

Flexible Regression and Smoothing

Centile Estimation

Mikis Stasinopoulos Bob Rigby

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

- 1 The Dutch boys data
- 2 The problem
- 3 The methods
- 4 The LMS model and its extensions
- 5 GAMLSS functions for centile estimation
- 6 Conclusions

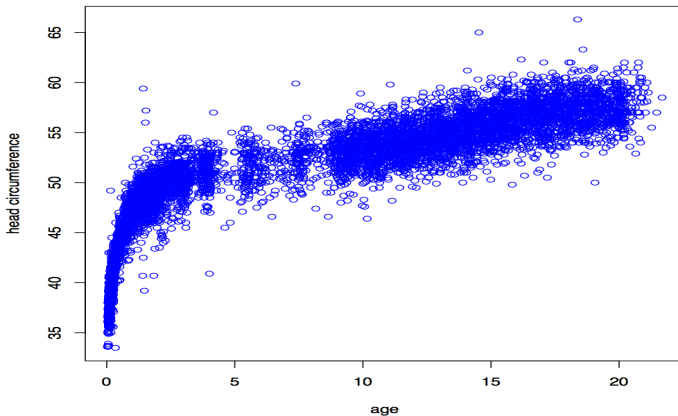
The Dutch boys data

`head` : the head head circumference of 7040 boys

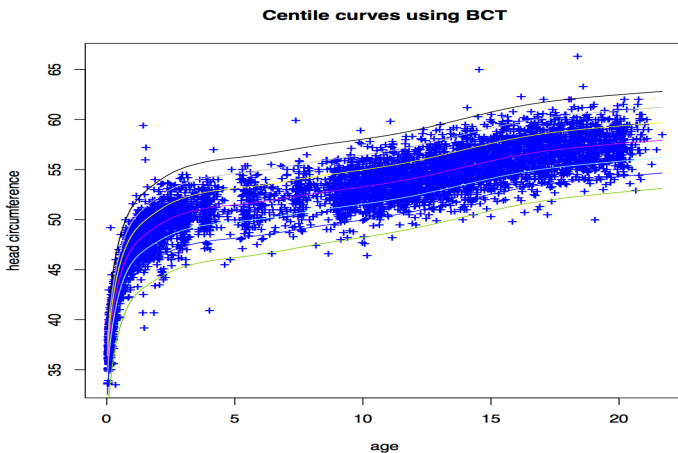
`age` : the age in years

Source: Buuren and Fredriks (2001)

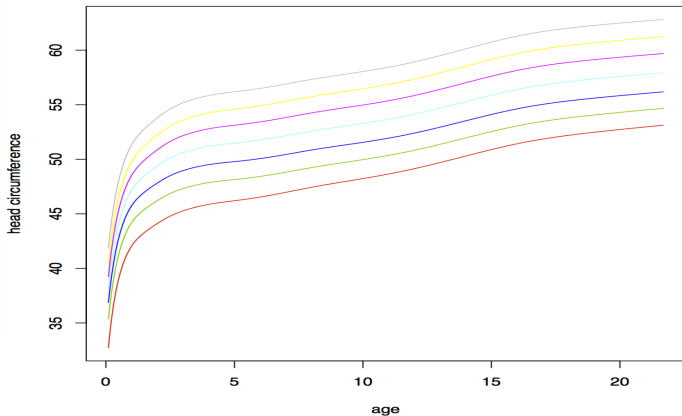
The Dutch boys data



The Dutch boys data with centiles



The Dutch boys data: centiles



The methods

- i) the non parametric approach of quantile regression (Koenker, 2005; Koenker and Bassett, 1978)
- ii) the parametric LMS approach of Cole (1988), Cole and Green (1992) and its extensions Rigby and Stasinopoulos (2004, 2006, 2007).

Centiles: the LMS method

$Y \sim f_Y(y|\mu, \sigma, \nu, \tau)$ where $f_Y()$ is any distribution

Y = head circumference and $X = AGE^\xi$

$$\mu = s(x, df_\mu)$$

$$\log(\sigma) = s(x, df_\sigma)$$

$$\nu = s(x, df_\nu)$$

$$\log(\tau) = s(x, df_\tau)$$

The LMS method and extensions

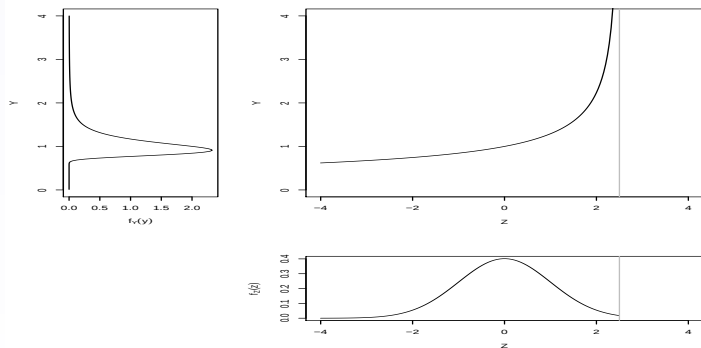
Let Y be a random variable with range $Y > 0$ defined through the transformed variable Z given by:

$$Z = \frac{1}{\sigma\nu} \left[\left(\frac{Y}{\mu} \right)^\nu - 1 \right], \quad \text{if } \nu \neq 0$$

$$= \frac{1}{\sigma} \log \left(\frac{Y}{\mu} \right), \quad \text{if } \nu = 0.$$

- ① if $Z \sim N(0, 1)$ then $Y \sim BCCG(\mu, \sigma, \nu) = \text{LMS method}$
- ② if $Z \sim t_\tau$ then $Y \sim BCT(\mu, \sigma, \nu, \tau) = \text{LMST method}$
- ③ if $Z \sim PE(0, 1, \tau)$ then $Y \sim BCPE(\mu, \sigma, \nu, \tau) = \text{LMSP method}$
adopted by WHO

The BC-CG transformation



Centiles: estimation of the smoothing parameters

We need to select the five values $df_\mu, df_\sigma, df_\nu, df_\tau, \xi$

- by trial and error
- minimize the generalized Akaike information criterion, $GAIC(\#)$
- minimize the the validation global deviance VGD
- using local selection criteria, i.e. CV, ML

Diagnostics **should** be used in all above cases

GAMLSS functions for centile estimation

`lms()` automatic selection of an appropriate LMS method

`centiles()` to plot centile curves against an x-variable.

`calibration()` adjust for sample quantiles to create centiles

`centiles.com()` to compare centiles curves for more than one object.

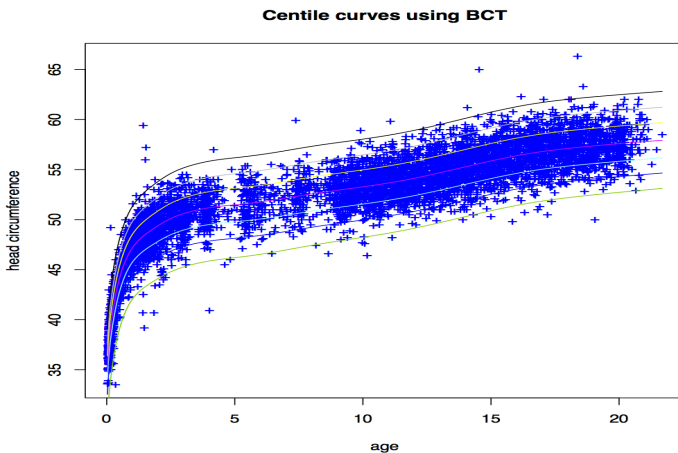
`centiles.split()` as for `centiles()`, but splits the plot at specified values of x.

`centiles.fan()` fan plot as in `centiles()`.

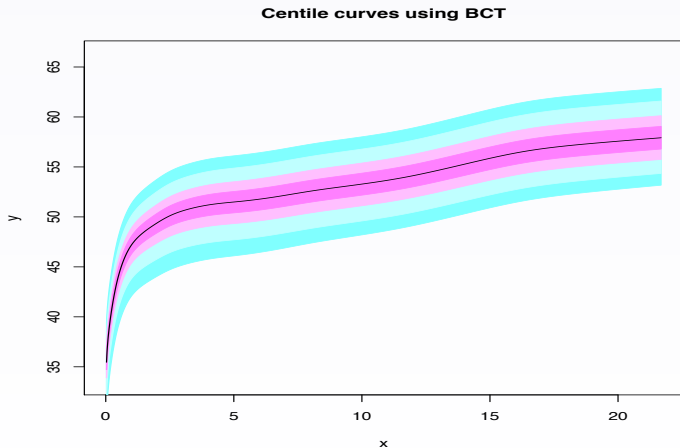
`centiles.pred()` to predict and plot centile curves for new x-values.

`fitted.plot()` to plot fitted values for all the parameters against an x-variable

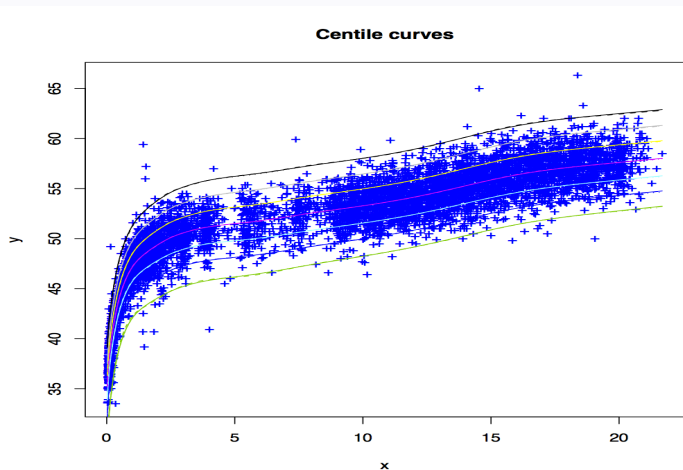
The head circumference data: centiles()



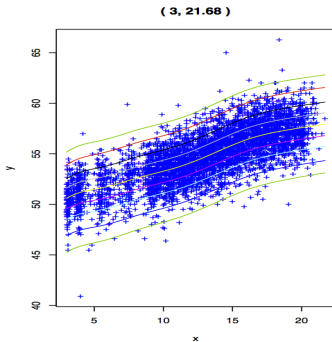
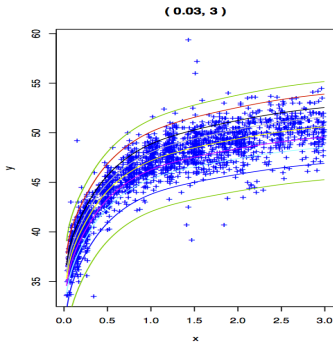
The head circumference data: `centiles.fan()`



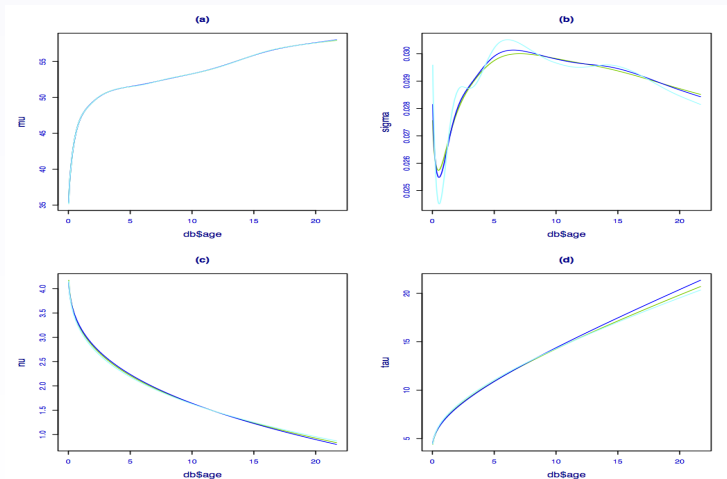
The head circumference data: comparison of the centiles, centiles.com()



The head circumference data: `centiles.split()`



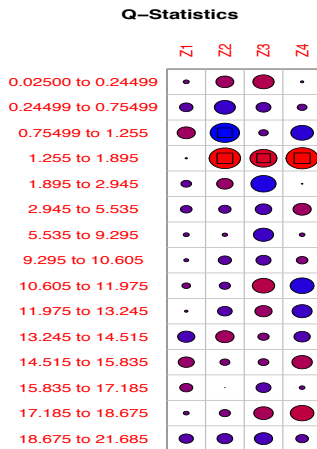
The head circumference data: `fitted.plot()`



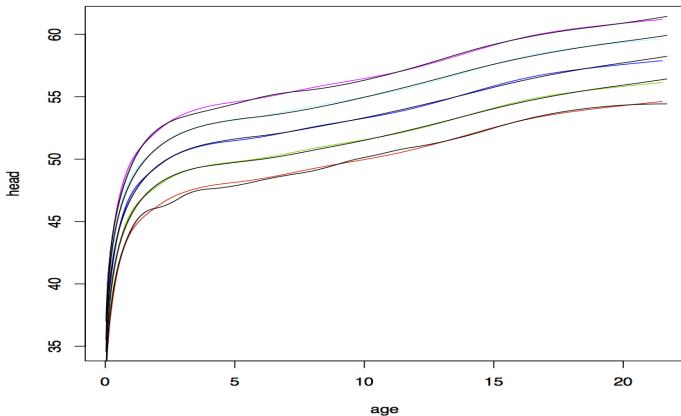
The head circumference data: The Q statistics

	Z1	Z2	Z3	Z4	Agostino	N
0.02500 to 0.24499	0.09	0.85	1.21	0.02	1.47	477
0.24499 to 0.75499	-0.48	-1.18	-0.58	-0.25	0.40	473
0.75499 to 1.255	0.88	-2.27	-0.24	-1.37	1.94	467
1.255 to 1.895	0.01	2.62	1.96	2.76	11.47	460
1.895 to 2.945	-0.29	0.72	-1.70	-0.01	2.90	473
2.945 to 5.535	-0.35	-0.40	-0.72	0.88	1.30	466
....						
15.835 to 17.185	0.45	0.00	-0.59	-0.08	0.35	466
17.185 to 18.675	0.09	0.32	1.01	1.48	3.23	472
18.675 to 21.685	-0.54	-0.65	-0.88	-0.42	0.96	467
TOTAL Q stats	2.85	16.90	15.43	18.04	33.46	7040
df for Q stats	1.95	12.01	12.35	13.00	25.35	0
p-val for Q stats	0.23	0.15	0.24	0.16	0.13	0

The head circumference data: Q.stats()



Comparing GAMLSS and QR: fitted centiles



Conclusions: GAMLSS

- 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 additive terms in the models for the distribution parameters
- The fitted algorithm is modular, where different components can be added easily
- Models can be fitted easily and fast
- Explanatory tool to find appropriate set of models
- 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	

END

for more information see

www.gamlss.org