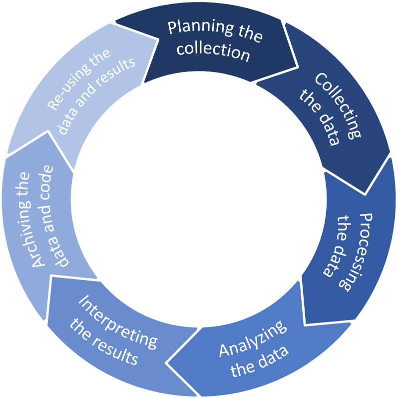
Visual tools for distributional regression models (GAMLSS)

Mikis Stasinopoulos and Bob Rigby

London Metropolitan University

Annual Meeting of the Brazilian Region International Biometric Society, Florianopolis, Nov 2022

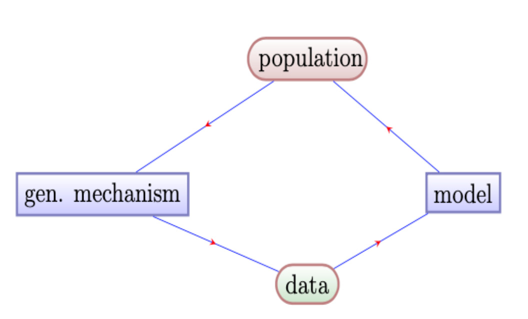
The Data Cycle.



Statistical Models: Basic Principles

1. "all models are wrong but some are useful"
2. the model should be fit for purpose
3. most data analytic models have a structural-mathematical part and a stochastic-probabilistic part
4. "all models are based on assumptions " [there is no free meal]
5. given the the assumptions are adequate the model can be useful

Statistical Inference



Model Cycle

1. assume a model
2. fit the model ('training the model')
3. check the model ('model diagnostics')
4. compare different models ('model selection')
5. interpret the model

Regression Models

$$X \quad \rightarrow \quad Y$$

explanatory terms $\rightarrow$ *response* variable(s)

supervised learning

regression models $\rightarrow$ Y *continuous or discrete*

classification models $\rightarrow$ Y *categorical*

we are trying to explain the behaviour of the response variable Y as a function of the explanatory X.
There one direction in the relationship no dynamic or reinforce modelling

Regression Models

The most common regression models are dealing with

$$E(Y|\mathbf{X}) = g(\mathbf{X})$$

The average value of Y depends on the values of the features $\mathbf{X}$ according to an unknown function $g()$.
More generally:

$$Ch(Y|\mathbf{X}) = g(\mathbf{X})$$

Some characteristic of the distribution of Y depends on the values of the features $\mathbf{X}$ according to an unknown mechanism $g()$

Distributional Regression Models

Let us assume a stochastic (probability) model where Y and X are random variables:

$$\underbrace{f(Y, X)}_{\text{joint}} = \underbrace{f(Y|X)}_{\text{conditional}} \underbrace{f(X)}_{\text{marginal}}$$

gamlss

Distributional Regression Models

- In a distributional regression model we assume that the response is coming from a theoretical probability (density) function i.e

$$\mathbf{y}|X \overset{\text{ind}}{\sim} f(\boldsymbol{\theta}(X))$$

- where $\boldsymbol{\theta}$ is a vector of distribution parameters which can depend on the explanatory variables

The following statistical model can be thought as distributional models:

- linear models (LM), under the assumption of normality,
- generalised linear models (GLM) and
- generalised additive model (GAM)
- Regression machine learning techniques i.e. regression trees e.t.c.

but they model only the mean parameter of the distribution

Distributional Regression: GAMLSS

- when the whole of the conditional distribution $f(Y|X)$ is of concern distributional regression, GAMLSS, models are appropriate
- GAMLSS allowing modelling all the feature of the conditional distribution
  - location
  - scale
  - skewness
  - kurtosis (long tails)
- most of the well known statistical models are modelling only the mean (or location).
- distributional regression models are beyond mean models
- they do need a strong assumption about $f(Y|X)$ but it is easy to check the assumptions

GAMLSS: Rigby and Stasinopoulos (2005)

The original GAMLSS model

$$
\begin{aligned}
\mathbf{y} &\overset{\text{ind}}{\sim} \mathcal{D}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) \\
g_1\,(\boldsymbol{\theta}_1) &= \mathbf{X}_1\boldsymbol{\beta}_1 + s_{11}(\mathbf{x}_{11}) + \ldots + s_{1J_1}(\mathbf{x}_{1J_1}) \\
g_2\,(\boldsymbol{\theta}_2) &= \mathbf{X}_2\boldsymbol{\beta}_2 + s_{21}(\mathbf{x}_{21}) + \ldots + s_{2J_2}(\mathbf{x}_{2J_2}) \\
g_3\,(\ldots) &= \ldots \\
g_k\,(\boldsymbol{\theta}_k) &= \mathbf{X}_k\boldsymbol{\beta}_k + s_{k1}(\mathbf{x}_{k1}) + \ldots + s_{kJ_k}(\mathbf{x}_{kJ_k})
\end{aligned}
$$

The R Implementation: Stasinopoulos and Rigby (2008)

The R- implementation GAMLSS model: additive

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

GAMLSS: Stasinopoulos at al (2017)

The model do not have to be additive.

$$
\begin{aligned}
\mathbf{y} &\stackrel{\text{ind}}{\sim} \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\
g_1(\boldsymbol{\mu}) &= ML(\mathbf{X}_1) \\
g_2(\boldsymbol{\sigma}) &= ML(\mathbf{X}_2) \\
g_3(\boldsymbol{\nu}) &= ML(\mathbf{X}_3) \\
g_4(\boldsymbol{\tau}) &= ML(\mathbf{X}_4)
\end{aligned}
$$

ML: Machine Learning techniques i.e. Regression Trees, Neural Networks, LASSO, Ridge Regression or Principal Component Regression (PCR) [Stasinopoulos at al (2020)].

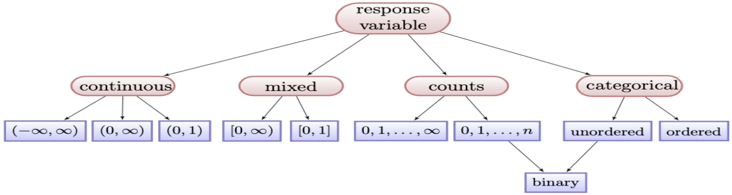
What is GAMLSS?

GAMLSS : are distributional based semi-parametric regression type models.

- regression type: since $\mathbf{X} \rightarrow \mathcal{D}(\mathbf{y})$
- distributional: a full parametric distribution assumption $\mathcal{D}(\mathbf{y})$ is made for the response variable
- semi-parametric: All the parameters of the distribution can be modelled as additive function of the explanatory variables i.e. $\mathbf{X}_k \boldsymbol{\beta}_k + s_{k1}(\mathbf{x}_{k1}) + \ldots + s_{kJ_1}(\mathbf{x}_{kJ_1})$
- GAMLSS philosophy: try different distributions, try different additive functions

LM, GLM and GAM models are special cases of GAMLSS

The Response (Target)



The *range* of values the **response** variable, is crucial, in developing a good model.

GAMLSS: Distributions

There are more than 100 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 continuous distribution in $(0, \infty)$ or $(0, 1)$ can be inflated at 0 or/and 1

gamlss

The Features

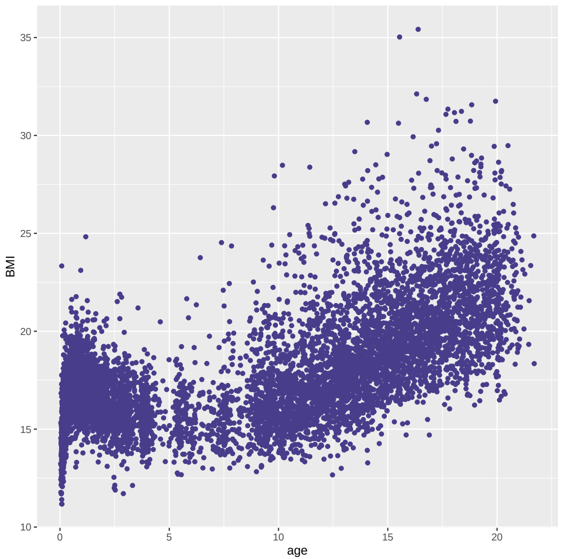
- The **explanatory** terms, $(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots, \mathbf{x_{p-1}}, \mathbf{x_p})$, can be:

  Continuous covariates: taking a variety of different values i.e. age or `height`.

  factors: Factors are categorical variables which can be item[unordered, i.e. `gender` where the *levels* of the factor, ‘male’, ’female’ do not have a specific order or ordered i.e. `disease_level` where the levels ”severe”, ”mild’, ”light’ do have a specific order.

gamlss

The Dutch boys data: 7294 observations

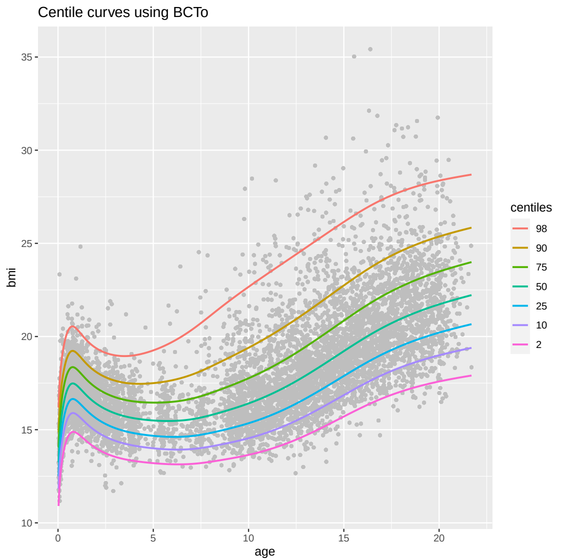


The Dutch Boys Data

> `BMI` : the BMI of 7294 boys
>
> `age` : the age in years

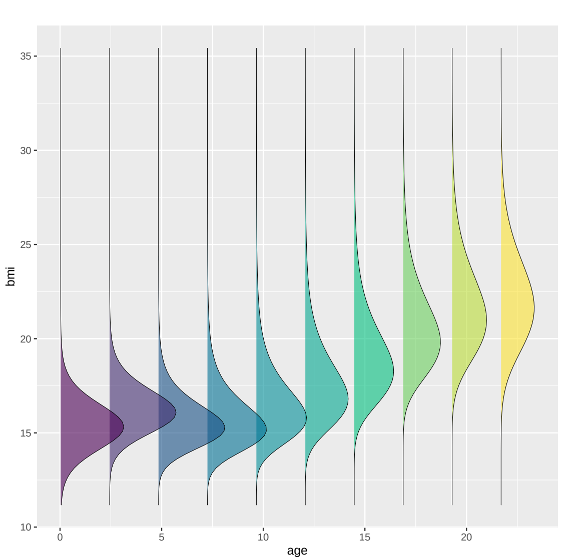
Source: van Buuren and Fredriks (2001) with the purpose of constructing growth (centile) curves

The Dutch Boys Data: Centile Curves

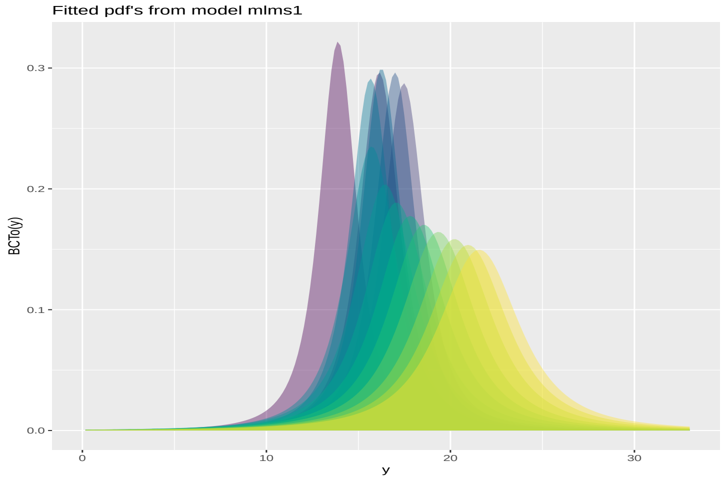


gamlss

The Dutch Boys Data: Fitted Distributions



How the distribution of BMI changes with age



gamlss

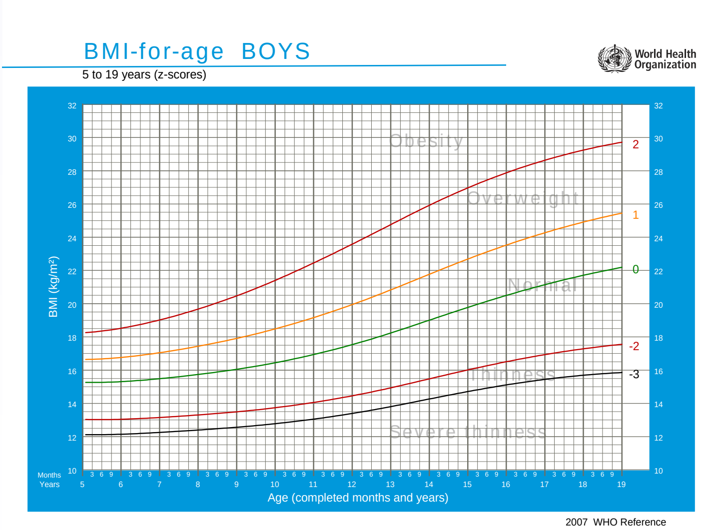
GAMLSS and the BCT distribution

The model

$$
\begin{aligned}
bmi &\overset{\text{ind}}{\sim} BCTo(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\
\log(\boldsymbol{\mu}) &= s_{\mu}(age^{p}) \\
\log(\boldsymbol{\sigma}) &= s_{\sigma}(age^{p}) \\
\boldsymbol{\nu} &= s_{\nu}(age^{p}) \\
\log(\boldsymbol{\tau}) &= s_{\tau}(age^{p})
\end{aligned}
$$

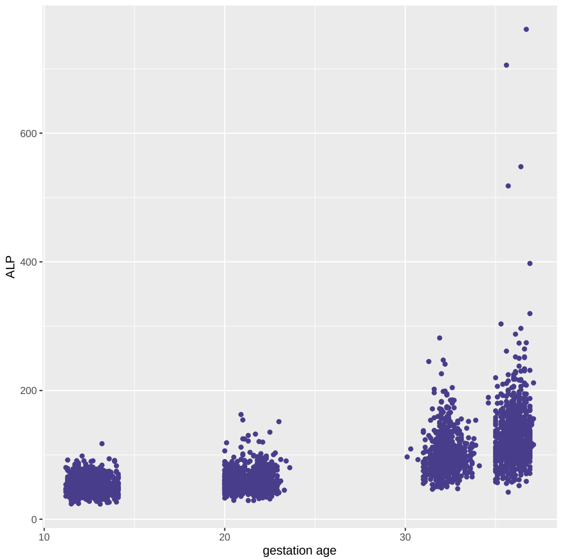
$p = .33$

World Health Organisation Child Growth Standards: Boys

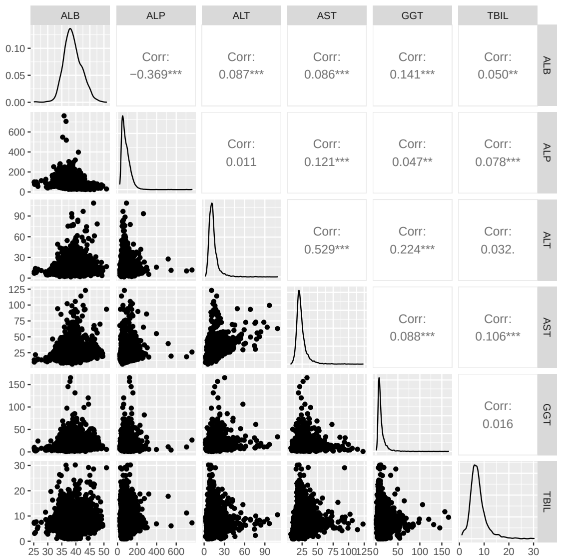


gamlss

The King's College Hospital Data: 3429 observations



The King's College Hospital Data; 6 responses (liver)



gamlss

The King's College Hospital Data; explanatory variables

`ga` : gestation age

`age` : the age of the mother

`race` : the race

`ht` : the height of the mother

`wt` : the weight of the mother

`conception` : the type of conception

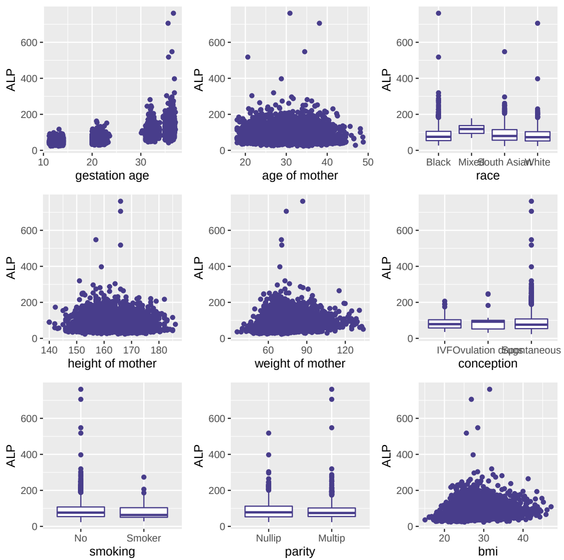
`smoking` : whether smoking

`parity` : the number of children

`bmi` : the BMI of the mother

Source: Maternal-Fetal Medicine, King's College

The King's College Hospital Data: ALT response

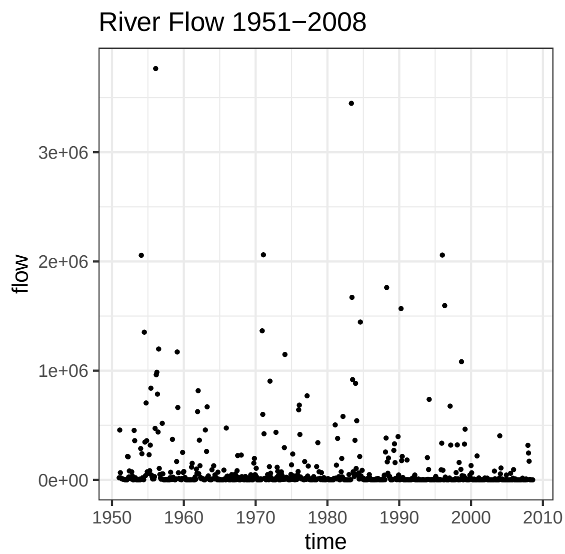


gamlss

The King's College Hospital Data: the final model

$$
\begin{aligned}
ALP &\stackrel{\text{ind}}{\sim} BCTo(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\
\log \hat{\boldsymbol{\mu}} &= 5.37 - 0.0601 \text{ ga} + 0.0025 \text{ ga}^2 + 0.0214 \text{ bmi} + \\
&\quad 0.007 \text{ ga*bmi} - 0.000033 \text{ ga}^2 * \text{bmi} - \\
&\quad 0.126 \text{ age} + 0.0035 \text{ age}^2 - 0.000033\text{age}^3 - \\
&\quad 0.0037 \text{ wt} + 0.06469 \text{ race.SA} \\
\log \hat{\boldsymbol{\sigma}} &= -1.6503 + 0.008814 \text{ bmi} \\
\hat{\boldsymbol{\nu}} &= -0.2742018 \\
\log \hat{\boldsymbol{\tau}} &= 6.9594177 - 0.1359 \text{ ga}
\end{aligned}
$$

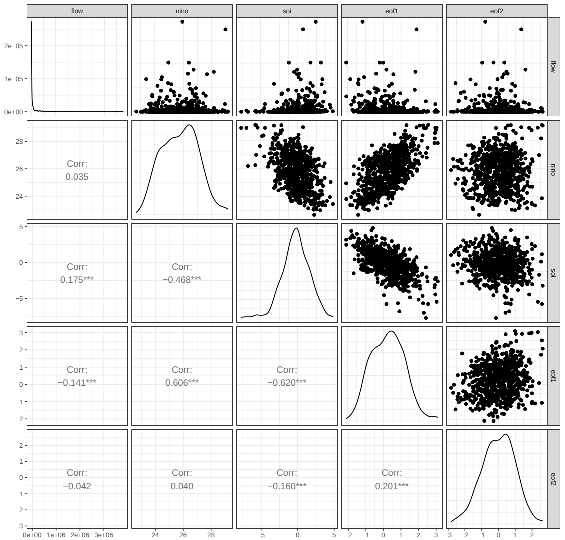
River Flow Data



River Flow Data

- the response contains zeros
- the positive river flow values are very skew to the right
- this is typical to a lot of enviromental data
- you need a zero-adjusted (mixed) distribution for the response
- the data are time series
- the interest lies in the left tail (drought?) and the right tail (flooding?)

River Flow Data; explanatory variables



River Flow Data; explanatory variables

- `nino` : average sea surface temperatures in the region
- `soi` : the southern oscillation index monthly air pressure difference between Darwin and Tahiti
- `eof1` : first empirical orthogonal functions of sea surface temperature
- `eof2` : second empirical orthogonal functions of sea surface temperature
- `t` : time
- `m` : month
- `F1-5` : log-lags for river flow
- `P1-5` : lags whether it rains

A total of 12.4 of the response have zero values
Source: Van Ogtrop et al (2011)

River Flow Data; The final model

$$
\begin{aligned}
\textbf{flow} &\overset{\text{ind}}{\sim} \text{zAdjBCTo}(\boldsymbol{\pi}_0, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\
\log\left(\frac{\hat{\boldsymbol{\pi}}_0}{1-\hat{\boldsymbol{\pi}}_0}\right) &= -11.021 - 0.402\ F_{t-1} + 0.108\ F_{t-2} + s(t) - 0.269\ soi \\
\log \hat{\boldsymbol{\mu}} &= 57.38 + 0.52\ F_1 + s_c(m) + 0.25\ soi - 0.026\ t + \\
&\quad 5.085\ P_1 \\
\log \hat{\boldsymbol{\sigma}} &= -9.81 - 0.116\ F_1 + s_c(m) - 1.008\ P_1 + \\
&\quad 0.29\ P_2 + 0.005\ t \\
\hat{\boldsymbol{\nu}} &= s_c(m) - 0.017\ F_1 \\
\log \hat{\boldsymbol{\tau}} &= 2.504
\end{aligned}
$$

Visual Tools: Introduction

- Diagnostics and model interpretation tools are a posteriori model tools that are used after the fitting of a model or models.
- Diagnostic tools are checking the assumptions of the model.
- Interpretation tools try to interpret the fitted model using different scenarios
- Note: "Models are based on assumptions, the stronger the assumptions, more diagnostic tools usually exist to check those assumptions".

gamlss

Visual Model Diagnostic Tools

- Visual model diagnostic tools are trying to answer two questions;
- is the distribution for the response adequate for the data?
- how the model will change if we slightly change certain aspect of the data set?
  - Those aspect usually are parts of the rows or the columns of the data matrix .
  - Diagnostics tools can be used within a model to check specific properties of the model or between models for model comparison.

gamlss

Visual Model Diagnostic Tools

Checking the distribution using residuals

- is the distribution adequate for the data?
  - RF-residual plots
  - RF-worm-plot
  - RF-multiple worm-plot
  - RF-model worm-plot
  - RF-detrended Own's plot
- are the distribution parameters θ_k fitted properly? worm plot
- is skewness and kurtosis still present?
  Use the bucket plot De Bastiani *et. al* (2022)
  - RF-moment bucket plots
  - RF-centile bucket plots
  - RF-model bucket plots

gamlss

Visual Model Diagnostic Tools

Checking the structural part of the model

- are the explanatory terms fitted properly?
  - do we need linear or non-linear smoothing terms?
  - do we need interactions?
- are observations in the response unusual? KC-outliers KC-deviance increnments
- are observations in explanatory variables unusual? KC-leverage
- are observations influencial in the fit? (need more work)
- is the model robust?

gamlss

Visual Model Interpretation Tools

- Interpretation tools are used for trying to understand why and how the model works.
- Transparent box provides an explanation on how $\mathbf{X}$ effects Y
- Black box models are difficult to interpret but mostly the interest lies on whether are good for prediction
- GAMLSS are transparent as far the interpretaion of parameters is concern but not straight forward because it is not always clear what part of the distribution is affected and by which explanatory terms and how.
- The main tool for interpretation are partial effects

Partial Effect

let

- $\mathbf{x_j}$ the term of interest
- $\mathbf{x_{-j}}$ all other terms
- $\omega(\mathcal{D})$ a characteristic of the assumed distribution i.e. 90% quantile
- S a scenario and a function $g()$ how to treat the extra terms $\mathbf{x_{-j}}$
- Them the partial effect that the term $\mathbf{x_j}$ has on the characteristic of the distribution $\omega(\mathcal{D})$, given the specific values of the rest of the terms, say $g(\mathbf{x_{-j}})$, under the scenario S, is denoted as:

$$\mathrm{PE}_{\omega(\mathcal{D})}\left(\mathbf{x}_j | S\left[g(\mathbf{x_{-j}})\right]\right)$$

Partial Effect; different $\omega(\mathcal{D})$

- the predictor η_{θ_k}: term.plot(), fitted_terms() (not-appropriate for interactions) KC-η_μ
- the predictor η_{θ_k}: pe_param(..,type="eta") KC-intaraction-η_μ, KC-intaraction-η_μ
- the parameter θ_k: pe_param(..,type="param") KC-μ, KC-μ interaction
- the partial mean and variance of Y; pe_moment()
- the partial quantile of Y; pe_quantile() KC-gestation age, KC-race, KC-all
- the partial distribution of Y pe_pdf() KC-continuous, KC-factor

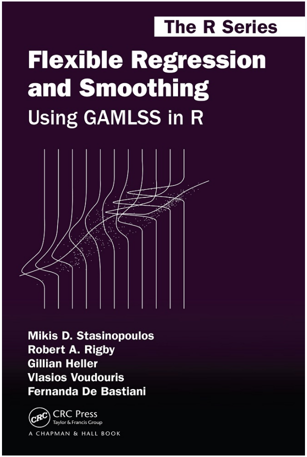
Partial Effect; different scenarios $S\left[g(\mathbf{x_{-j}})\right]$

- fixed $\mathbf{x_{-j}}$
  - at predetermine values
  - at the mean or median of $\mathbf{x_{-j}}$
- average $\mathbf{x_{-j}}$
  - over all $\mathbf{x_{-j}}$ values, leading to Partial Dependence Plots (PDF) of Friedman (2001)
  - over values in the neighbour of $\mathbf{x_{-j}}$, leading to Marginal-plots (M-plots)
  - over the derivative values of $\mathbf{x_{-j}}$, leading to Accumulated Local Effects, (ALE) plots, Apley (2020).

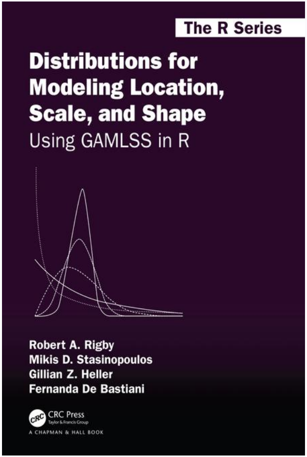
Conclusions

- GAMLSS is a very flexible statistical model (but a rather complex one)
- 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 additive terms in the models for the distribution parameters
- Allow a variety of statistical and graphical tools for checking the assumptions and intepreting the model

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
Distributional Regression, with Applications

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

Users: academic

GAMLSS can model highly skewed and kurtotic data which are common in environmental and financial data. GAMLSS is used:

- general modelling: "linguistics", "agro-industrial waste", "spam", "real estate", "energy efficiency ratings", "wall insulation" etc.
- environment modelling: "fire", "flood frequency", "annual flood peaks", "extreme rainfalls", "extreme precipitation", "extreme minimum temperatures", "satellite telemetry in wildlife research", "water extreme events", "flood frequency , "drought hazard"
- economy: "commercial fisheries ", "book markets", "claim counts" , " unpaid claims reserves", "stress testing", e.t.c.
- centile estimation and growth curves: is the most popular use of GAMLSS

Users: non-academic

- WHO Child Growth Standards which have been adopted by more than 150 countries,
- Global Lung Function Initiative to produce standards for lung capacity (based on age, height, gender and ethnicity)
- International Monetary Fund IMF
- Bank of England, BE
- Standard Chartered Bank, Bank of America
- Great Barrier Reef Marine Park Authority for modelling the health and condition of seagrass in the Great Barrier Reef.

gamlss

This is a collaborative work: A list of people who help with GAMLSS and the R impementation

active	past
Gillian Heller	Popi Akantziliotou
Fernanda De Bastiani	Vlasios Voudouris
Julian Merder	Fiona McElduff
Nikos Kametas	Raydonal Ospina
Luiz Nakamura	Abu Hossain
Andreas Mayr	Daniil Kiose
Thomas Kneib	Nicoleta Mortan
Nadja Klein	Konstantinos Pateras
Tim Cole	Majid Djennad
Achim Zaileis	Nikos Georgikopoulos
	Paul Eilers
	María Xosé Rodríguez-Álvarezi
	Dea-Jin Lee

More to be done

the END

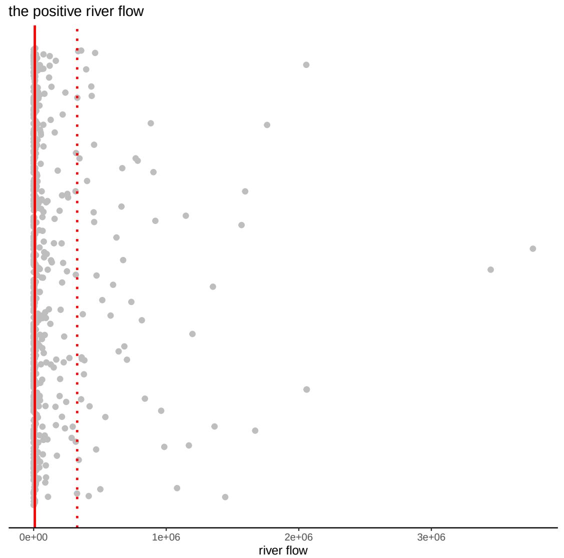
for more information see

www.gamlss.com

gamlss

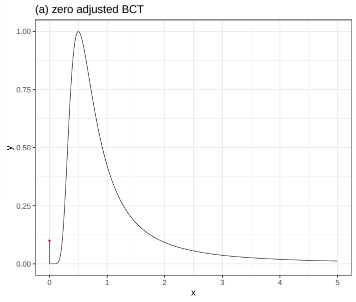
River flow data

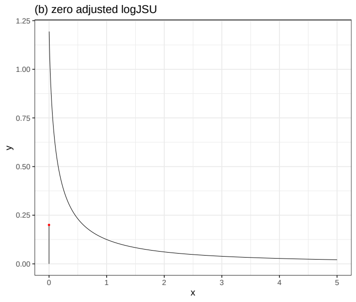
Go back



gamlss

Mixed distributions Go back





The Data

Go back

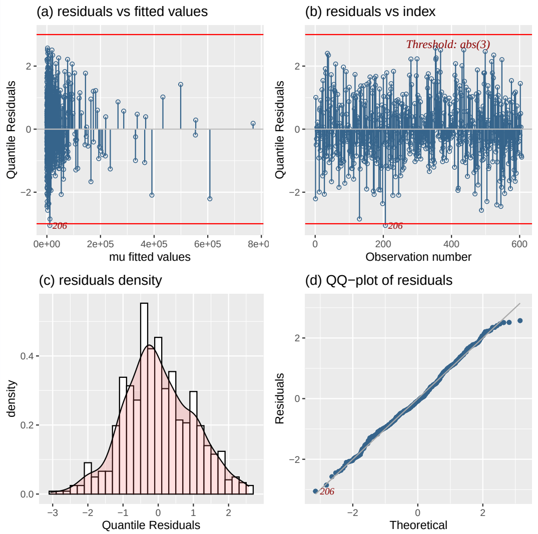
Table: Table showing typical data sets for regression models

obs.	$\mathbf{y}$	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	...	$\mathbf{x_{p-1}}$	$\mathbf{x_p}$
1	y_1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	...	$x_{1,p-1}$	$x_{1,p}$
2.	y_2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	...	$x_{2,p-1}$	$x_{2,p}$
3	y_3	$x_{3,3}$	$x_{3,2}$	$x_{3,3}$.	...	$x_{3,p-1}$	$x_{3,p}$
...	...	...	...	...	...	...	...
n-2	y_{n-2}	$x_{n-2,1}$.	$x_{n-2,2}$	$x_{n-2,3}$	...	$x_{n-2,p-1}$	$x_{n-2,p}$
n-1	y_{n-2}	$x_{n-1,1}$	$x_{n-2,2}$	$x_{n-1,3}$	...	$x_{n-1,p-1}$	$x_{n-1,p}$
n	y_n	$x_{n,1}$.	$x_{n,2}$.	$x_{n,3}$	...	$x_{n,p-1}$	$x_{n,p}$

gamlss

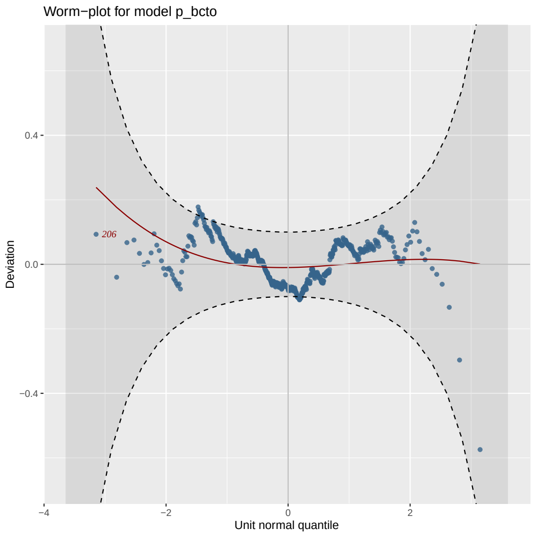
Residuals plots (river flow)

Go back



gamlss

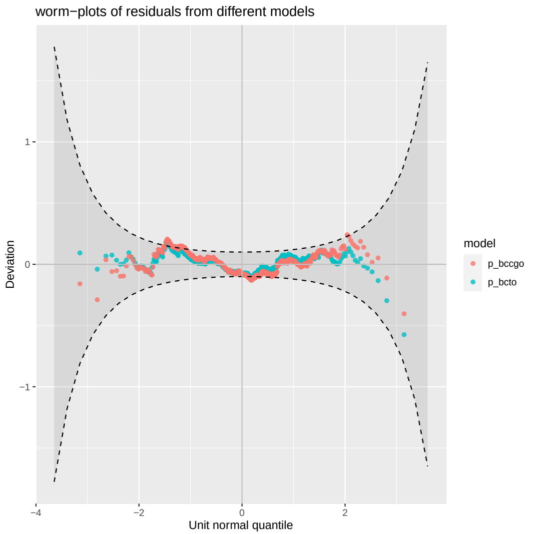
Worm plot (river flow) Go back



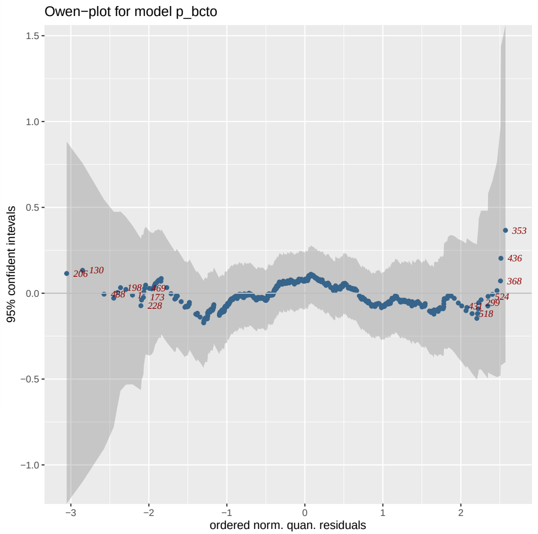
gamlss

Model worm plot (river flow)

Go back

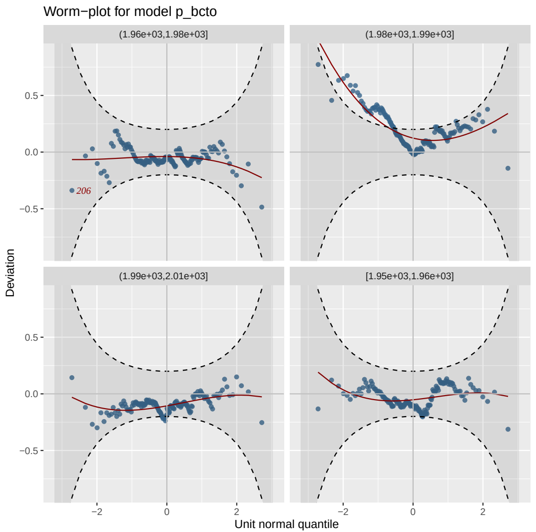


De-trended Own;s plots (river flow) Go back



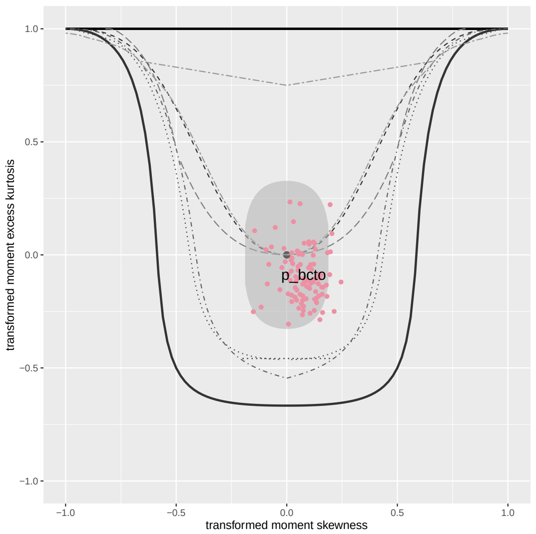
Multiple worm plots (river flow)

Go back



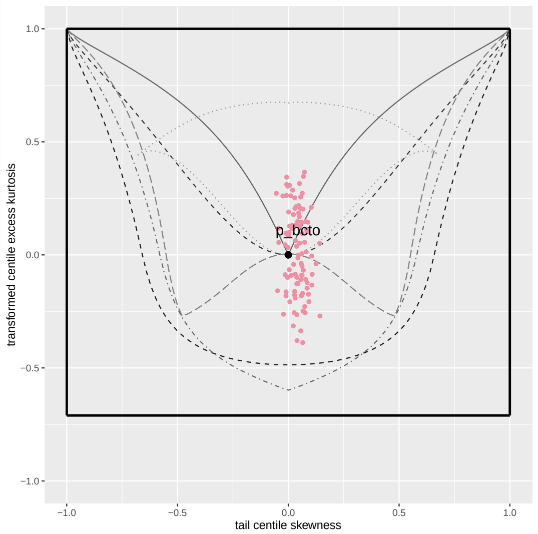
Bucket plot (river flow)

Go back



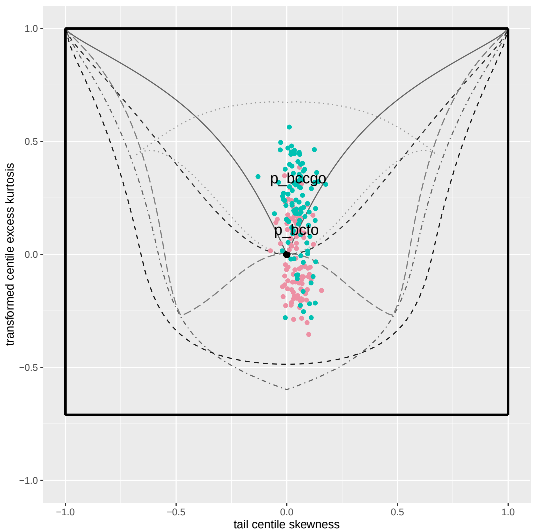
Centile bucket plot (river flow)

Go back



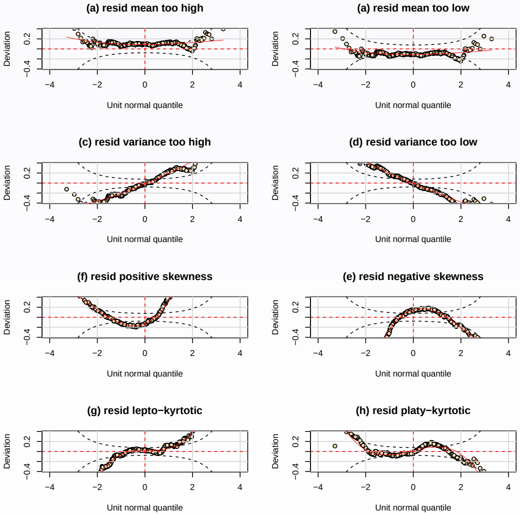
Model bucket plot (river flow)

Go back



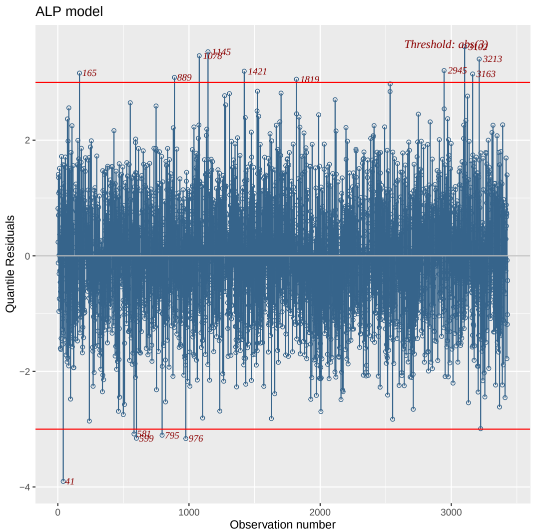
Worm plot for diagnostics

Go back



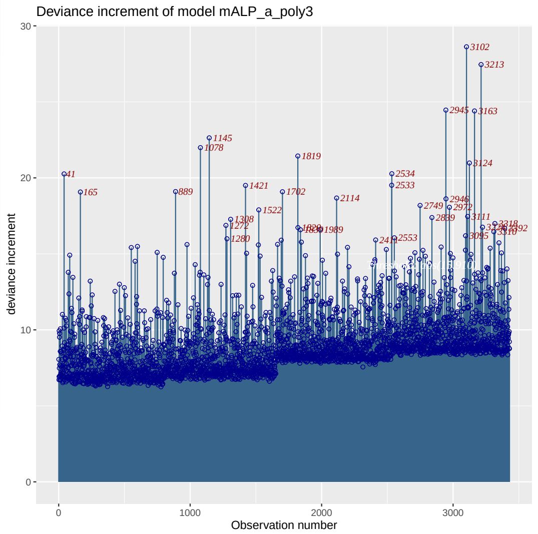
Outliers (King's college data)

Go back



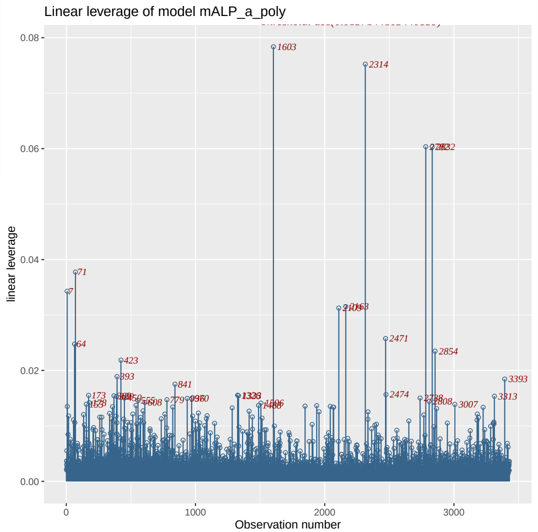
Deviance increments (King's college data)

Go back

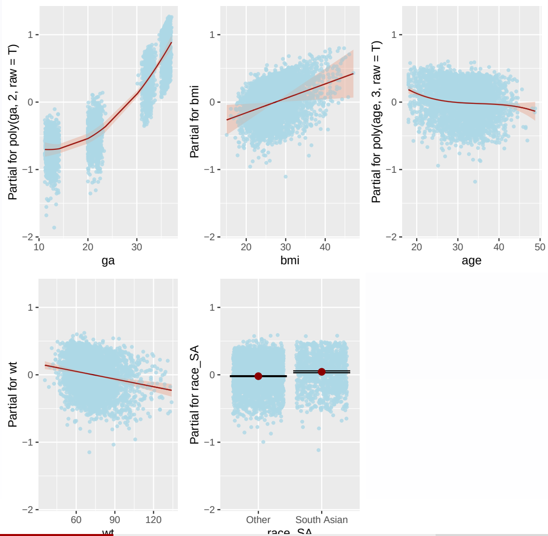


Leverage (King's college data)

Go back

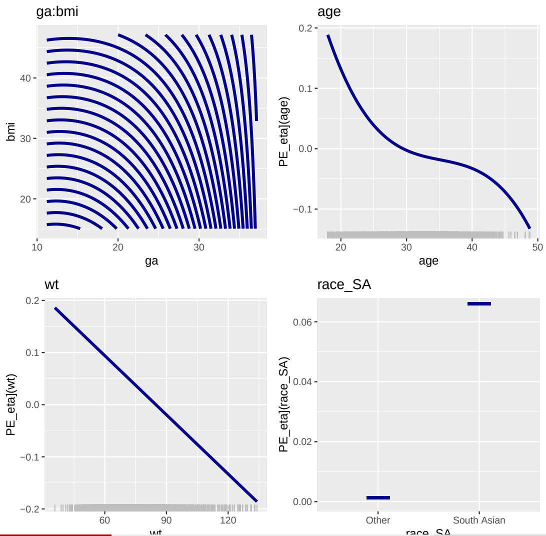


Partial plots for predictor for μ using `fitted_terms()` (King's college data) Go back



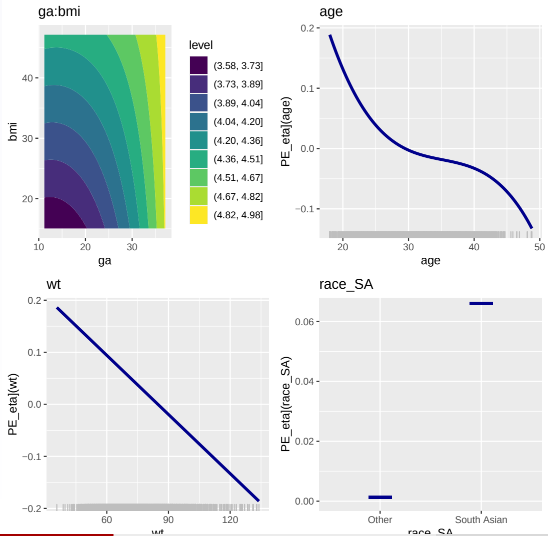
gamlss

Partial plots for predictor for μ using `pe_param(.., type="eta")`(King's college data) Go back



gamlss

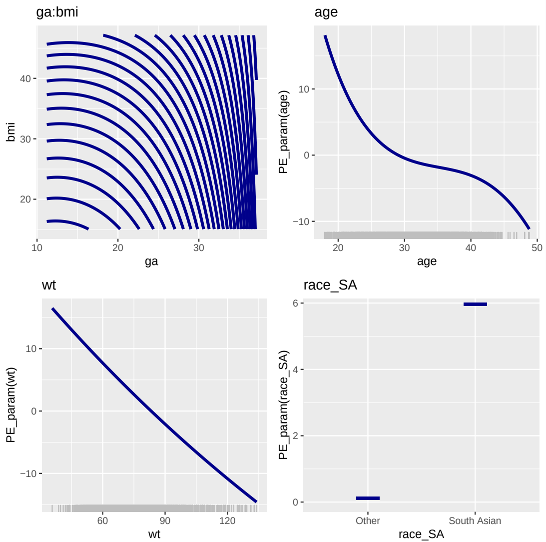
Partial plots for predictor for μ using `pe_param(.., type="eta", filled=TRUE)` (King's college data) Go back



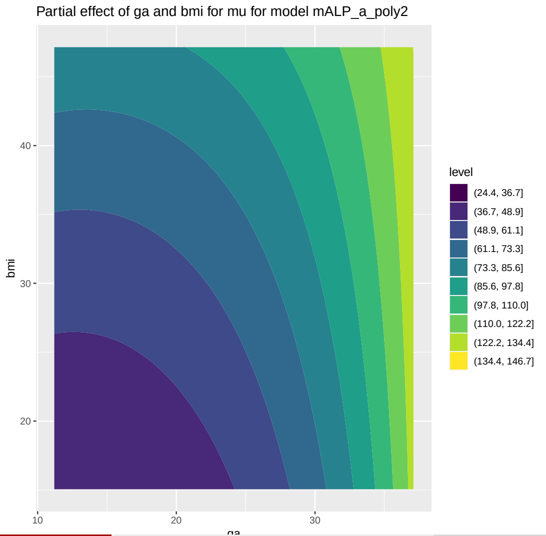
gamlss

Partial plots for μ using `pe_param()` (King's college data)

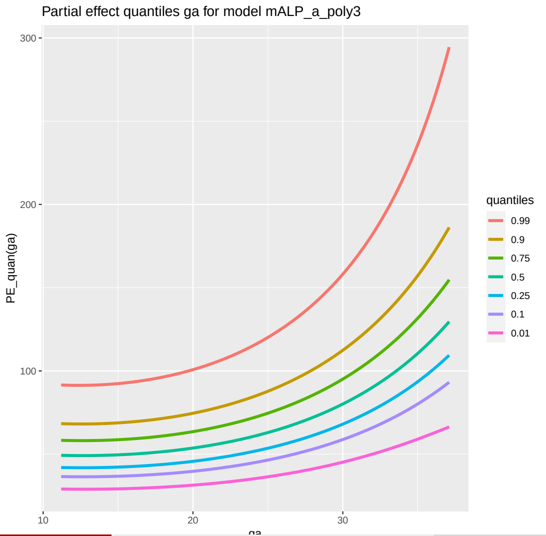
Go back



Partial interaction plot for μ using `pe_param()` (King's college data) Go back

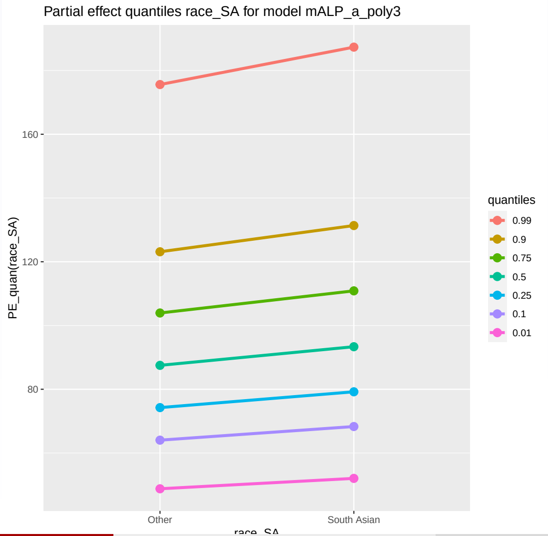


Partial quantile plots for ga using `pe_quantile()` (King's college data) Go back



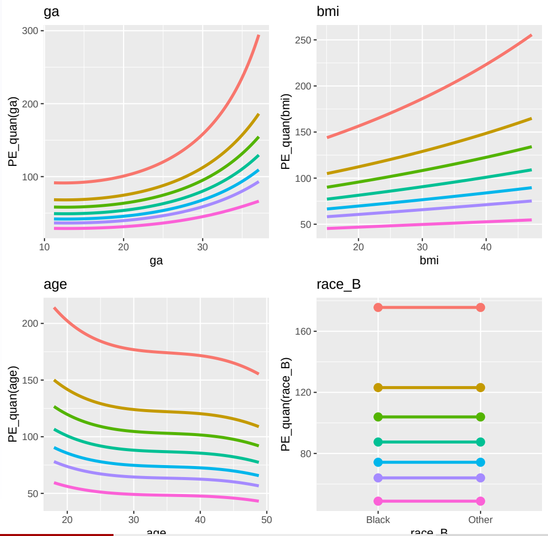
gamlss

Partial quantile for `race_SA` using `pe_quantile()` (King's college data) Go back

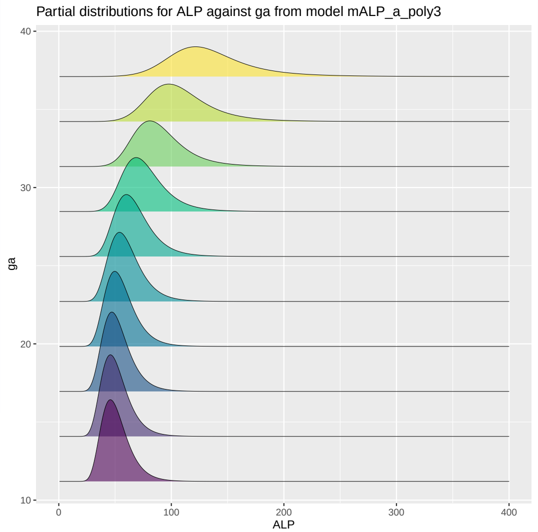


Partial quantile for all variables using `pe_quantile_grid()` (King's college data)

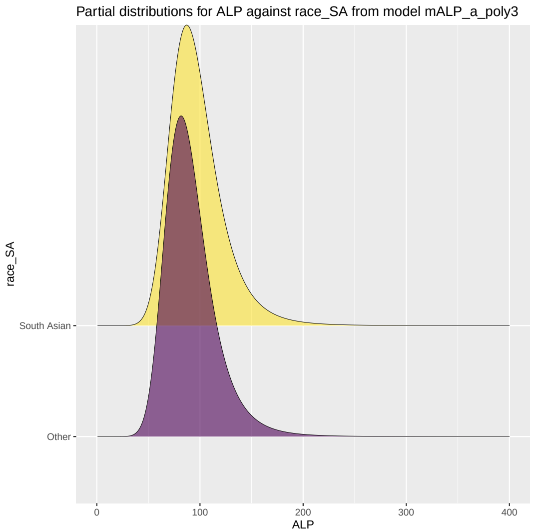
Go back



PDF for variable ga (King's College Data) Go back



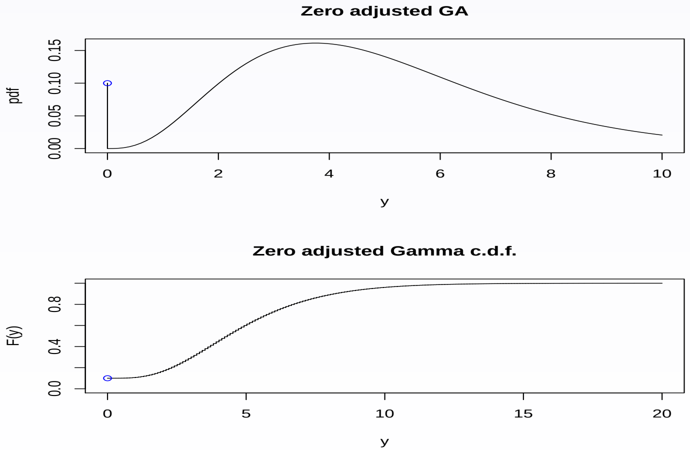
PDF for variable race_SA (King's College Data) Go back



gamlss

Example of mixed distribution distributions

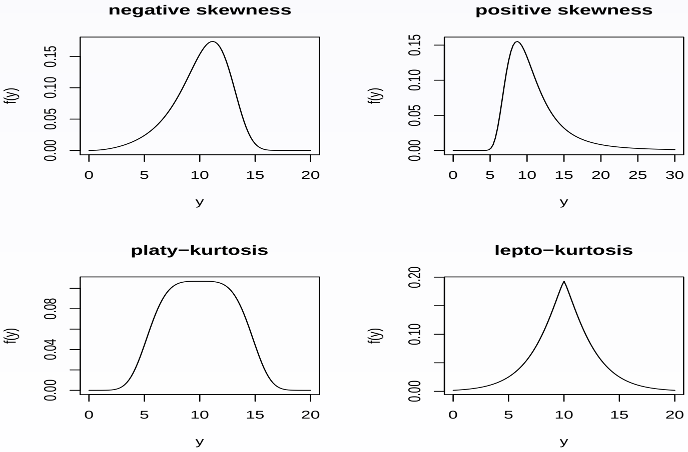
Go back Distributions



gamlss

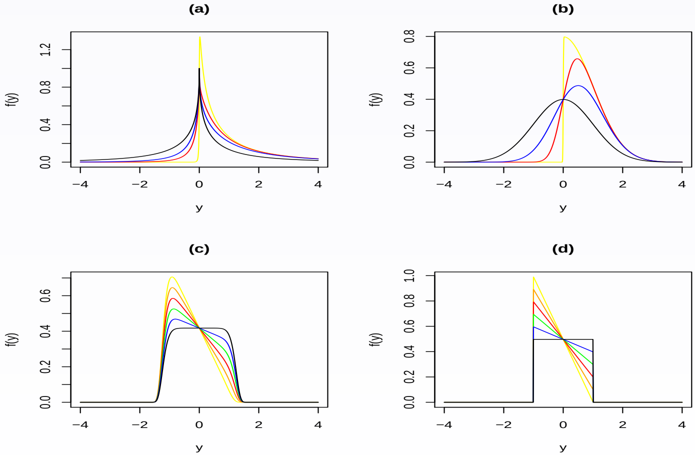
Continuous distributions: different shapes

Go back Distributions



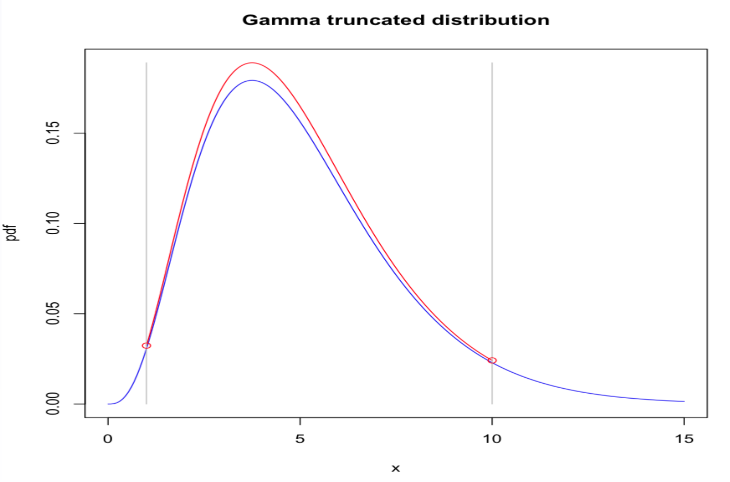
Continuous distributions: different types

Go back Distributions



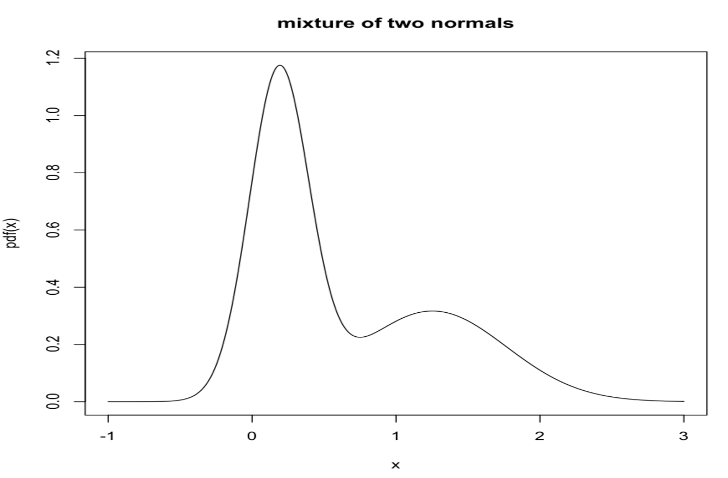
Continuous distributions: different types

Go back Distributions



Continuous distributions: different types

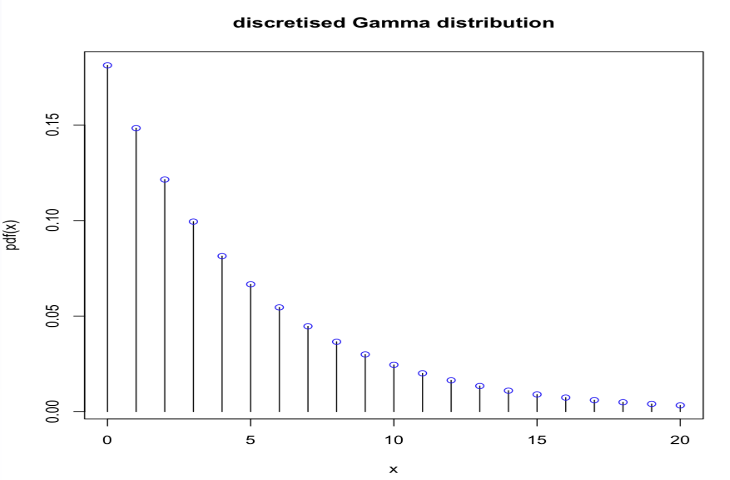
Go back Distributions



gamlss

Continuous distributions: different types

Go back Distributions



The stylometric data, Go back to distributions