 70-80

# The generalised linear model

Generalised Linear Model, Nelder and Wedderburn(1972)

$g(\boldsymbol{\mu}) = \mathbf{X}\beta$ where $\mathbf{y} \sim \mathit{ExpFamily}(\boldsymbol{\mu}, \phi)$

where $g()$ is the link function

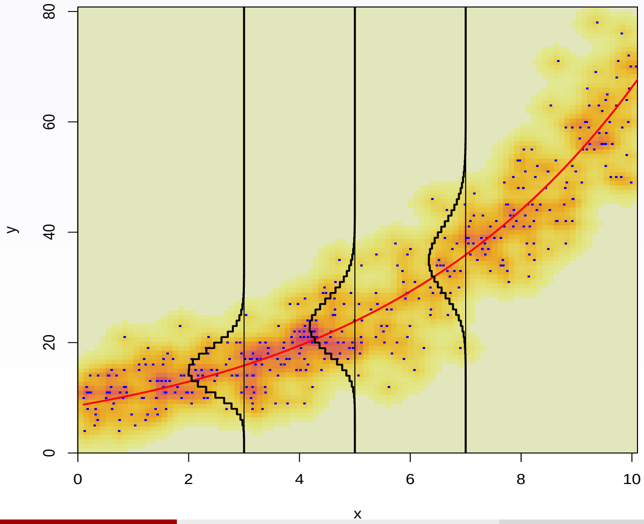
The exponential family

1. normal
2. Gamma
3. inverse Gaussian
4. Poisson
5. binomial

Go back 70-80 Next page

# The generalised linear model

Go back 70-80 Next page




gamlss

# The generalised linear model

- estimation is achieved by Iterative Re-weighted Least Squares (IRLS)
- we can model discrete response variables
- we are still "mean lovers".

Go back 70-80

# The generalised additive model

Generalised additive model Hastie and Tibshirani (1990)

$\mathbf{y} \sim ExpFamily(\boldsymbol{\mu}, \phi)$

$g(\boldsymbol{\mu}) = \mathbf{X}\beta + h_1(\mathbf{x}_1) + ... + h_k(\mathbf{x}_k)$

where $h_j(\mathbf{x}_j)$ are smooth functions of the $X$'s.

Go back Historical Next page