Smooth centile curves for skew and kurtotic data

Bob Rigby

Mikis Stasinopoulos

Ch12 Centile estimation and diagnostics

12.1 Introduction to centile estimation

12.2 Modelling the distribution

12.3 Modelling the parameters

12.4 Model selection

12.5 Modelling head circumference against age

12.6 Model diagnostics

12.7 Conclusions

12.1 Introduction to centile estimation

12.1.1 Data example

Variables

Head circumference (HEAD)

against age,

for 7040 males under 22 years,

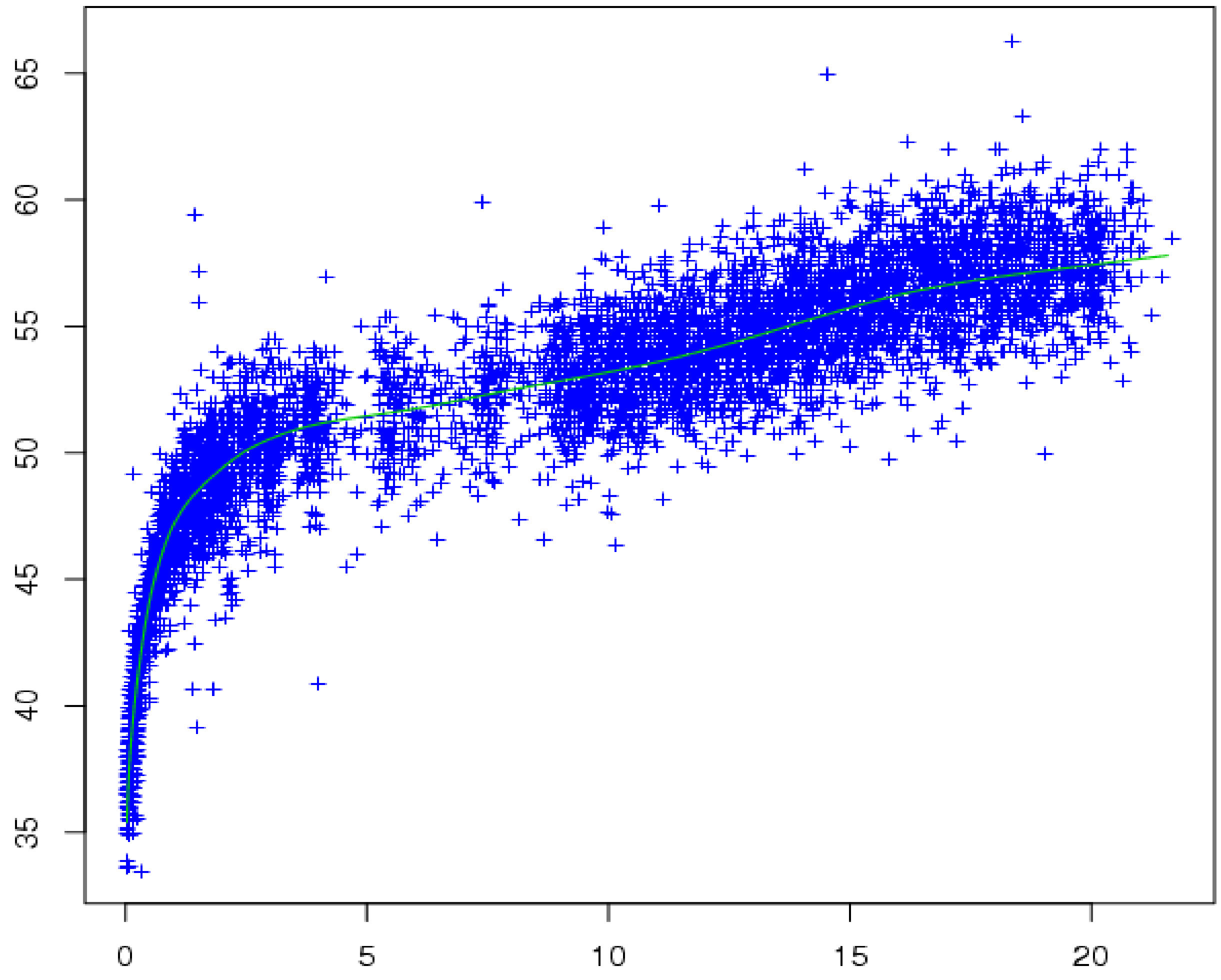
Study

cross sectional study,

'The Fourth Dutch Growth Study',

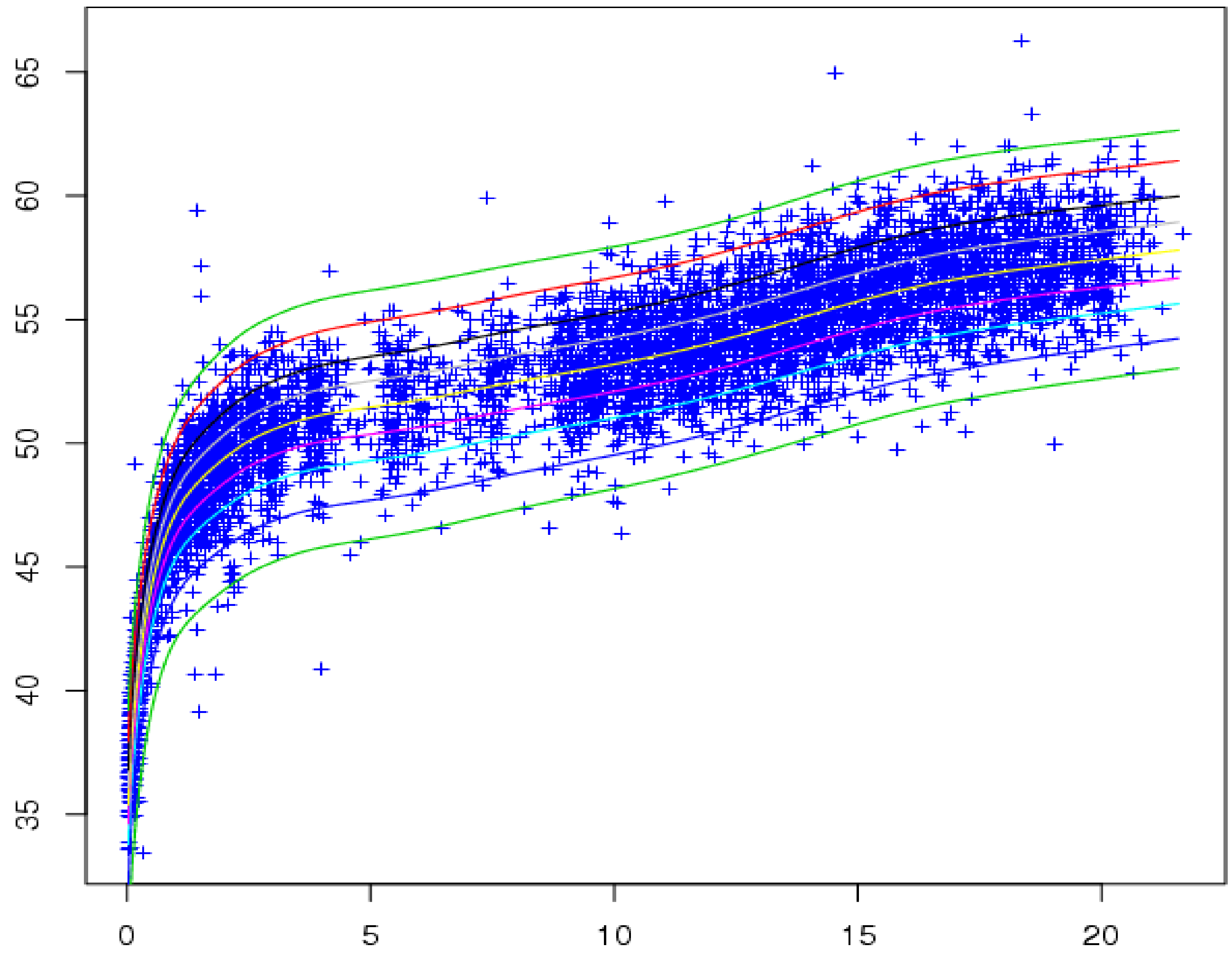
Fredriks *et al.* (2000)

HEAD against AGE for Dutch males



12.1.2 Objective

To obtain centile curves of HEAD against AGE



12.1.3. Methods of centile estimation

Methods based on treating x (AGE) continuously

(i) without distribution model for Y (HEAD)

e.g. HRY method, Healey et al. (1988)

e.g. Quantile regression,

Koenker and Bassett (1978),

Heagerty and Pepe (1999)

(ii) with distribution model for Y (HEAD)

e.g. LMS method, Cole and Green (1992)

e.g. LMSP method, Rigby and Stasinopoulos (2004)

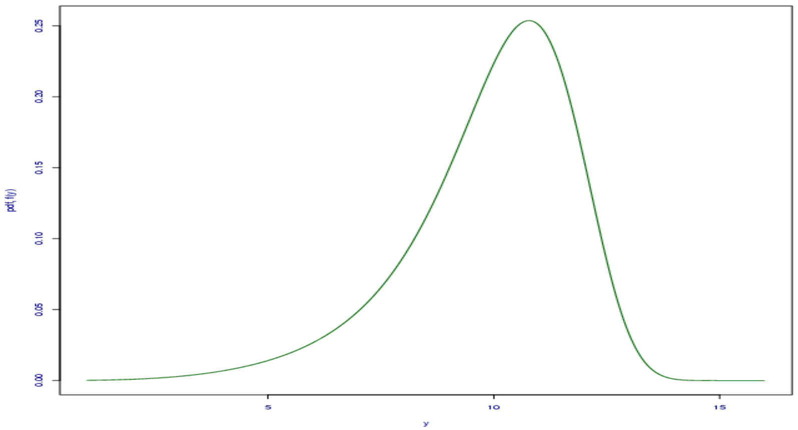
12.2 Modelling the distribution of Y

Four parameters

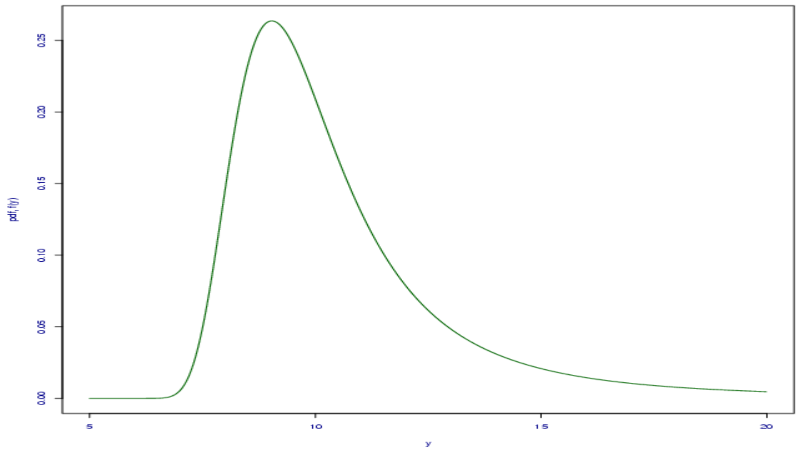
μ	location
σ	scale
ν	skewness
τ	kurtosis

Skewness and kurtosis

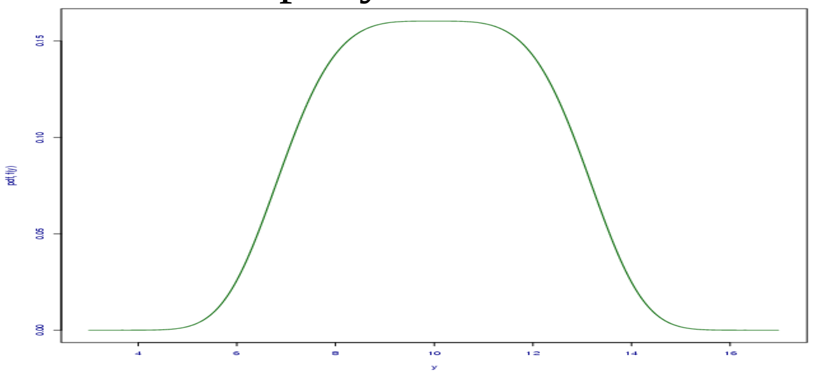
negative skewness



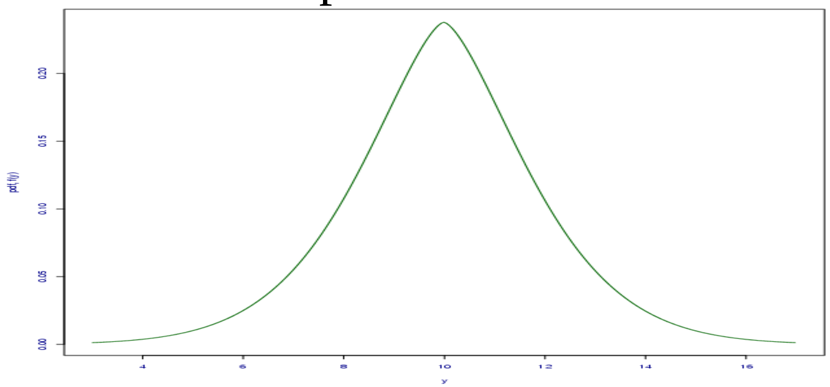
positive skewness



platykurtosis



leptokurtosis



gamlss
for statistical modelling

Continuous distributions for *Y*

Two parameter distributions

BE	Beta
GA	Gamma
GU	Gumbel
LO	Logistic
LNO	Log Normal
NO	Normal
IG	Inverse Gaussian
RG	Reverse Gumbel
WEI	Weibull (also WEI2, WEI3)

Continuous distributions for *Y*

Three parameter distributions

BCCG	Box-Cox Normal (Cole and Green, 1992)
PE	Power Exponential (Box and Tiao, 1962)
TF	*t* family (Lange *et al.*, 1989)

gamlss
for statistical modelling

Continuous distributions for *Y*

Four parameter distributions

BCT	Box-Cox *t* (Rigby and Stasinopoulos, 2006)
BCPE	Box-Cox Power Exponential (Rigby and Stasinopoulos, 2004)
JSU	Johnson Su (Johnson, 1949)
SHASH	Sinh Arc Sinh (Jones, 2005)
SEP	Skew Exponential Power (DiCiccio and Monti,2004)
ST3	Skew *t* (Jones and Faddy, 2003)

Continuous distributions for *Y*

Two parameter distributions

N	unable to model skewness or kurtosis

Three parameter distributions

BCCG	unable to model kurtosis
TF, PE	unable to model skewness

Four parameter distributions

BCT, JSU, ST3	unable to model platykurtosis
BCPE, SHASH, SEP	OK

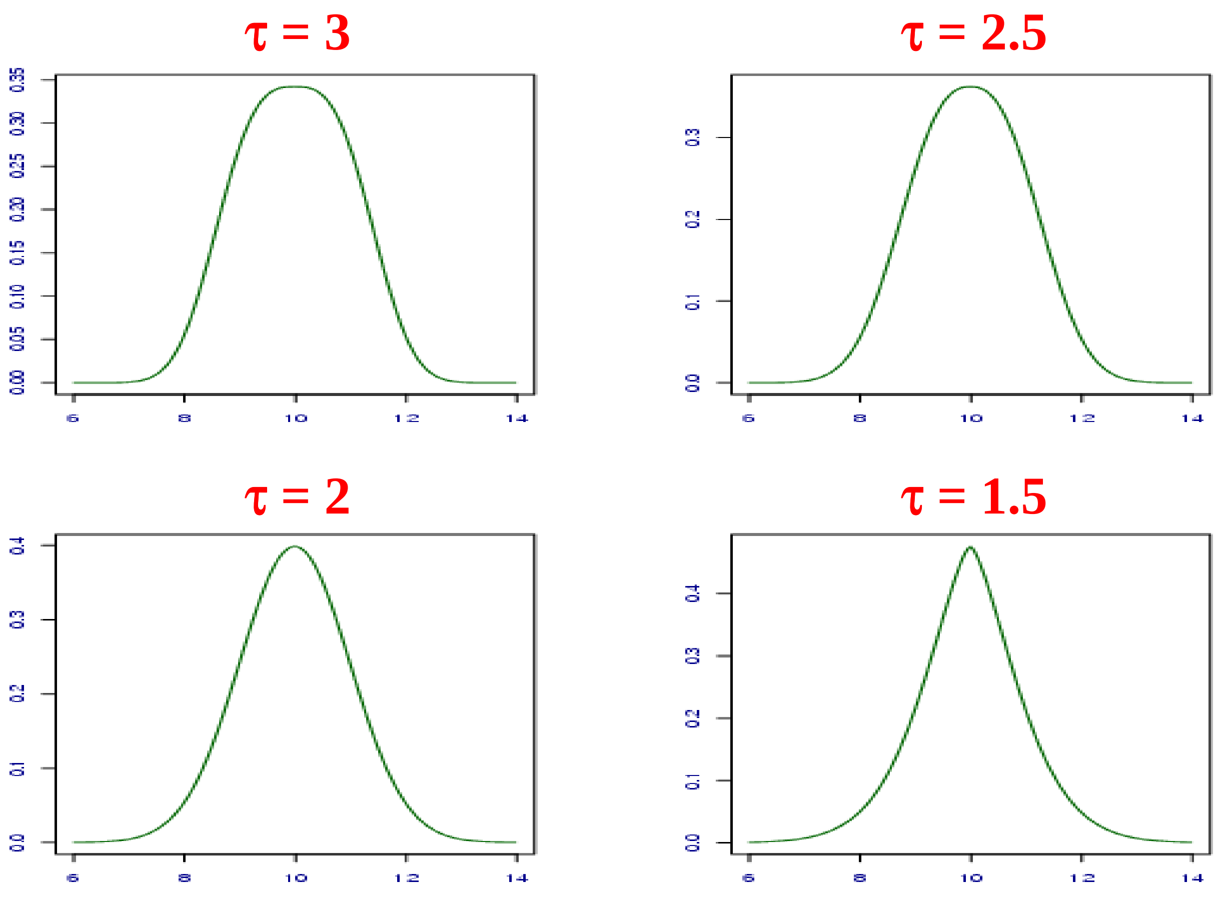
gamlss
for statistical modelling

Box-Cox Power Exponential distribution

parameter	distribution shape
μ	median
σ	coef. of variation (approx)
$\nu>1$ $\nu=1$ $\nu<1$	negatively skew symmetric positively skew
$\tau>2$ $\tau=2$ $\tau<2$	platykurtic mesokurtic leptokurtic

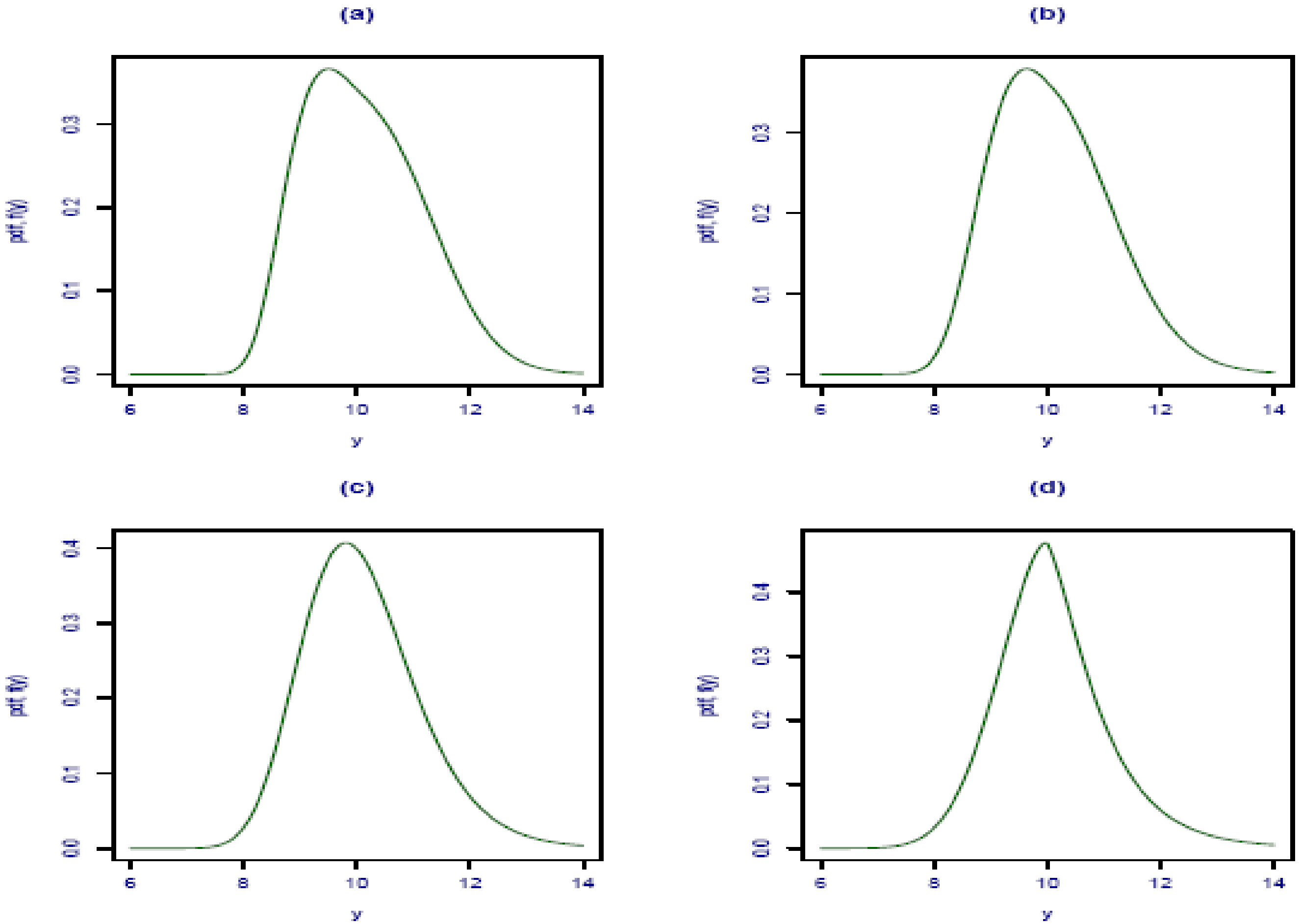
BCPE distribution for ν = 1, symmetric case

[$\mu = 10$, $\sigma = 0.1$, $\nu = 1$, $\tau = 3$, 2.5, 2, 1.5]



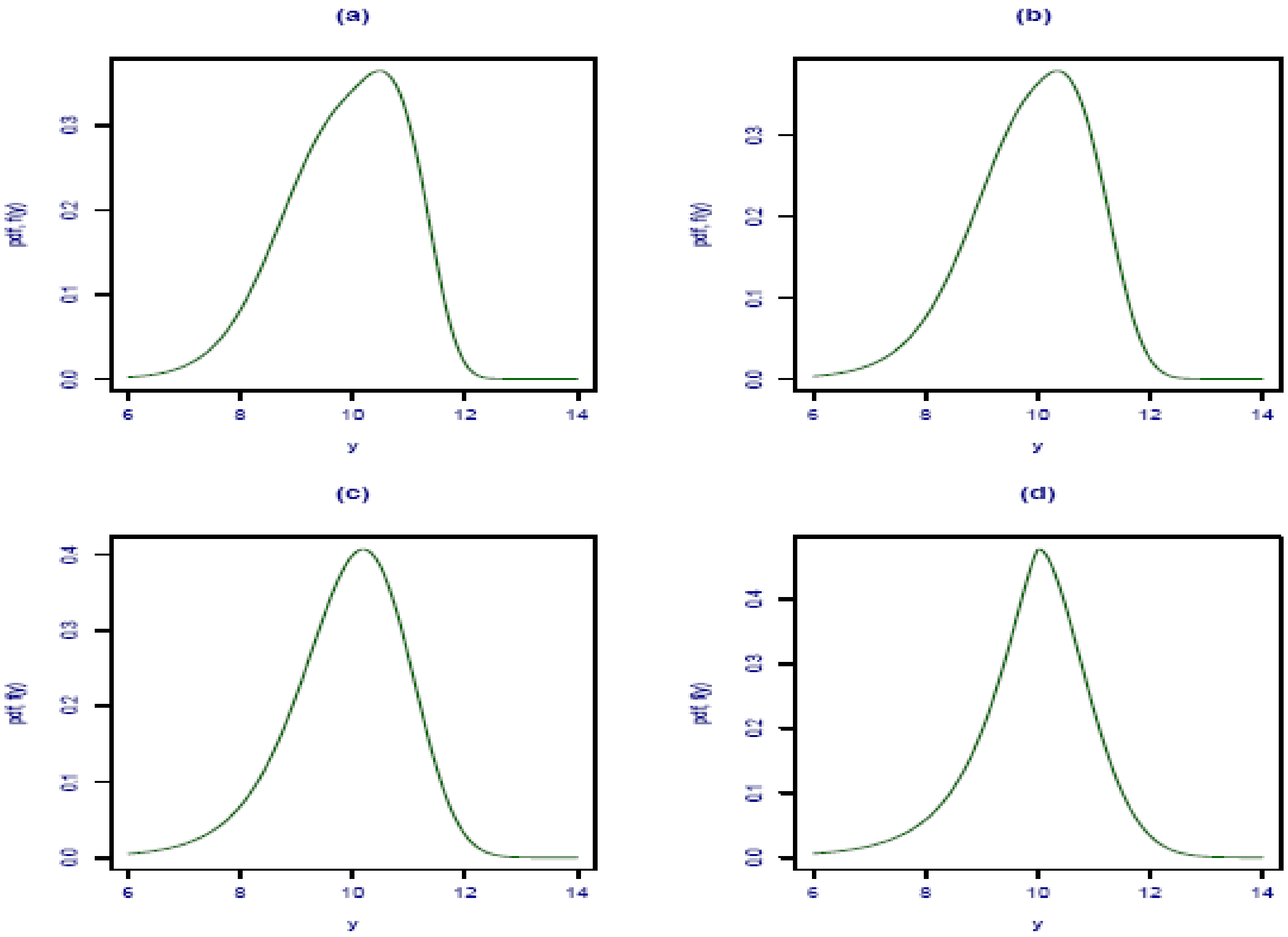
BCPE distribution for ν < 1, i.e. positively skew case

[$\mu = 10$, $\sigma = 0.1$, $\nu = -1$, $\tau = 3$, 2.5, 2, 1.5]



BCPE distribution for $\nu > 1$, i.e. negatively skew case

[$\mu = 10$, $\sigma = 0.1$, $\nu = 3$, $\tau = 3$, 2.5, 2, 1.5]



Definition of distributions

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

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

If $Z \sim N(0,1)$ then $Y \sim BCCG(\mu,\sigma,\nu)$

If $Z \sim t_{\tau}$ then $Y \sim BCT(\mu,\sigma,\nu,\tau)$

If $Z \sim PE(0,1,\tau)$ then $Y \sim BCPE(\mu,\sigma,\nu,\tau)$

gamlss
for statistical modelling

Definition of Johnson Su distribution

Let Y be a random variable with range $-\infty < Y < \infty$, defined through the transformed variable Z given by

$$Z = \nu + \tau \sinh^{-1}\left(\frac{y-\mu}{\sigma}\right)$$

where $\sinh^{-1}(u) = \log[u + (u^2+1)^{1/2}]$

if $Z \sim N(0,1)$, then $Y \sim JSUo(\mu, \sigma, \nu, \tau)$

the original parameterization of Johnson(1949).

12.3 Model for smooth centile curves

$Y \sim D(\mu, \sigma, \nu, \tau)$ where D is any distribution, and

where $Y = HEAD$ and $x = AGE^{\xi}$ and

$$\mu = cs(x, df_{\mu})$$

$$log(\sigma) = cs(x, df_{\sigma})$$

$$\nu = cs(x, df_{\nu})$$

$$log(\tau) = cs(x, df_{\tau})$$

gamlss
for statistical modelling

12.4 Automatic model selection of degrees of freedom and ξ

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

An automatic procedure is used to select the values which minimize the generalized Akaike information criterion GAIC

GAIC(#) = Deviance + #.df

where # is a penalty for each degree of freedom used in the model and df is the total degrees of freedom used in the model, Akaike (1983)

Choosing the penalty

Akaike information criterion AIC uses #=2

Schwartz Bayesian Criterion SBC uses #=log(n)=8.9

Alternative values of # can be used, e.g. 3.

Higher penalty # $\Rightarrow$ smoother but biased centiles.

Lower penalty # $\Rightarrow$ rougher but less biased centiles.

gamlss
for statistical modelling

12.5. Modelling head circumference against age

12.5.1 Choosing the penalty

Model distribution: HEAD ~ BCT(μ, σ, υ, τ)

For each fixed #,

select the optimum degrees of freedom and ξ.

criterion	penalty #	dfμ	dfσ	dfν	dfτ	ξ
AIC	2	20.4	4.7	3.0	8.8	0.09
GAIC	3	12.3	5.7	2	2	0.33
SBC	8.9	8.9	1	1	2	0.41

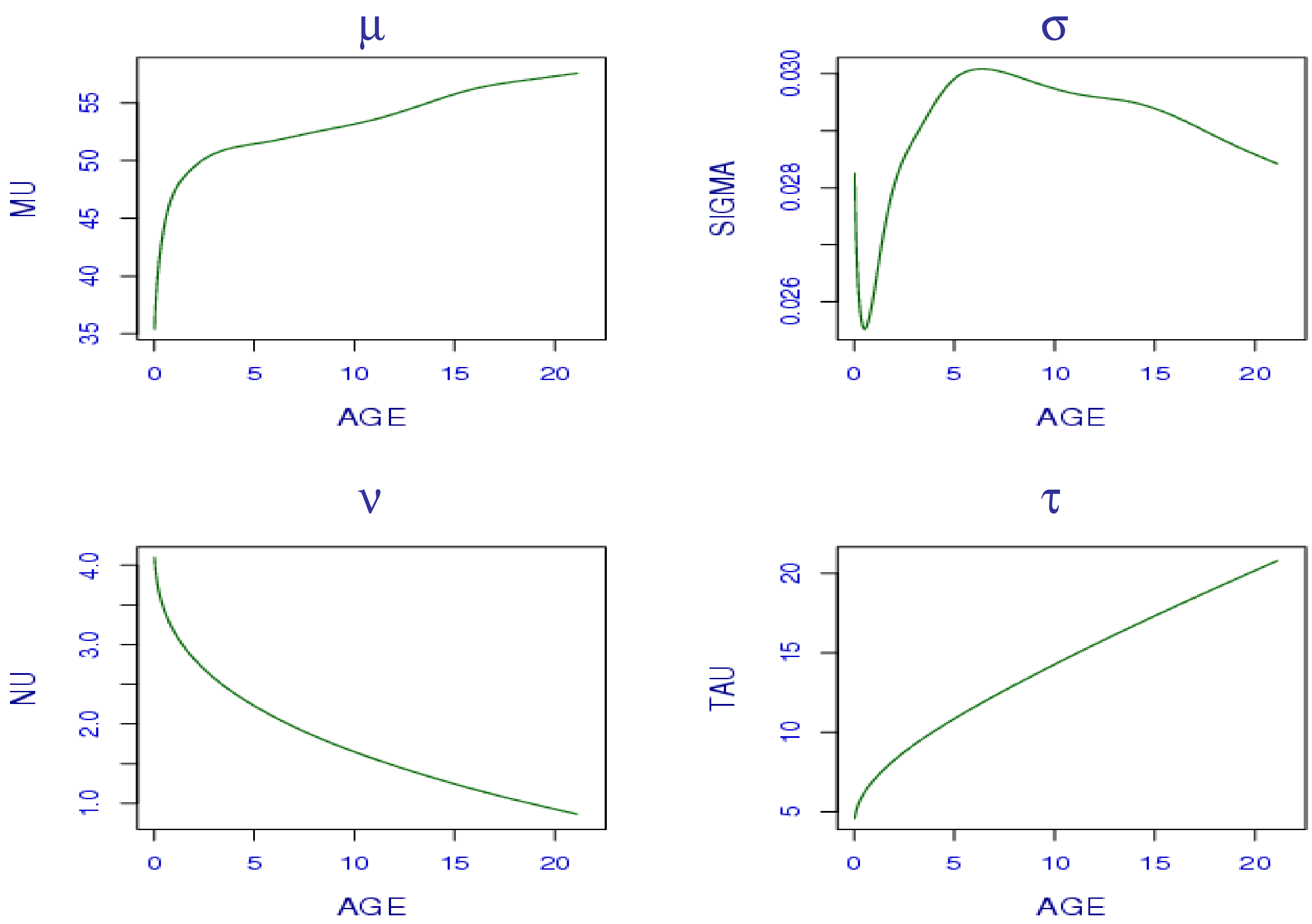
gamlss
for statistical modelling

Box-Cox *t* (BCT) distribution

parameter	distribution shape
μ	median
σ	related to coef. of variation
$\nu>1$ $\nu=1$ $\nu<1$	negatively skew symmetric positively skew
$\tau>0$ $\tau\rightarrow\infty$	leptokurtic mesokurtic

12.5.2 Fitted parameters μ, σ, υ, τ against AGE

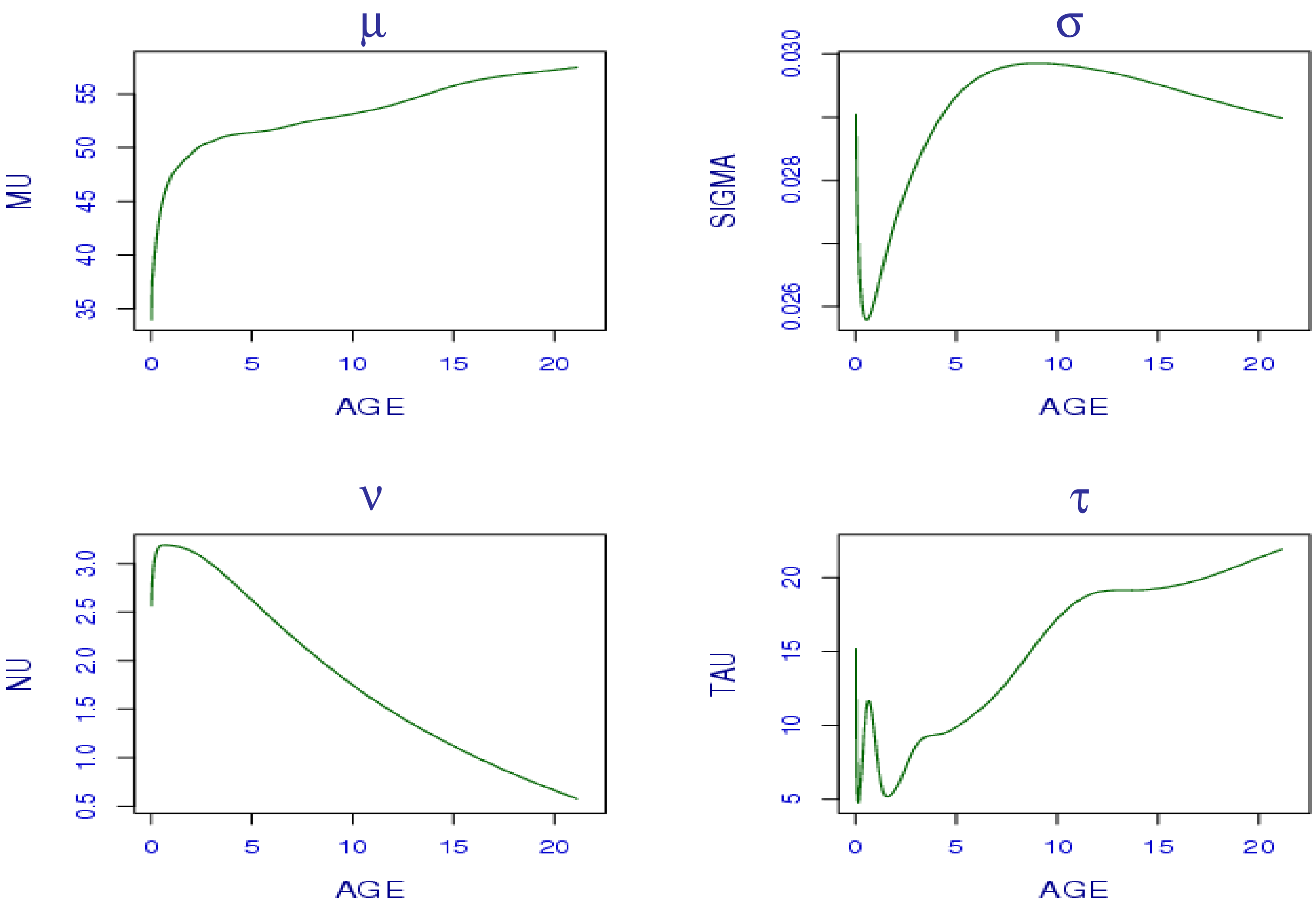
(for BCT model chosen with penalty # = 3)



gamlss
for statistical modelling

Fitted parameters μ, σ, υ, τ against AGE

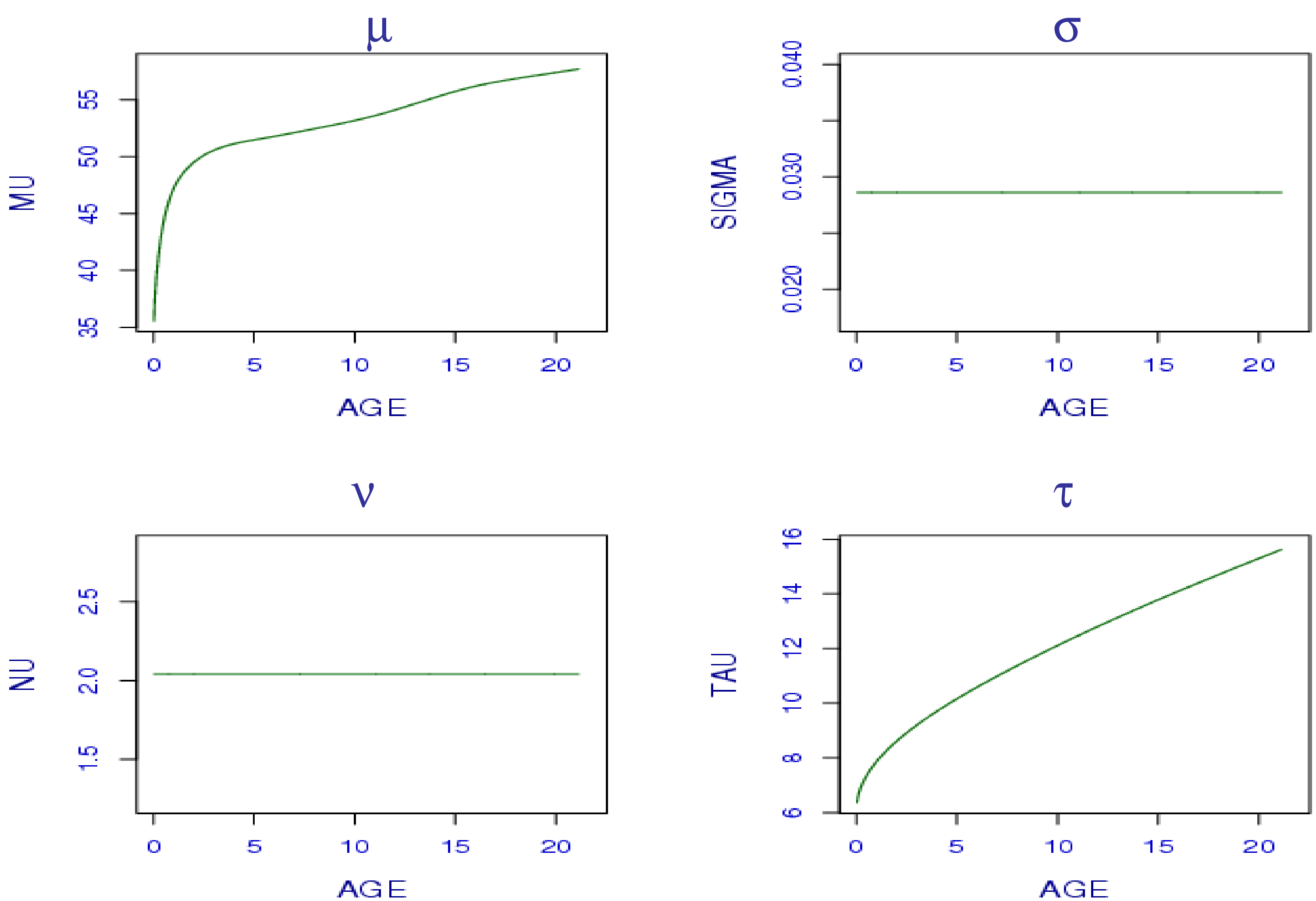
(for BCT model chosen with penalty # = 2)



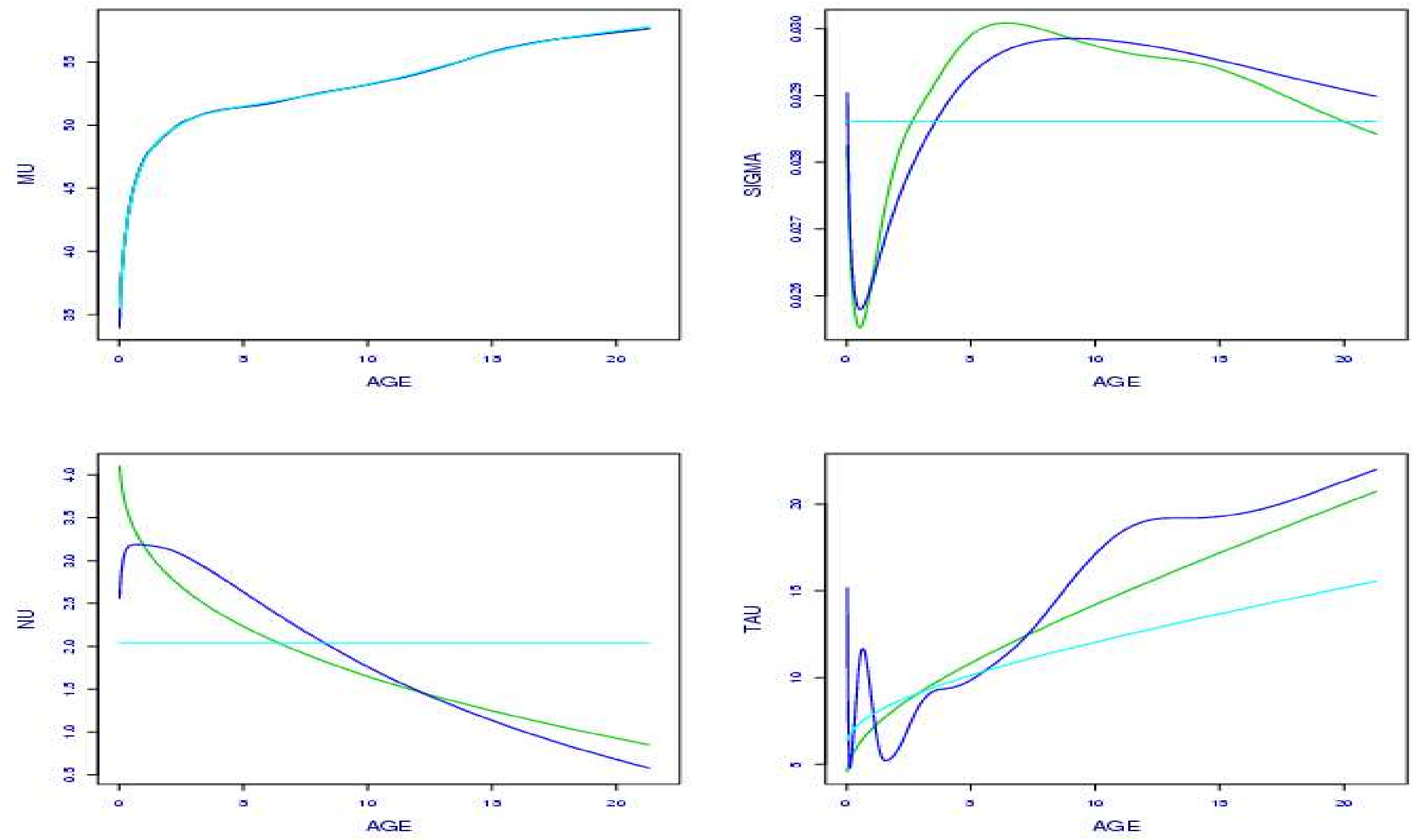
gamlss
for statistical modelling

Fitted parameters μ, σ, υ, τ against AGE

(for BCT model chosen with penalty # = 8.9)



Comparison of fitted models for # = 2, 3, 8.9



12.5.3 Choosing the distribution

distribution	GAIC(3)-26814.7	dfμ	dfσ	dfν	dfτ	ξ
NO	221.6	16.4	30	-	-	0.001
BCCG (LMS)	172.9	16.7	20	14.7	-	0.01
BCPE (LMSP)	81.7	12.2	7.9	2	2	0.34
SEP	71.7	11.7	3.7	2	2	0.40
TF	4.8	13.1	2.9	-	3.1	0.27
JSU	3.4	11.7	3.4	2	2	0.46
BCT (LMST)	0	12.3	5.7	2	2	0.33

gamlss
for statistical modelling

12.5.4 Centile estimates

The 100α centile y_α of Y is given by

$$y_\alpha = \begin{cases} \mu[1+\sigma\nu z_\alpha]^{1/\nu} & \text{if } \nu \neq 0 \\ \mu \exp[\sigma z_\alpha] & \text{if } \nu = 0, \end{cases}$$

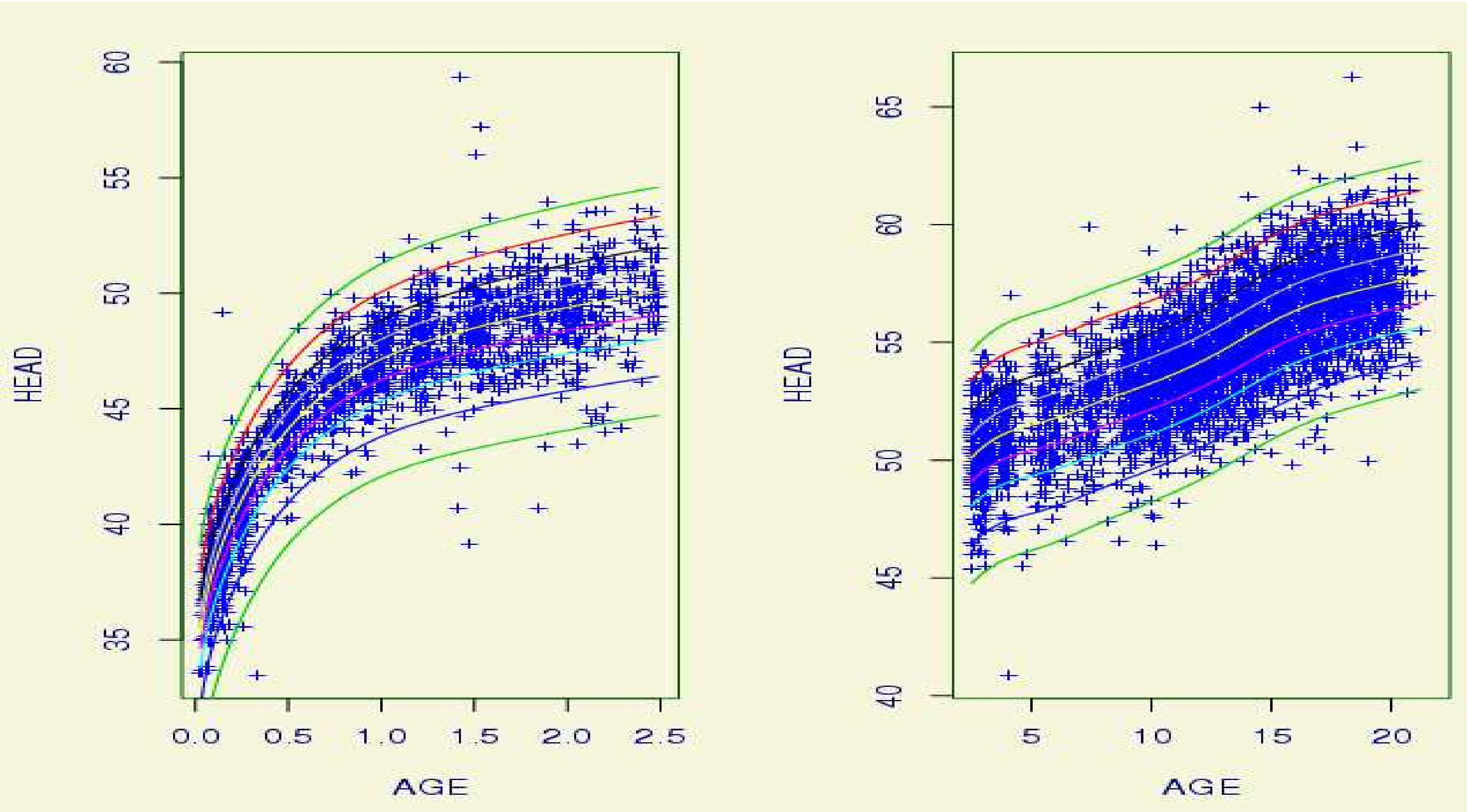
1) if $Y \sim BCCG(\mu,\sigma,\nu)$, then z_α is the $N(0,1)$ centile

2) if $Y \sim BCT(\mu,\sigma,\nu,\tau)$, then z_α is the centile of a t_τ distribution

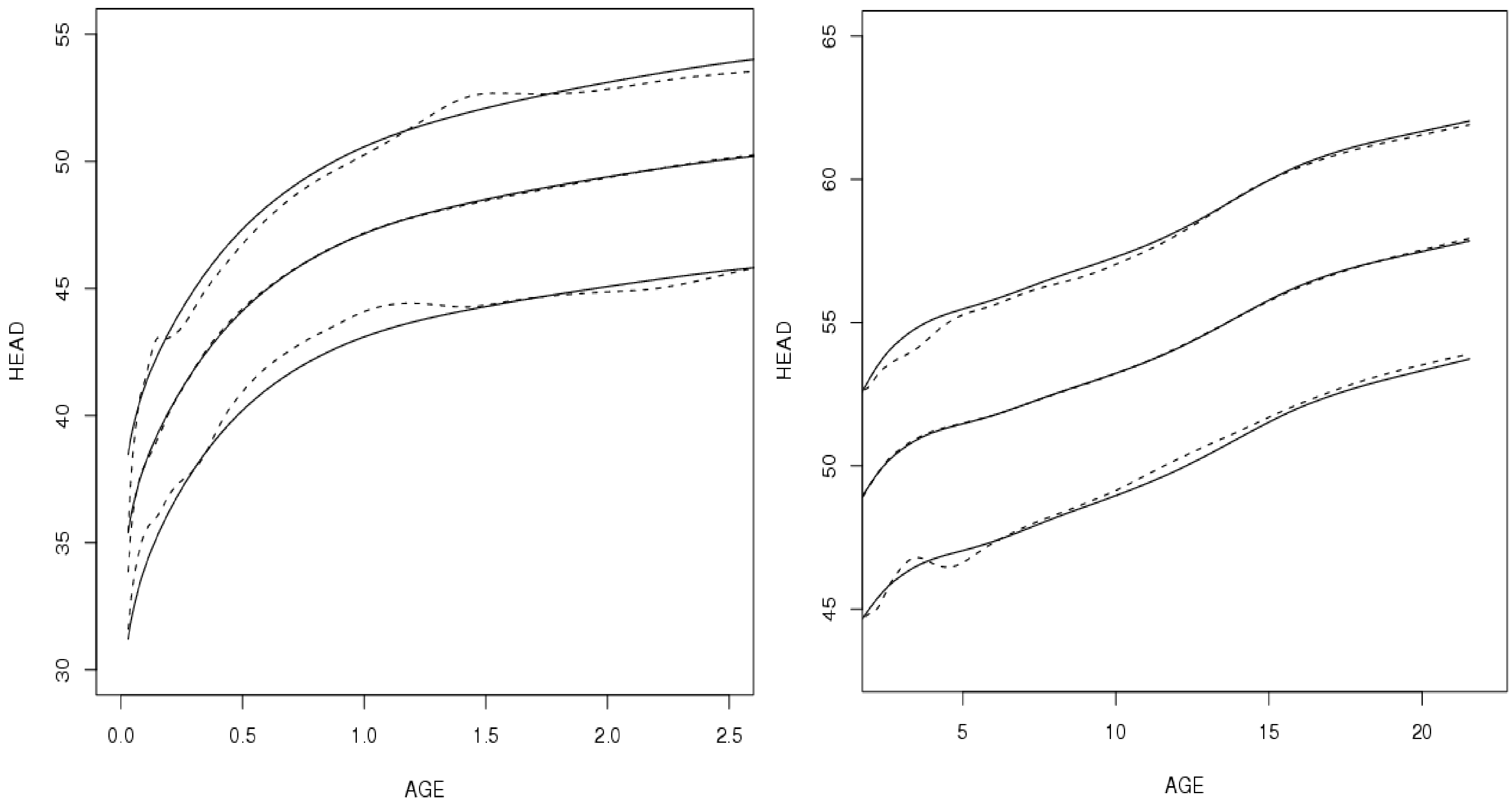
3) if $Y \sim BCPE(\mu,\sigma,\nu,\tau)$, then z_α is the centile of $PE(0,1,\tau)$ a standard power exponential distribution.

Centiles for BCT model chosen with # = 3

(0.4, 2, 10, 25, 50, 75, 90, 98, 99.6) %



Comparison of centiles , (1, 50, 99)%, BCT (—) and BCCG (---) for # = 3



12.6 Model diagnostics

12.6.1 (Normalized quantile) residuals

$$\hat{r} = \Phi^{-1}\left[\hat{F}(y)\right] \quad \text{if Y is continuous}$$

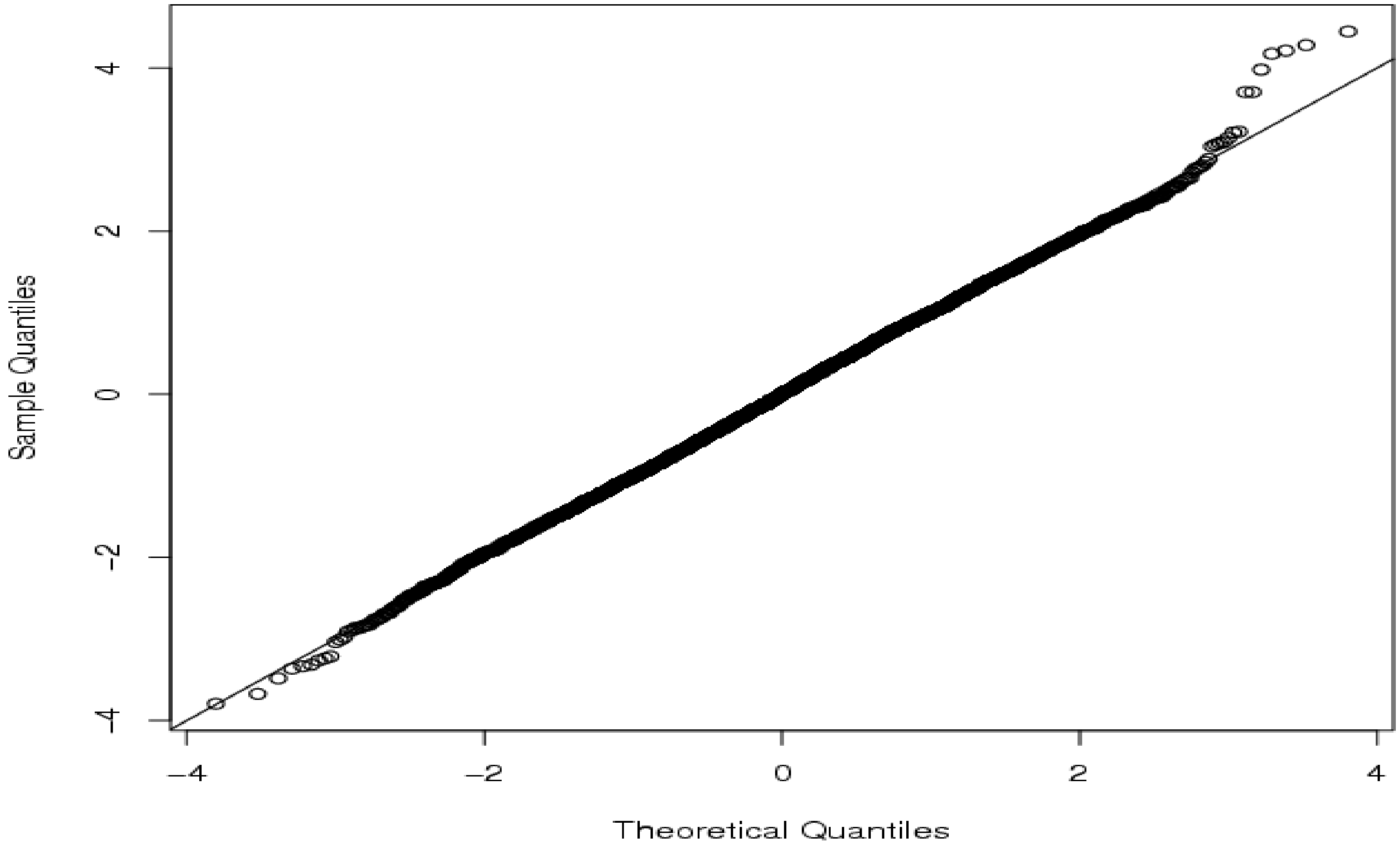
$$\hat{r} = \Phi^{-1}[u] \quad \text{if Y is discrete, where } u \sim U[\hat{F}(y-1), \hat{F}(y)]$$

and Φ^{-1} is the inverse cdf of a N(0,1) variable

and $\hat{F}(y)$ is the fitted model cdf of Y

The true residuals r from the model have a N(0,1) distribution, whatever the original correct distribution of Y.

12.6.2 Residual QQ plot
(for BCT model chosen with penalty # = 3)



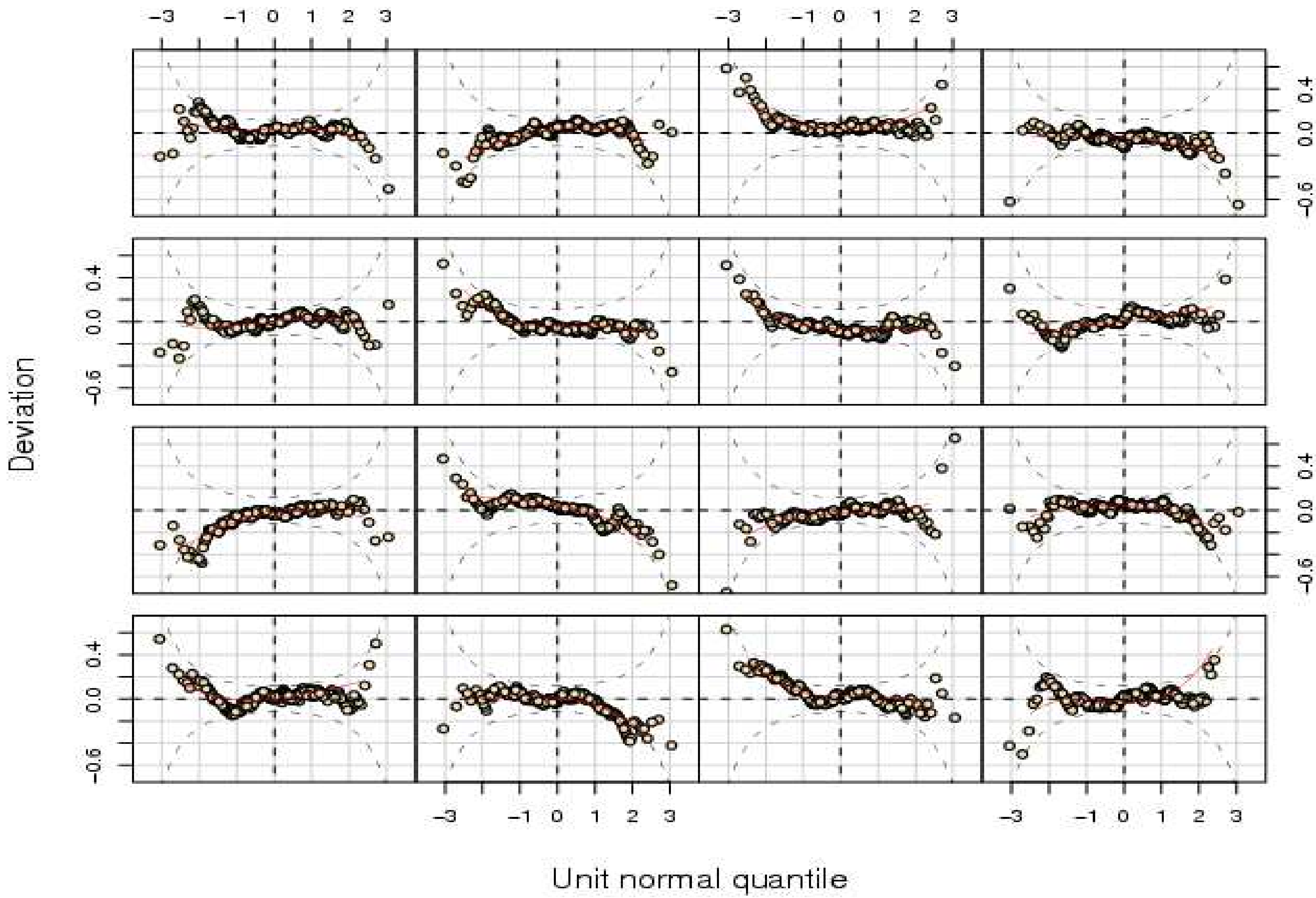
12.6.3 Worm plots

Purpose is detect inadequacies in the model fit within specific age ranges, (van Buuren and Fredriks, 2001).

The age range is split into 16 age range intervals,
with the same number of observations in each age range.

A detrended QQ plot of the residuals in each age range is obtained.

Worm plots (for BCT model with penalty # = 3)



Interpretation of worm coefficients
(scaled by factor 100)

In a particular age range:

b0 < -10	$\Rightarrow$	residual mean too low	$\Rightarrow$	model μ too high
b0 > 10	$\Rightarrow$	residual mean too high	$\Rightarrow$	model μ too low
b1 < -10	$\Rightarrow$	residual sd too low	$\Rightarrow$	model σ too high
b1 > 10	$\Rightarrow$	residual sd too high	$\Rightarrow$	model σ too low
b2 < -5	$\Rightarrow$	residual skewness too low	$\Rightarrow$	model ν too high
b2 > 5	$\Rightarrow$	residual skewness too high	$\Rightarrow$	model ν too low
b3 < -3	$\Rightarrow$	residual kurtosis too low	$\Rightarrow$	model τ too high
b3 > 3	$\Rightarrow$	residual kurtosis too high	$\Rightarrow$	model τ too low

Worm coefficients (scaled by 100)

gamlss
for statistical modelling

group	age range	b0	b1	b2	b3
1	0 - 0.2	-1	4	2.8	-0.6
2	0.2 – 0.7	2	-3	-2.9	-0.5
3	0.7 – 1.2	1	-1	2.0	-1.4
4	1.2 – 1.8	-2	-2	3.1	**2.2**
5	1.8 – 2.5	-1	5	-2.9	0.1
6	2.5 – 3.9	4	**-6**	-1.9	-0.6
7	3.9 – 7.9	-1	2	-0.8	0.7
8	7.9 – 10.0	5	-1	-2.4	-0.2
9	10.0 – 11.2	1	4	-0.7	-0.6
10	11.2 – 12.5	-5	0	2.1	-1.3
11	12.5 – 13.7	**-8**	-1	2.4	-0.9
12	13.7 – 14.8	2	4	0.2	0.2
13	14.8 – 16.1	2	2	0.0	-1.0
14	16.1 – 17.4	5	4	-3.1	-0.1
15	17.4 – 18.8	2	0	**3.7**	-0.1
16	18.8 – 22.0	-4	-1	-1.3	-0.4

12.6.4 Z statistics

Purpose is detect inadequacies in the model fit within specific age ranges, (Royston and Wright, 2000).

AGE is split into 16 ranges. Within each age range,

statistics Z1, Z2, Z3 and Z4 are calculated to test whether the mean, variance, skewness and kurtosis of the residuals are different from their model values, i.e. 0, 1, 0, 3.

In each age range:

Z1 = test statistic for mean 0 of residuals, [i.e. Ho: mean=0 , H_1: mean≠0]

Z2 = test of variance 1 of residuals

Z3 = D'Agostino *et al.* (1990) test statistic for skewness 0 of residuals

Z4 = D'Agostino *et al.* (1990) test statistic for kurtosis 3 of residuals

gamlss
for statistical modelling

Interpretation of Z statistics

In a particular age range:

$Z1 < -2$	$\Rightarrow$	residual mean too low	$\Rightarrow$	model μ too high
$Z1 > 2$	$\Rightarrow$	residual mean too high	$\Rightarrow$	model μ too low
$Z2 < -2$	$\Rightarrow$	residual sd too low	$\Rightarrow$	model σ too high
$Z2 > 2$	$\Rightarrow$	residual sd too high	$\Rightarrow$	model σ too low
$Z3 < -2$	$\Rightarrow$	residual skewness too low	$\Rightarrow$	model ν too high
$Z3 > 2$	$\Rightarrow$	residual skewness too high	$\Rightarrow$	model ν too low
$Z4 < -2$	$\Rightarrow$	residual kurtosis too low	$\Rightarrow$	model τ too high
$Z4 > 2$	$\Rightarrow$	residual kurtosis too high	$\Rightarrow$	model τ too low

Z statistics (for BCT model penalty # = 3)

gamlss
for statistical modelling

group	age range	Z1	Z2	Z3	Z4
1	0 - 0.2	0.3	0.7	1.5	0.5
2	0.2 – 0.7	-0.3	-1.4	-1.5	-0.3
3	0.7 – 1.2	0.6	-1.4	1.0	-1.5
4	1.2 – 1.8	0.2	1.6	**2.0**	**2.9**
5	1.8 – 2.5	-0.9	1.7	-1.4	0.3
6	2.5 – 3.9	0.3	**-2.2**	-1.0	-0.4
7	3.9 – 7.9	-0.5	1.2	-0.4	0.9
8	7.9 – 10.0	0.5	-0.4	-1.3	0.0
9	10.0 – 11.2	0.0	0.7	-0.4	-0.5
10	11.2 – 12.5	-0.8	-1.2	1.1	-1.3
11	12.5 – 13.7	**-1.2**	-1.0	1.2	-0.7
12	13.7 – 14.8	0.4	1.5	0.3	0.9
13	14.8 – 16.1	0.5	-0.3	0.0	-0.8
14	16.1 – 17.4	0.4	1.1	-1.5	0.1
15	17.4 – 18.8	**1.2**	-0.1	**2.0**	0.9
16	18.8 – 22.0	-1.1	-0.7	-0.7	0.0

12.6.5 Q statistics

Q statistics are calculated by summing the squared Z statistics over the 16 age ranges, i.e.

$$Q_j = \sum_{g=1}^{G} Z_{gj}^2$$

for j = 1, 2, 3, 4.

Q statistics

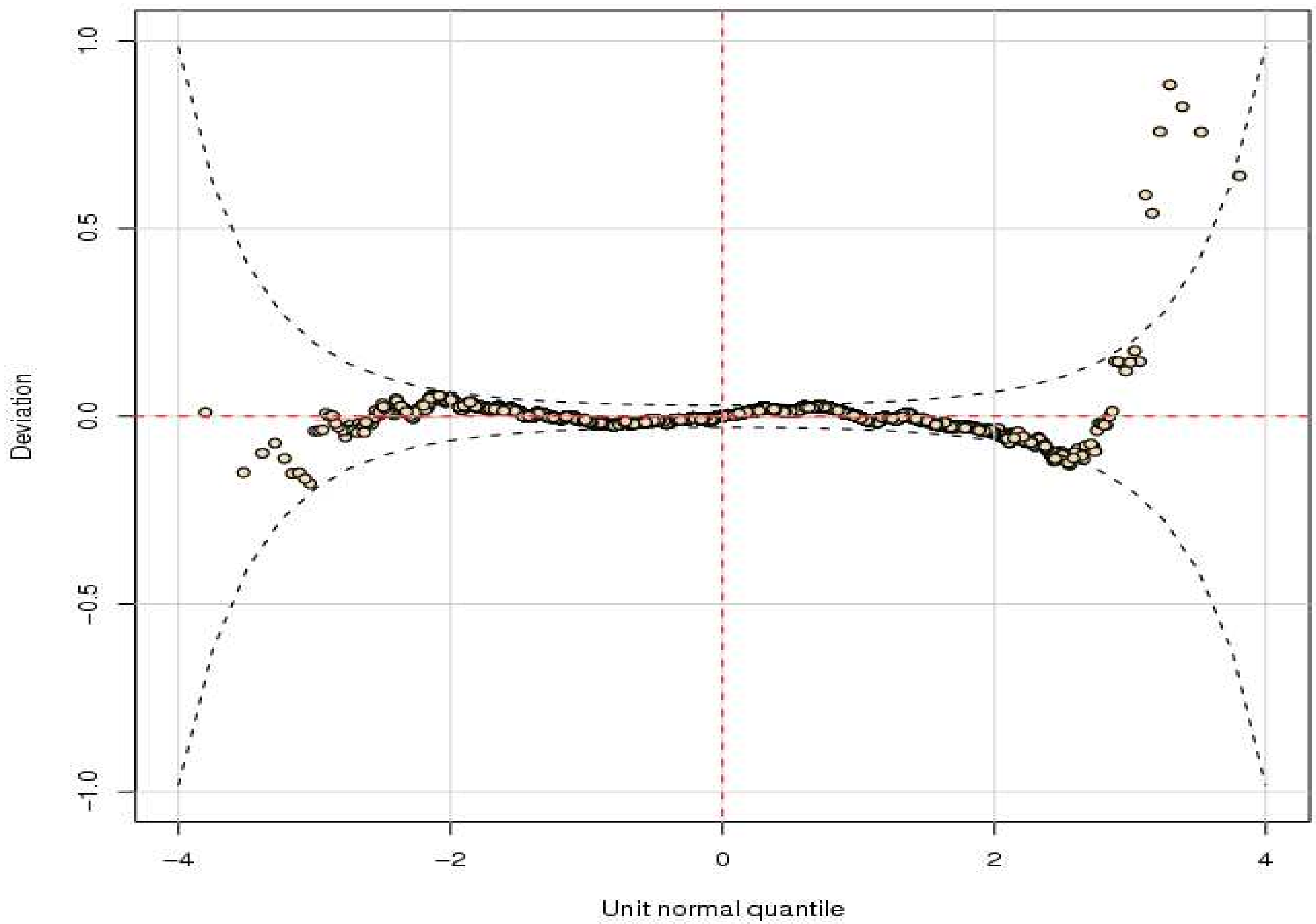
(for BCT model chosen with penalty # = 3)

	Q1	Q2	Q3	Q4
value	7.2	23.3	23.9	16.7
df	3.7	12.7	14.0	14.0
p-value	0.10	0.03	0.05	0.28

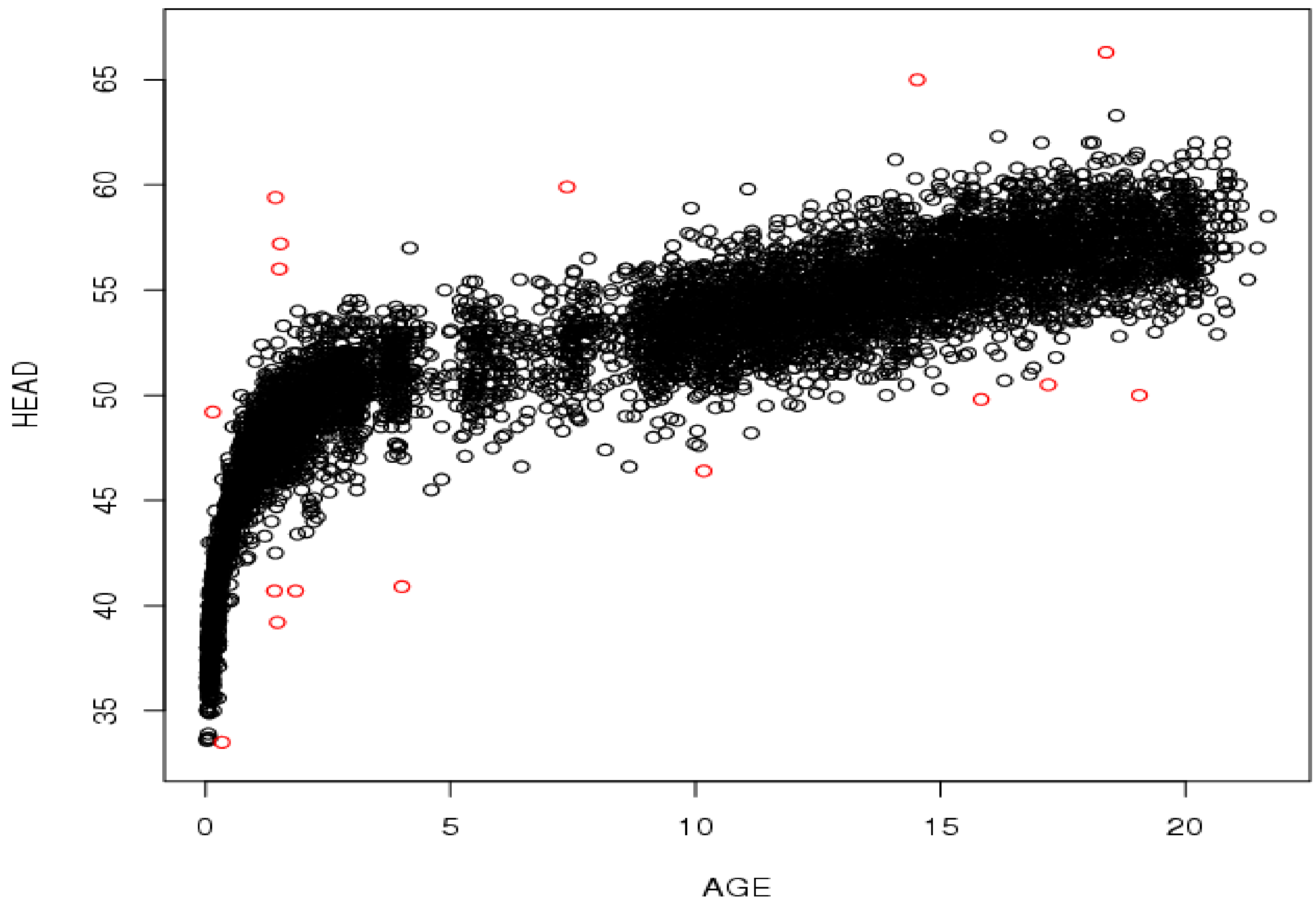
gamlss for statistical modelling

Detrended QQ plot

(for BCT model with penalty # = 3 for all data)

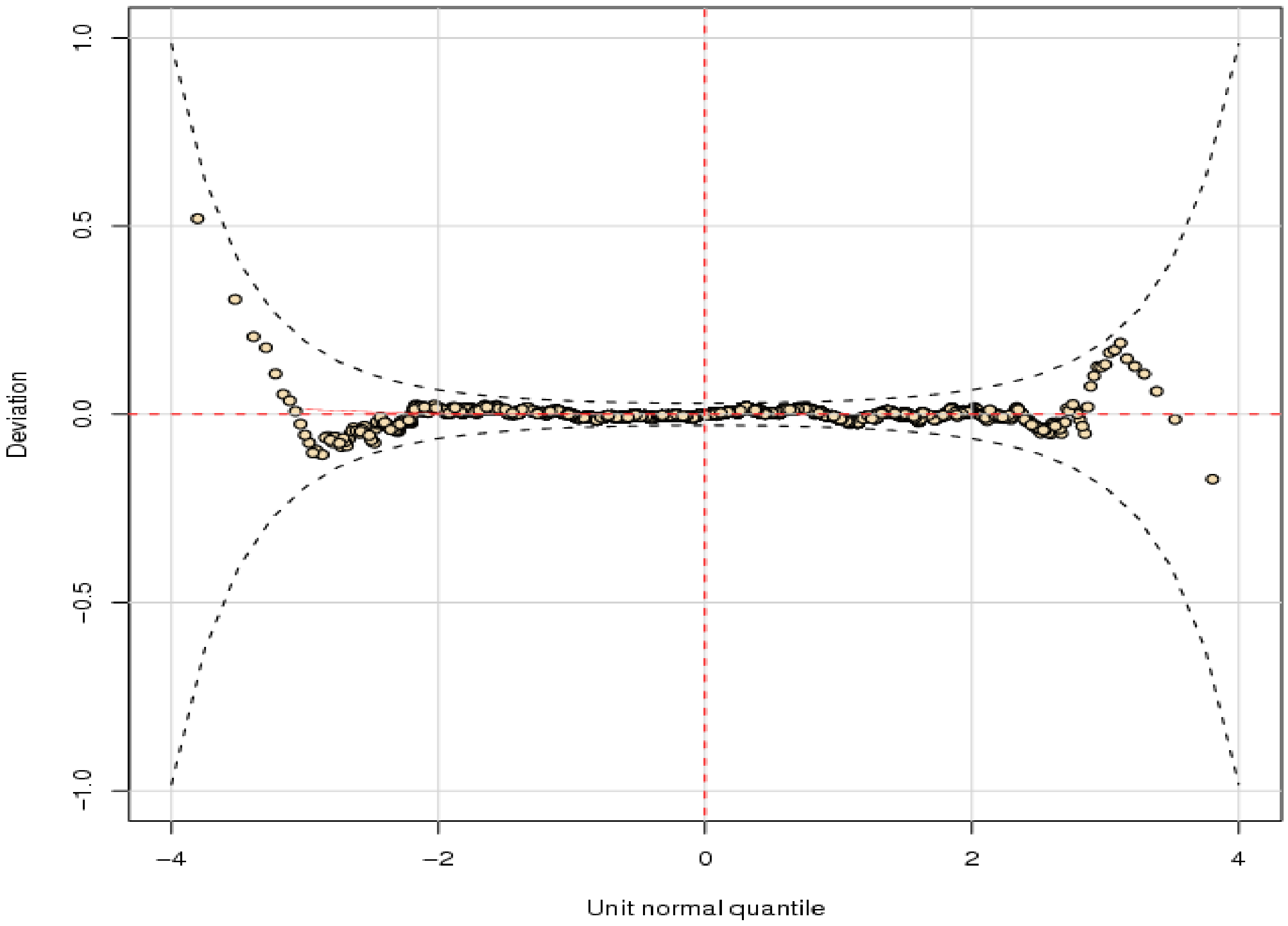


Outliers identified



Detrended QQ plot

(for BCT model with penalty # = 3 when 16 outliers removed)



Q statistics

(for BCT model with penalty # = 3 when 16 outliers removed)

	Q1	Q2	Q3	Q4
value	6.5	18.2	15.6	5.2
df	3.2	12.9	14	14.0
p-value	0.10	0.14	0.34	0.98

gamlss
for statistical modelling

12.7 Conclusion

GAMLSS allows flexible centile estimation,
by flexible modelling of both:

1) distribution of Y, including models for skewness (positive and negative) and kurtosis (platykurtosis or leptokurtosis)

2) dependence of the parameters μ, σ, υ, τ on explanatory variable(s).

gamlss
for statistical modelling

References

Akaike, H. (1983) Information measures and model selection. *Bulletin of the International Statistical Institute*, **50**, 277-290.

Box, G. E. P. and Tiao, G. C. (1962) A further look at robustness via Bayes's theorem. *Biometrika*, **49**, 419-432.

Cole, T. J. (1988) Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society, Series A*, **151**, 385-418.

Cole, T. J. and Green, P. J. (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*, **11**, 1305-1319.

Fredriks, A. M. *et al.* (2000) Body index measurements in 1996-7 compared with 1980. *Archives of Childhood*, **82** , 107-112.

Hansen, B. (1994) Autoregressive conditional density estimation. *International Economic Review*, **35**, 705-730.

Heagerty, P. J. and Pepe, M. S. (1999) Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Applied Statistics*, **48**, 533-551.

Healey, M. J. R., Rasbash, J. and Yang, M. (1988) Distribution free estimation of age-related centiles. *Annals of Human Biology*, **15**, 17-22.

Johnson, N. L. (1949) Systems of frequency curves generated by methods of translation. *Biometrika*, **36**, 149-176.

Koenker, R. W. and Bassett, G. W. (1978) Regression quantiles. *Econometrica*, **46**, 33-50.

Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989) Robust statistical modelling using the *t* distribution. *Journal of the American Statistical Association*, **84**, 881-896.

Rigby, R. A. and Stasinopoulos, D. M. (2004a) Smooth centile curves for skew and kurtotic data modelled using the Box-Cox Power Exponential distribution. (to appear in *Statistics and Medicine*).

Rigby, R. A. and Stasinopoulos, D. M. (2004b) Box-Cox *t* distribution for modelling skew and leptokurtic data with an application to centile estimation. (submitted for publication).

Rigby, R. A. and Stasinopoulos, D. M. (2004c) Generalized additive models for location, scale and shape. (accepted as a read paper by the U.K. *Royal Statistical Society*).

Royston, P. and Wright, E. M. (1998) A method of estimating age-specific reference intervals ('normal ranges') based on fractional polynomials and exponential transformation. *Journal of the Royal Statistical Society*, *Series A*, **161**, 79-101.

Royston, P. and Wright, E. M. (2000) Goodness-of-fit statistics for age-specific reference intervals. *Statistics in Medicine*, **19**, 2943-2962

van Buuren, S. and Fredriks, M. (2001) Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, **20**, 1259-1277.

Extra table

criterion	penalty	df_μ	df_σ	df_ν	df_τ	λ
AIC	$\sharp = 2$	11.0	12.2	8.8	2	0.36
GAIC	$\sharp = 3$	10.5	5.8	5.3	2	0.37
SBC	$\sharp = 8.9$	8.1	4.4	3.9	1	0.30

1.1 Data

Variables

Body Mass Index (BMI) against age,

for 7294 males under 22 years,

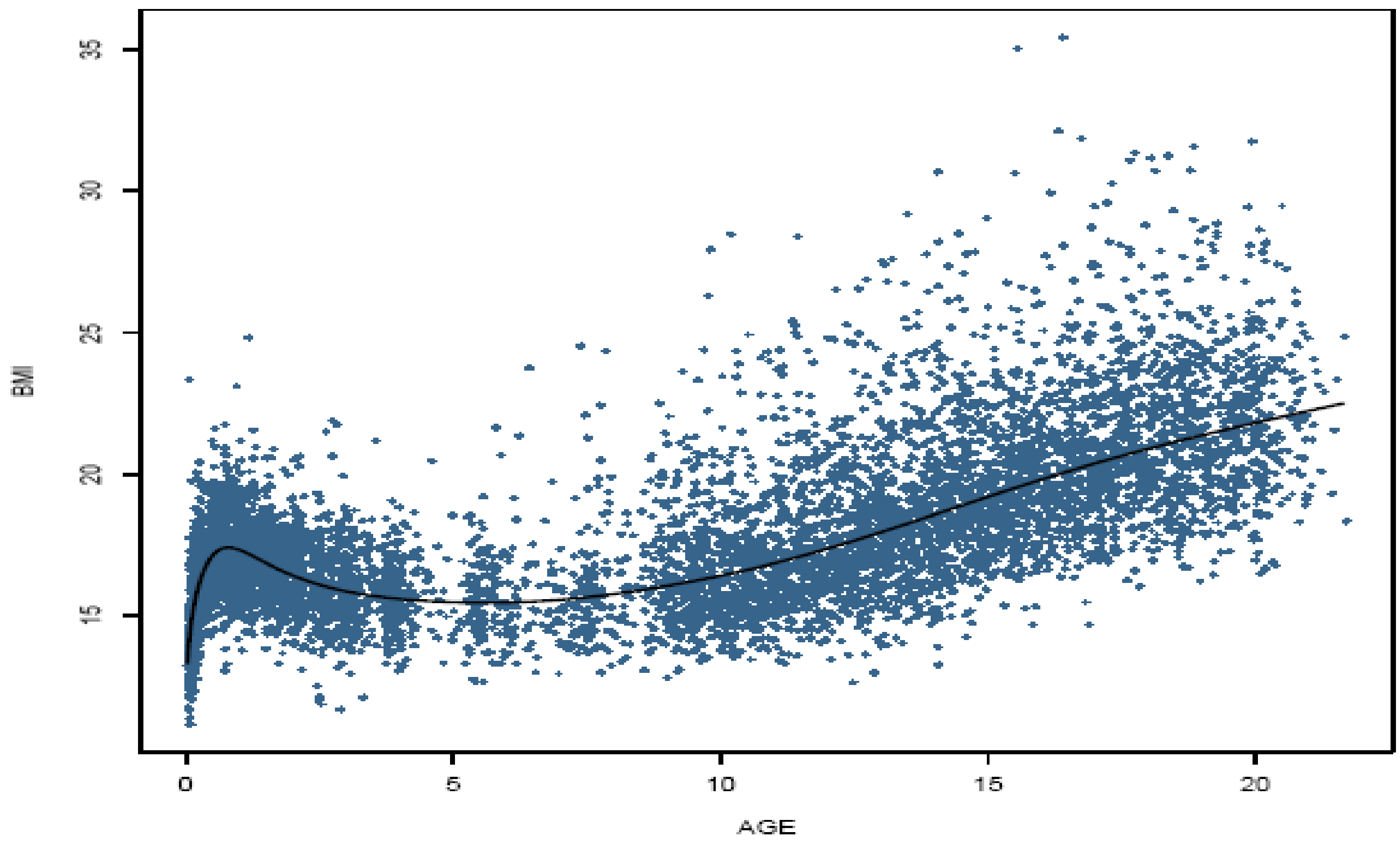
Study

cross sectional study,

‘The Fourth Dutch Growth Study’,

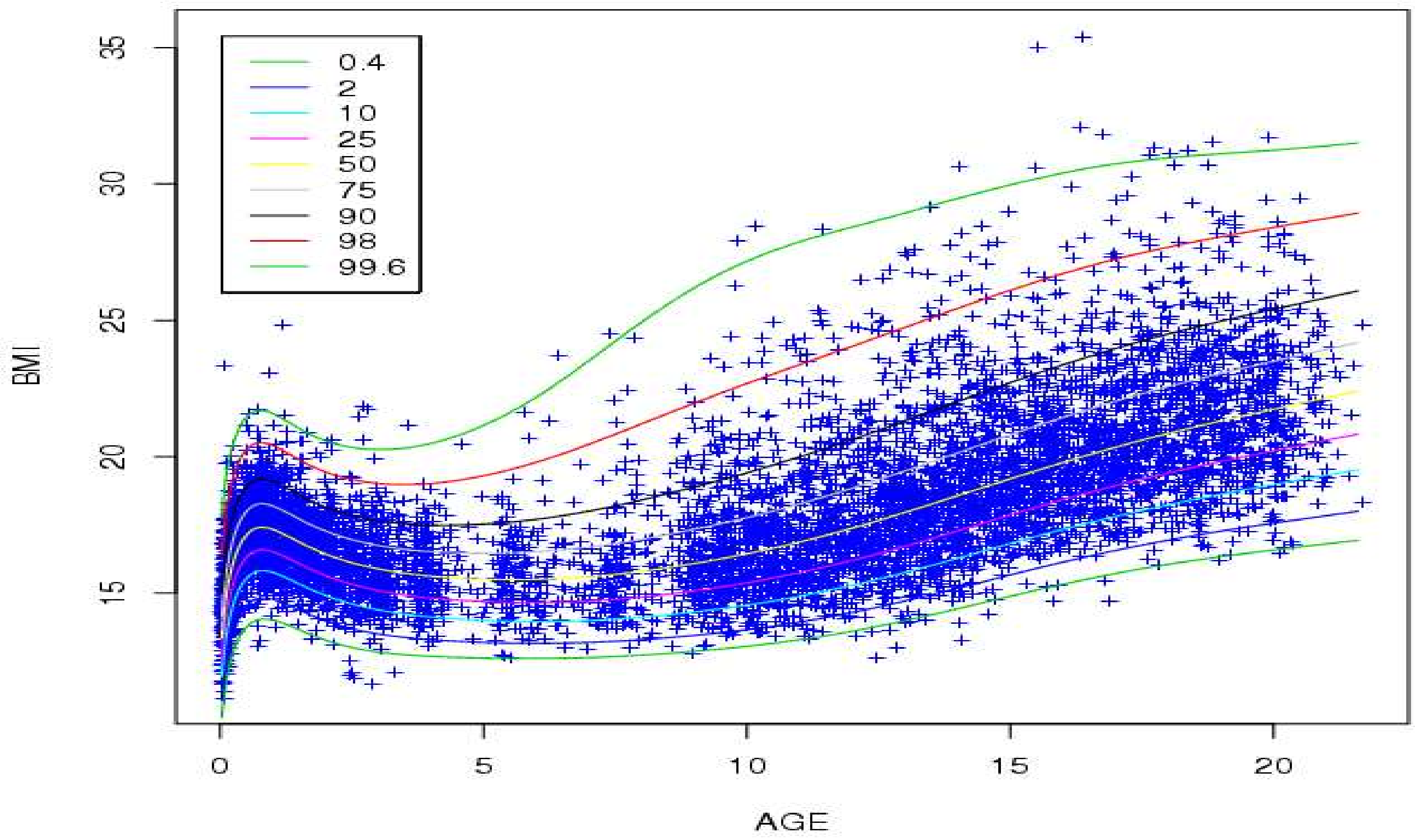
Fredriks et al. (2000)

BMI against AGE for Dutch males



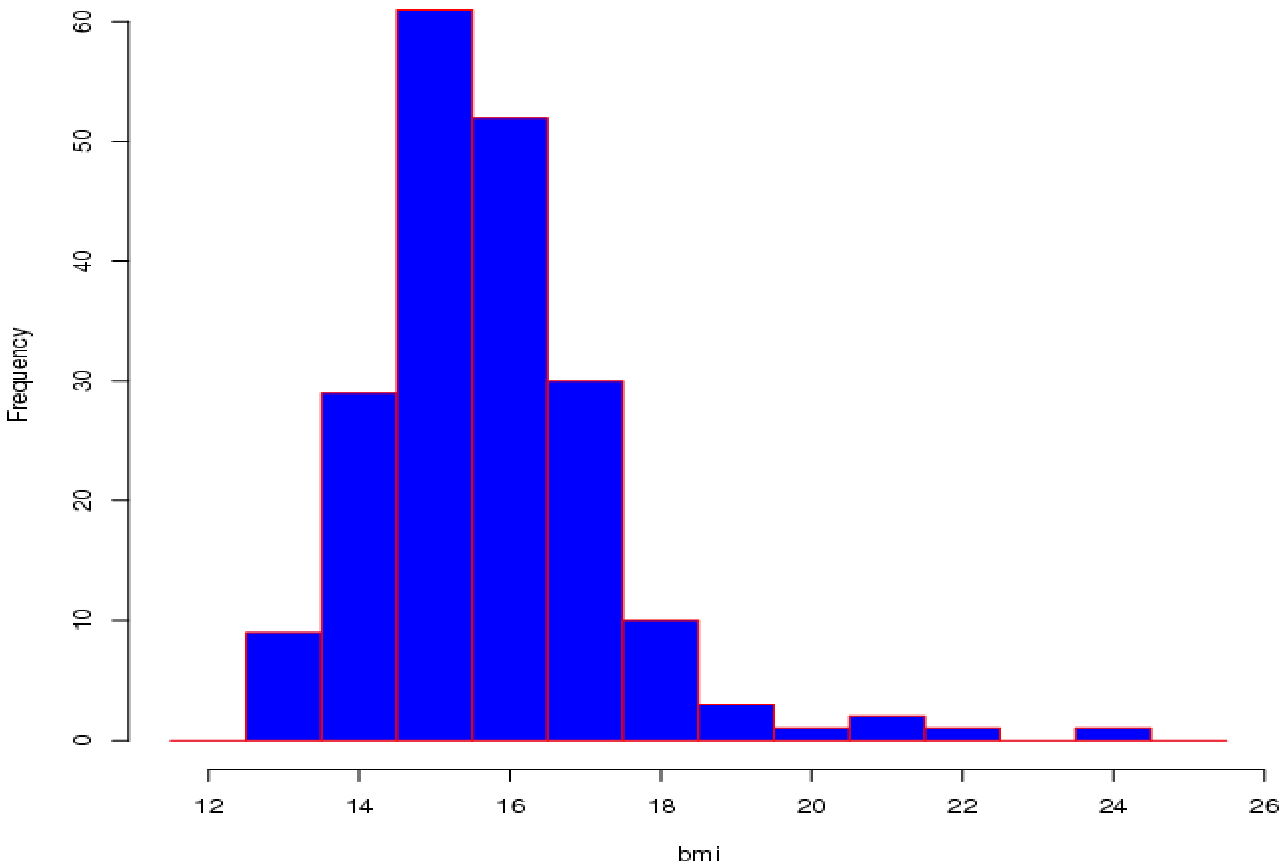
1.2. Objective

To obtain centile curves of BMI against AGE



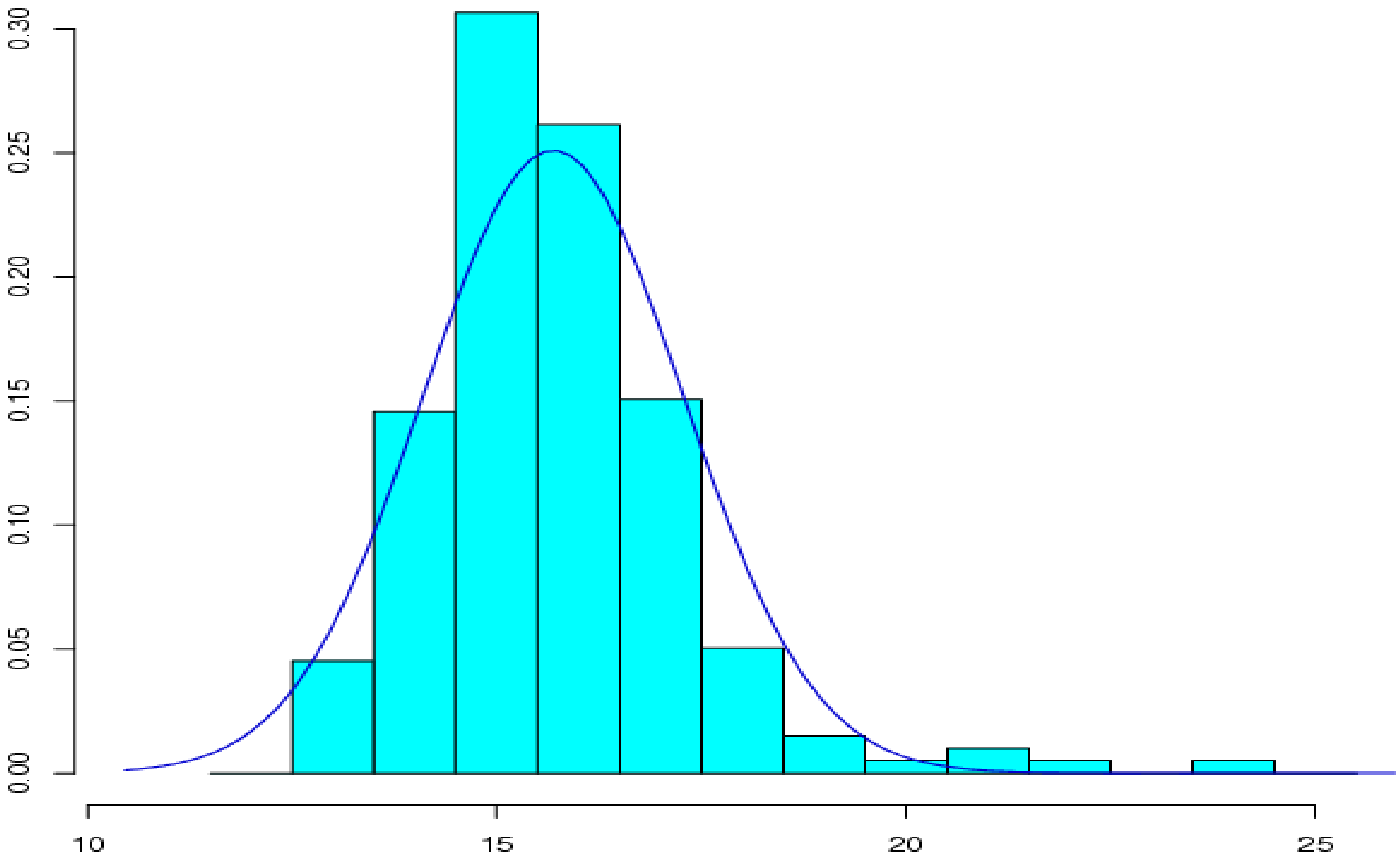
Example 2

Consider the values of BMI for ages from 5 up to 7 years as a single sample

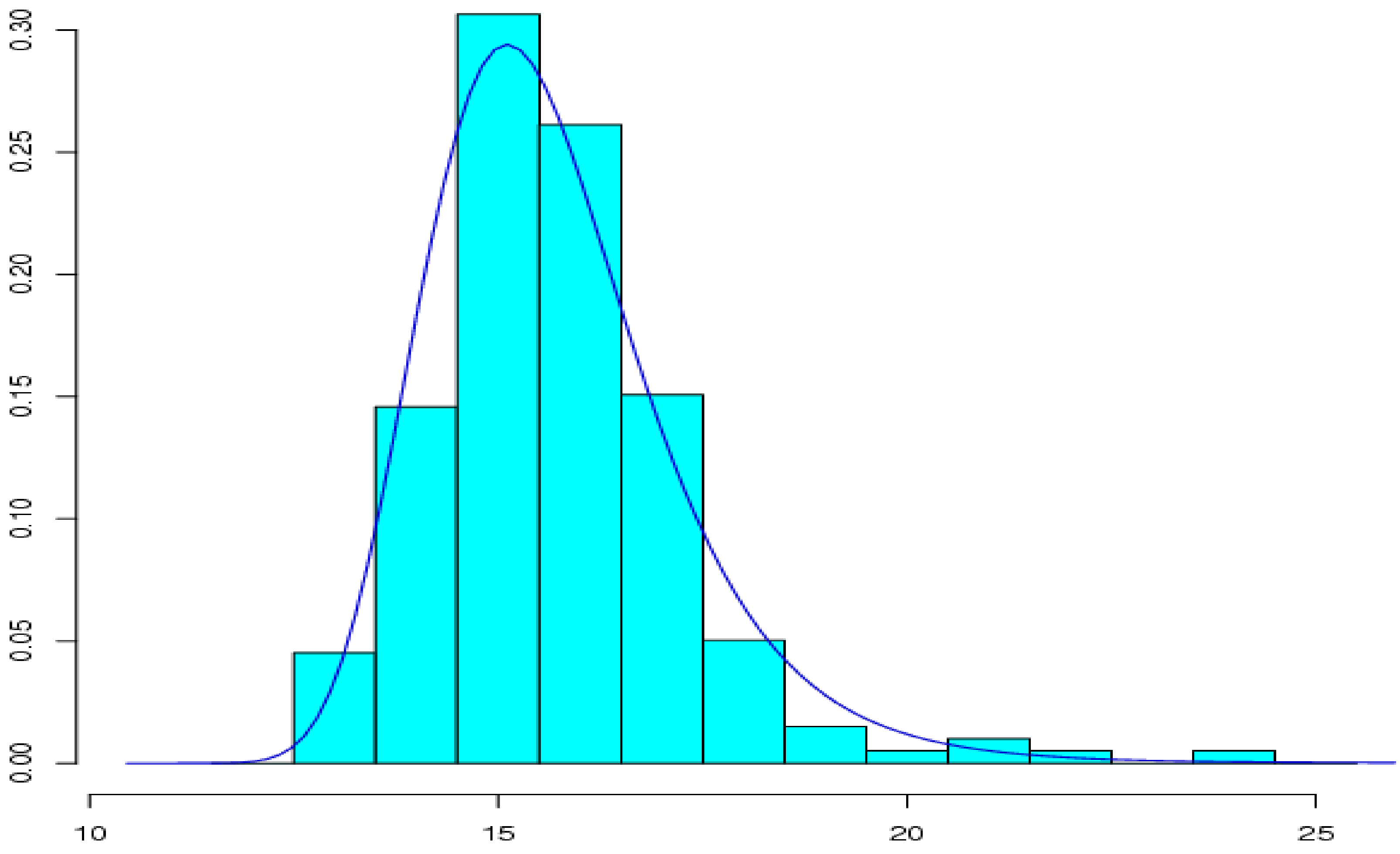


gamlss
for statistical modelling

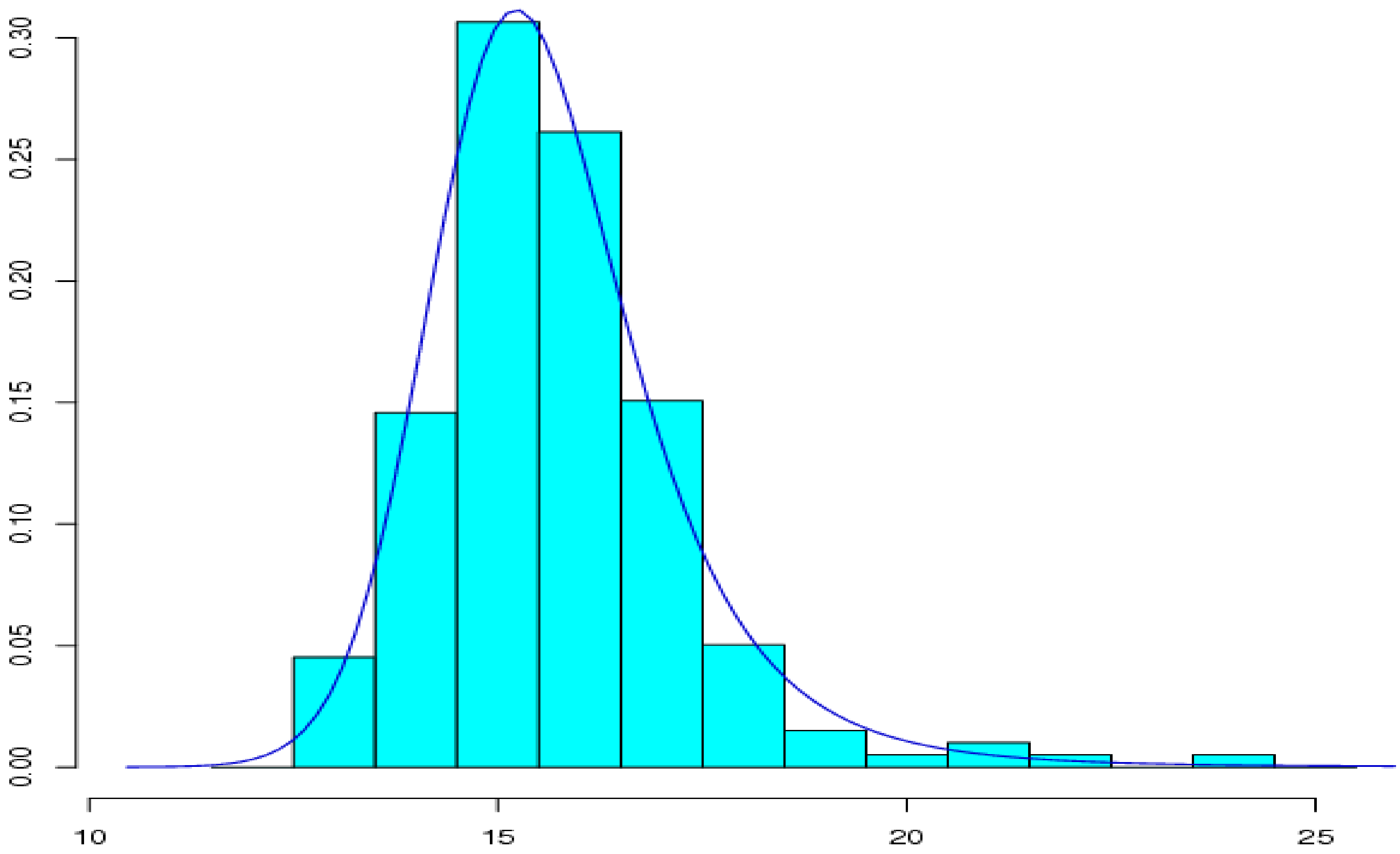
Normal (N) distribution fit to BMI sample



Box-Cox Normal (BCCG) fit to BMI sample



Box-Cox *t* (BCT) fit to BMI sample



gamlss
for statistical modelling

Table of deviances for different distribution fits

Distribution	Deviance - 710.5
N	38.5
BCCG	2.0
JSU	1.1
BCPE	0.2
BCT	0

gamlss
for statistical modelling

Model for smooth centile curves

$Y \sim D(\mu, \sigma, \nu, \tau)$ where D is a distribution,

where $x = \text{age}^{\lambda}$ and

$$\mu = h_1(x)$$

$$log(\sigma) = h_2(x)$$

$$\nu = h_3(x)$$

$$log(\tau) = h_4(x)$$

gamlss for statistical modelling

3. Modelling the parameters

Each parameter of the distribution μ, σ, ν, τ

is modelled in terms of the explanatory variable x

Model	Formula	gamlss code
simple linear	$\mu = \beta_0 + \beta_1 x$	x
polynomial	$\mu = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$	poly(x,3)
fractional polynomial	$\mu = \beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2}$	fp(x,2)
	for $p \in (-2, -1, 0, 0.5, 1, 2, 3)$	
e.g.	$\mu = \beta_0 + \beta_1 x^{0.5} + \beta_2 x^{-2}$	bfp(0.5,-2)
power polynomial	$\mu = \beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2}$	pp(x,2)
loess smoother		lo(x,3)
p-spline smoother		ps(x,3)
cubic spline smoother		cs(x,3)

gamlss
for statistical modelling

4. Modelling head circumference against age

Box-Cox *t* (BCT) distribution

parameter range	distribution shape
$\nu>1$ $\nu=1$ $\nu<1$	negatively skew symmetric positively skew
$\tau>0$ $\tau\to\infty$	leptokurtic mesokurtic

4.3 Choosing the distribution

criterion	penalty #	df_μ	df_σ	df_ν	df_τ	λ
AIC	2	20.4	4.7	3.0	8.8	0.09
GAIC	3	12.3	5.7	2	2	0.33
SBC	8.9	8.9	1	1	2	0.41

gamlss
for statistical modelling

4.3 Choosing the distribution

criterion	penalty #	dfμ	dfσ	dfν	dfτ	λ
AIC	2	20.4	4.7	3.0	8.8	0.09
GAIC	3	12.3	5.7	2	2	0.33
SBC	8.9	8.9	1	1	2	0.41

criterion	penalty	df_μ	df_σ	df_ν	df_τ	λ
AIC	$\sharp = 2$	20.4	4.7	3.0	8.8	0.09
GAIC	$\sharp = 3$	12.3	5.7	2	2	0.33
SBC	$\sharp = 8.9$	8.9	1	1	2	0.42

4.3 Choosing the distribution

Worm coefficients

group	age range	b0	b1	b2	b3
1	0 - 0.2	-1	4	2.8	-0.6
2	0.2 – 0.7	2	-3	-2.9	-0.5
3	0.7 – 1.2	1	-1	2.0	-1.4
4	1.2 – 1.8	-2	-2	3.1	**2.2**
5	1.8 – 2.5	-1	5	-2.9	0.1
6	2.5 – 3.9	4	**-6**	-1.9	-0.6
7	3.9 – 7.9	-1	2	-0.8	0.7
8	7.9 – 10.0	5	-1	-2.4	-0.2
9	10.0 – 11.2	1	4	-0.7	-0.6
10	11.2 – 12.5	-5	0	2.1	-1.3
11	12.5 – 13.7	**-8**	-1	2.4	-0.9
12	13.7 – 14.8	2	4	0.2	0.2
13	14.8 – 16.1	2	2	0.0	-1.0
14	16.1 – 17.4	5	4	-3.1	-0.1
15	17.4 – 18.8	2	0	**3.7**	-0.1
16	18.8 – 22.0	-4	-1	-1.3	-0.4

gamlss
for statistical modelling

Q statistics

(for BCT model chosen with penalty # = 3)

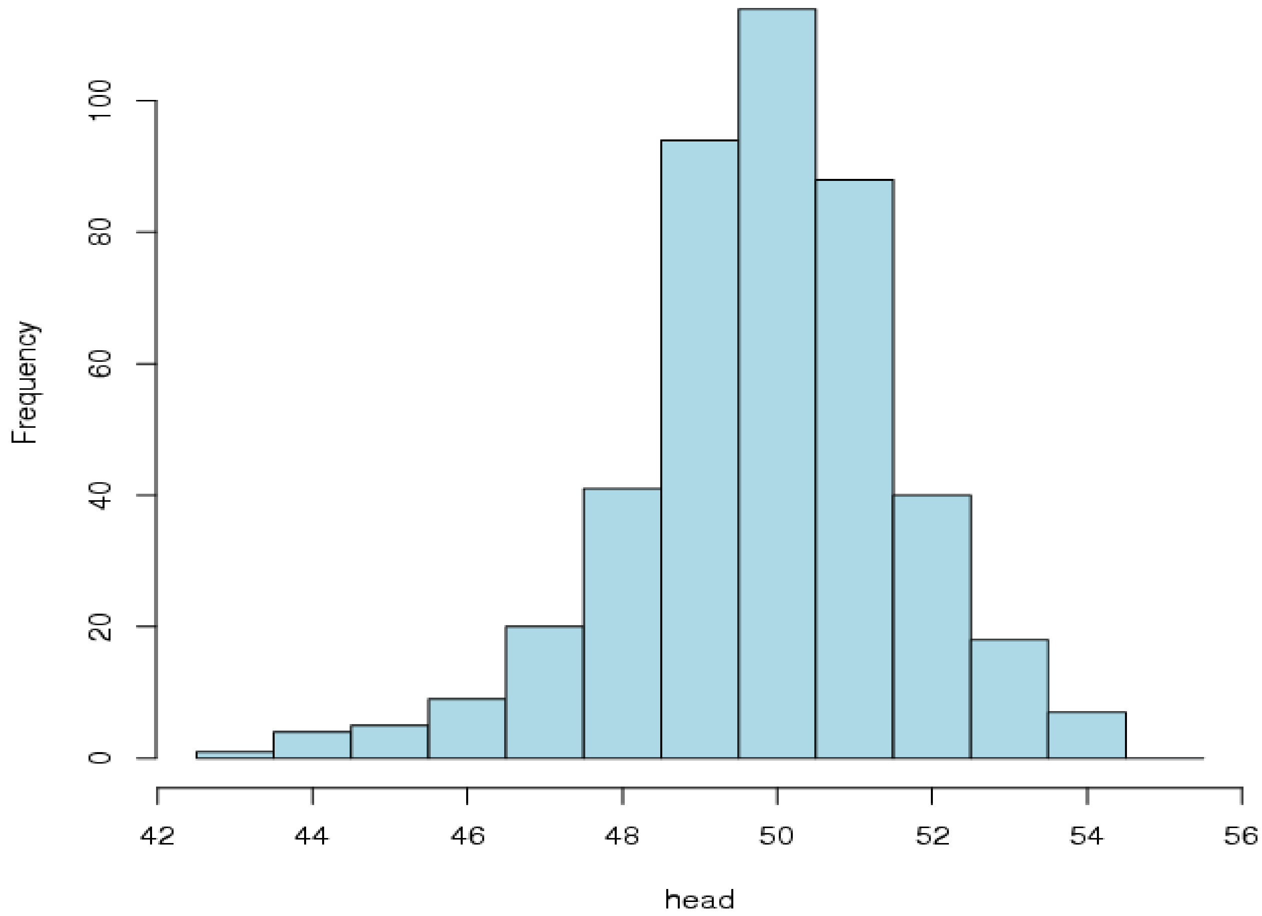
	Q1	Q2	Q3	Q4
value	7.2	23.3	23.9	16.7
df	3.7	12.7	14.0	14.0
p-value	0.10	0.03	0.05	0.28

Below with 7 extreme outliers removed

	Q1	Q2	Q3	Q4
value	7.2	19.7	17.3	4.2
df	3.7	12.7	14.0	14.0
p-value	0.10	0.10	0.24	0.994

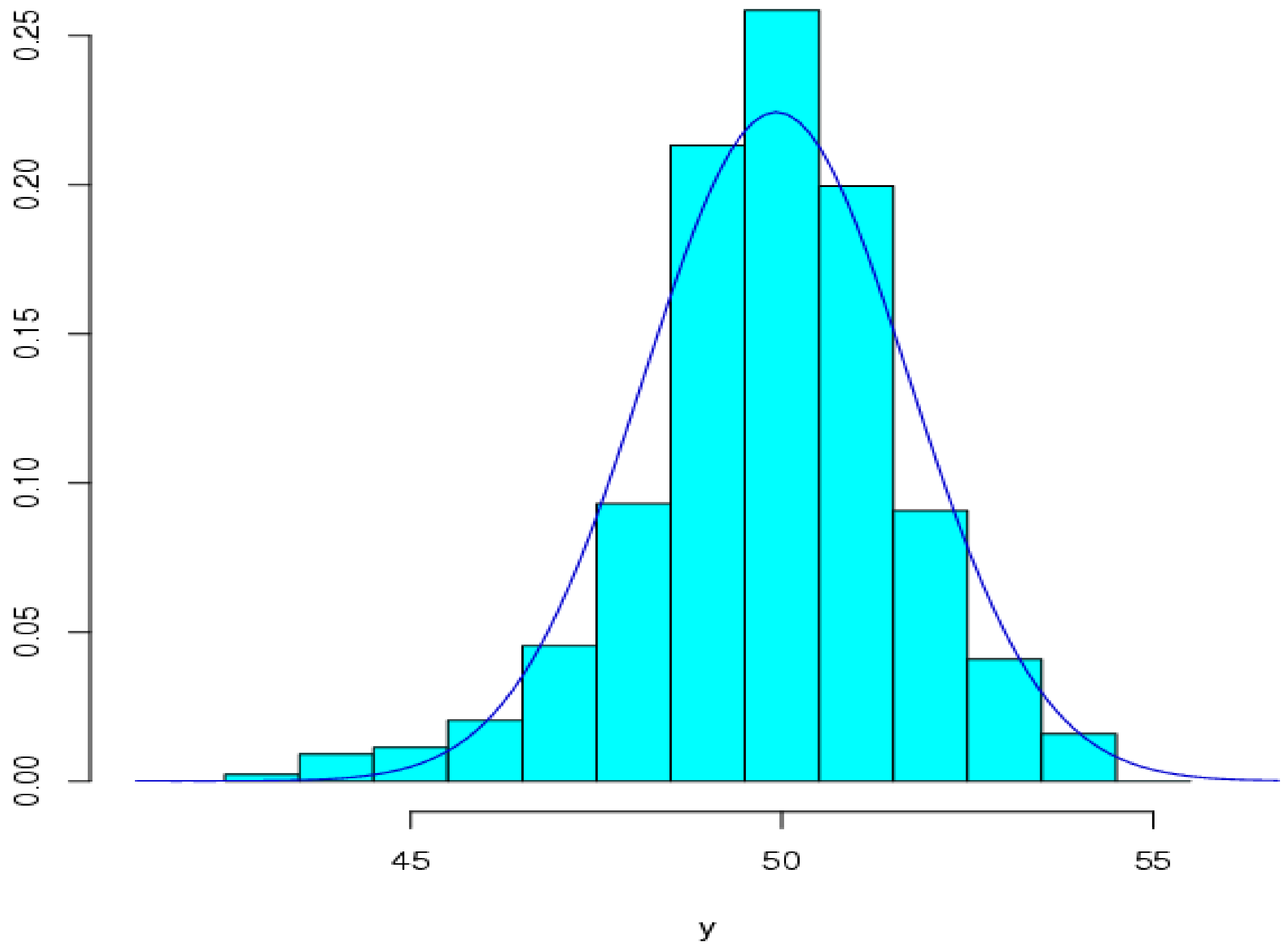
2.4 Example 1

Consider the values of head circumference for ages from 2 up to 3 years as a single sample



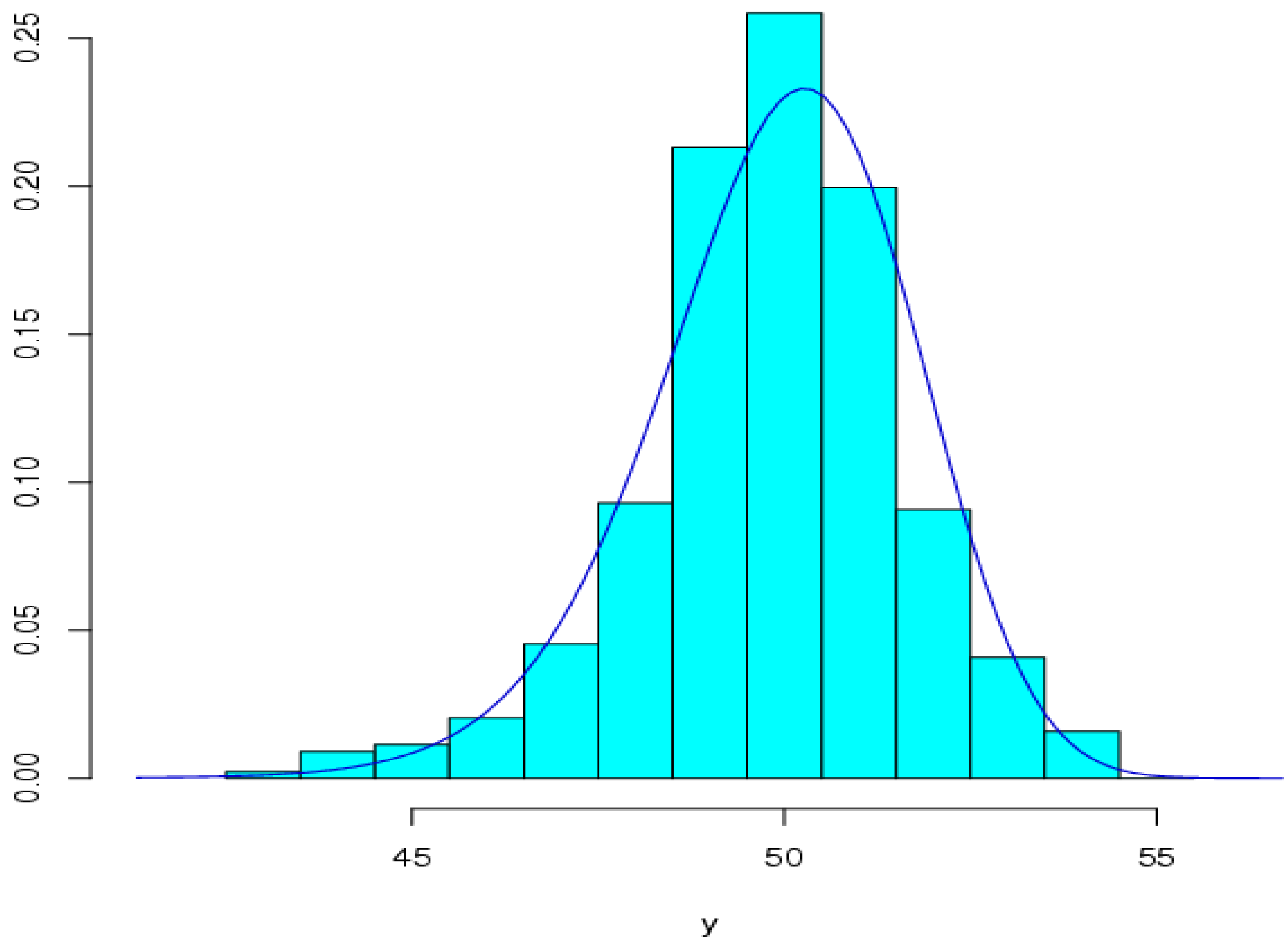
gamlss
for statistical modelling

Normal (N) fit to head circumference sample



gamlss
for statistical modelling

Box-Cox Normal (BCCG) fit to head sample



BCPE fit to head sample

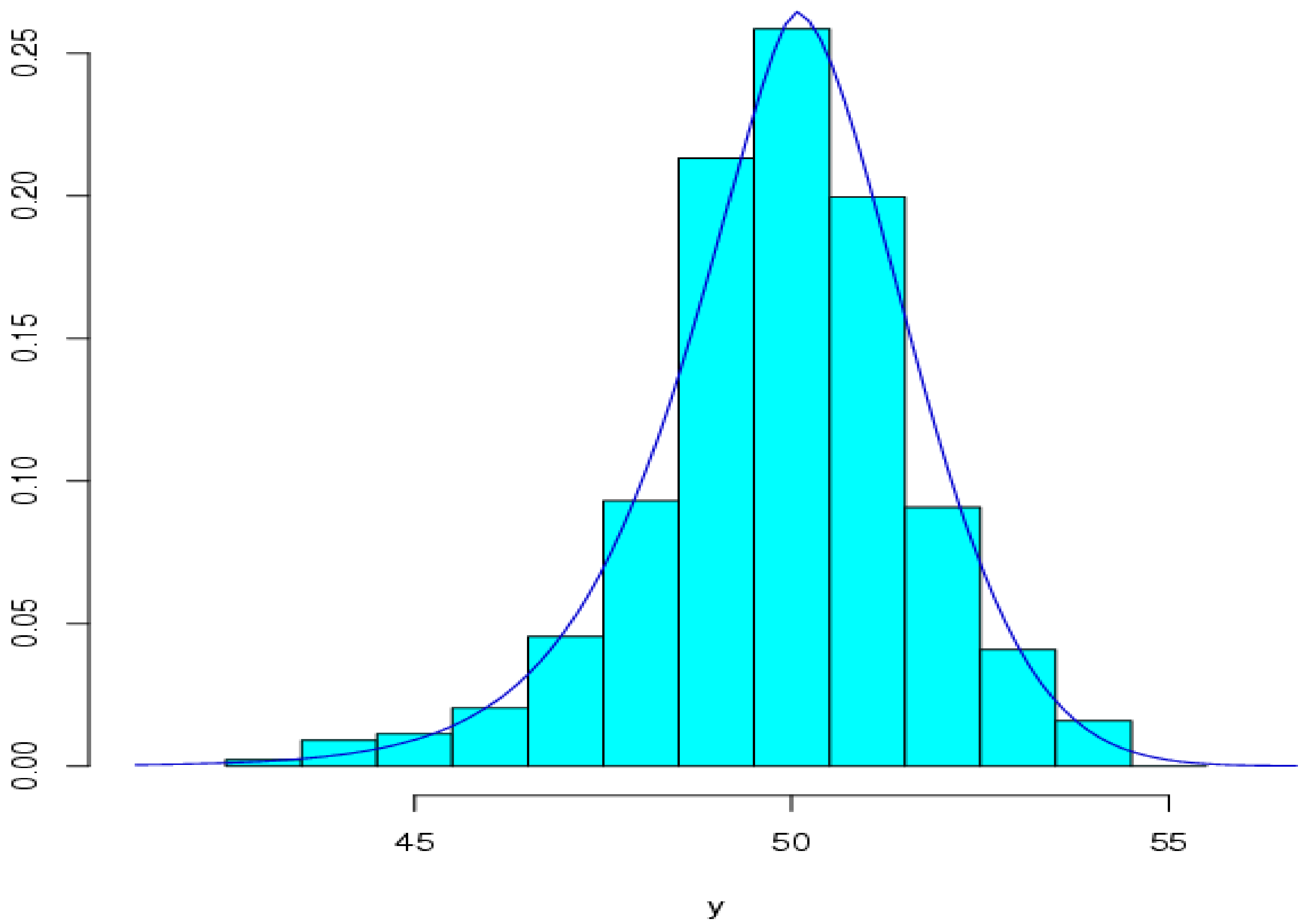


Table of deviances for different distribution fits

Distribution	Deviance – 1738.6
N	20.5
BCCG	5.0
JSU	1.7
BCT	1.6
BCPE	0

3. Modelling the parameters

Each parameter of the distribution μ, σ, ν, τ

is modelled in terms of the explanatory variable x

Model	Formula	gamlss code
simple linear	$\beta_0 + \beta_1 x$	x
polynomial	$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$	poly(x,3)
fractional polynomial	$\beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2}$	fp(x,2)
	for $p \in (-2, -1, 0, 0.5, 1, 2, 3)$	
e.g.	$\beta_0 + \beta_1 x^{0.5} + \beta_2 x^{-2}$	bfp(0.5,-2)
power polynomial	$\beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2}$	pp(x,2)
loess smoother		lo(x,3)
p-spline smoother		ps(x,3)
cubic spline smoother		cs(x,3)

3.1 GAMLSS model

$Y \sim D(\mu, \sigma, \nu, \tau)$ where D is any distribution and

$$g_1(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1})$$

$$g_2(\boldsymbol{\sigma}) = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} h_{j2}(\mathbf{x}_{j2})$$

$$g_3(\boldsymbol{\nu}) = \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} h_{j3}(\mathbf{x}_{j3})$$

$$g_4(\boldsymbol{\tau}) = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} h_{j4}(\mathbf{x}_{j4}).$$

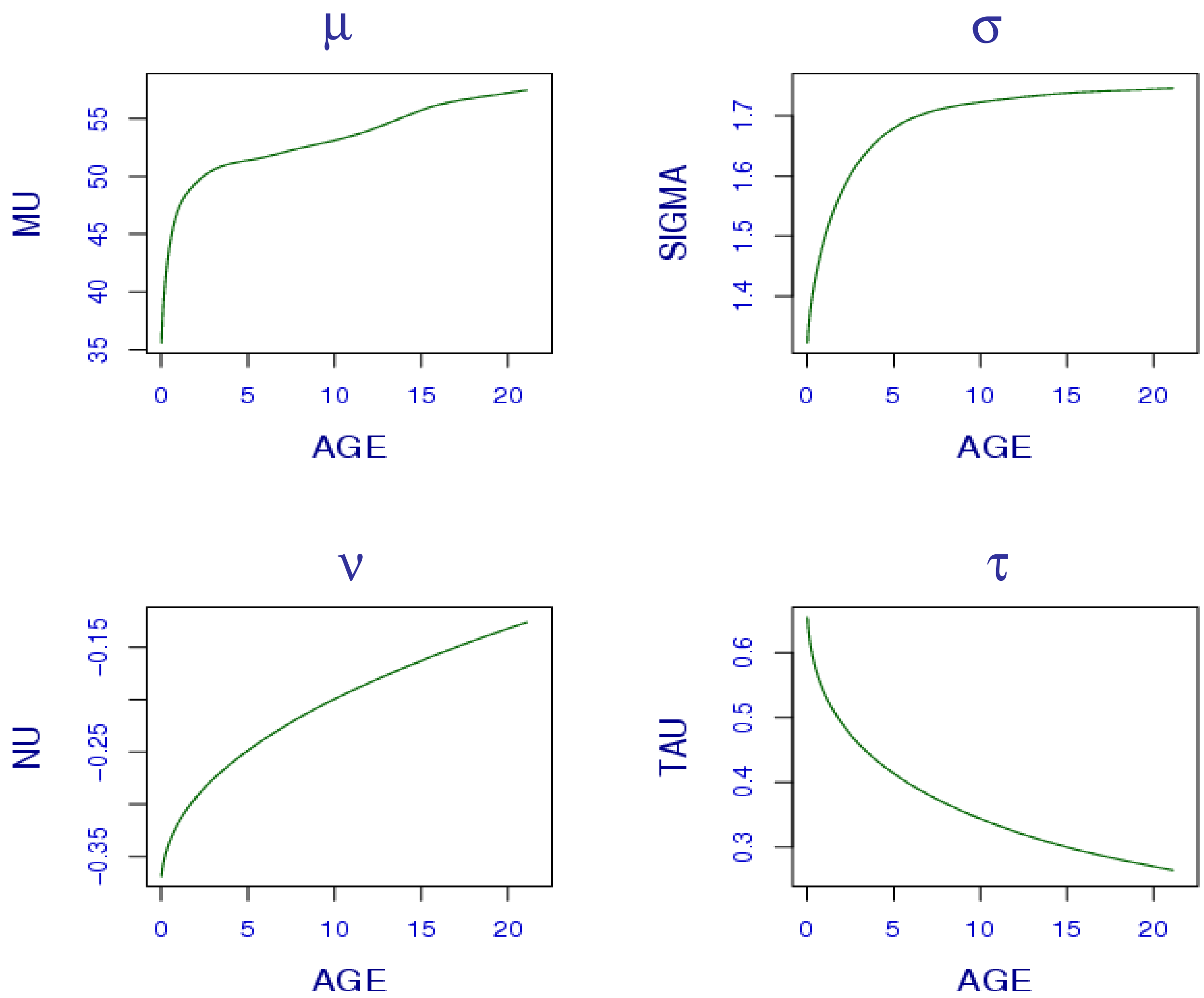
Johnson Su (JSU) distribution

parameter	distribution shape
μ	mean
σ	standard deviation
$\nu<0$ $\nu=0$ $\nu>0$	negatively skew symmetric positively skew
$\tau>0$ $\tau\to 0$	leptokurtic mesokurtic

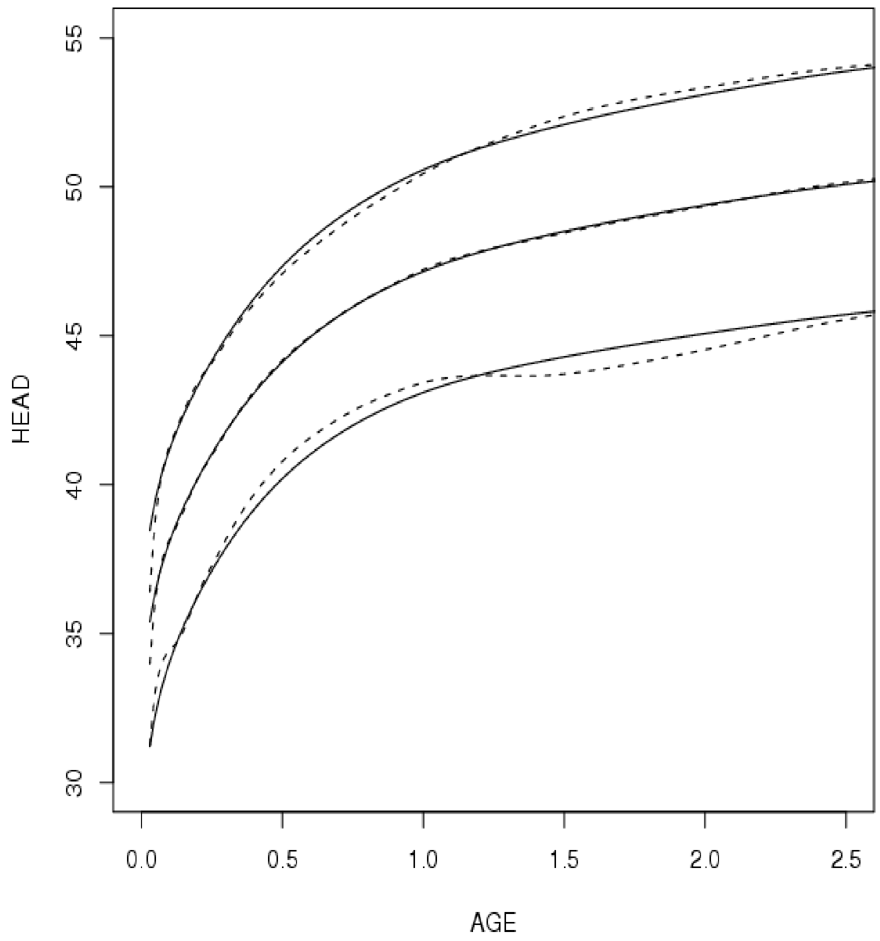
gamlss
for statistical modelling

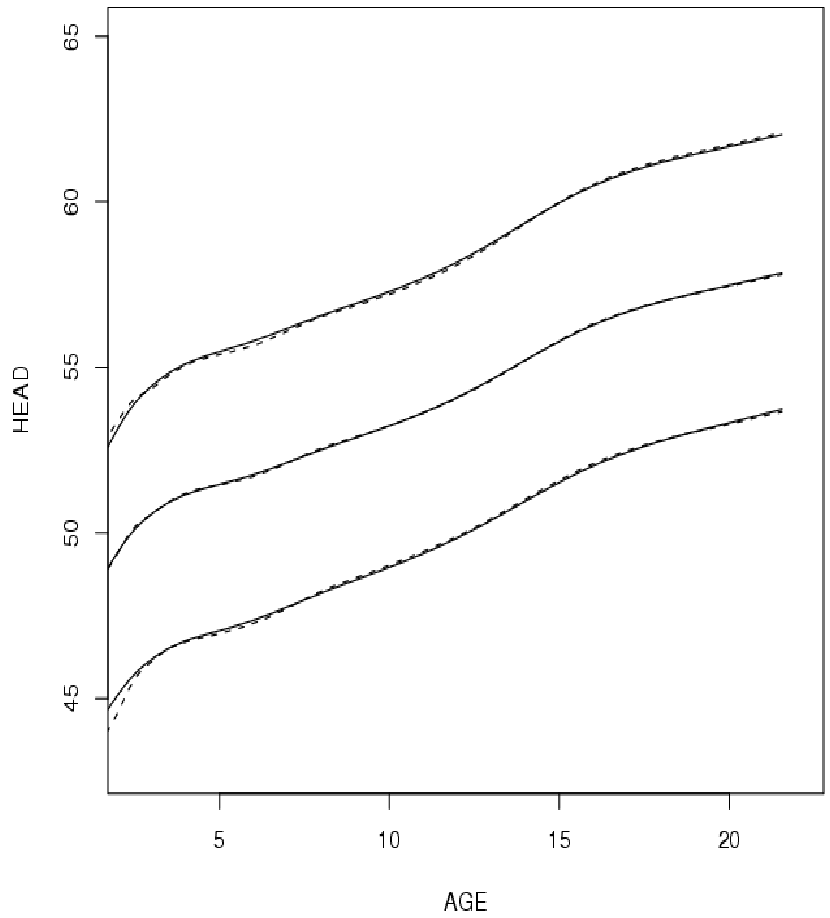
Fitted parameters μ, σ, υ, τ against AGE

(for JSU model chosen with penalty # = 3)

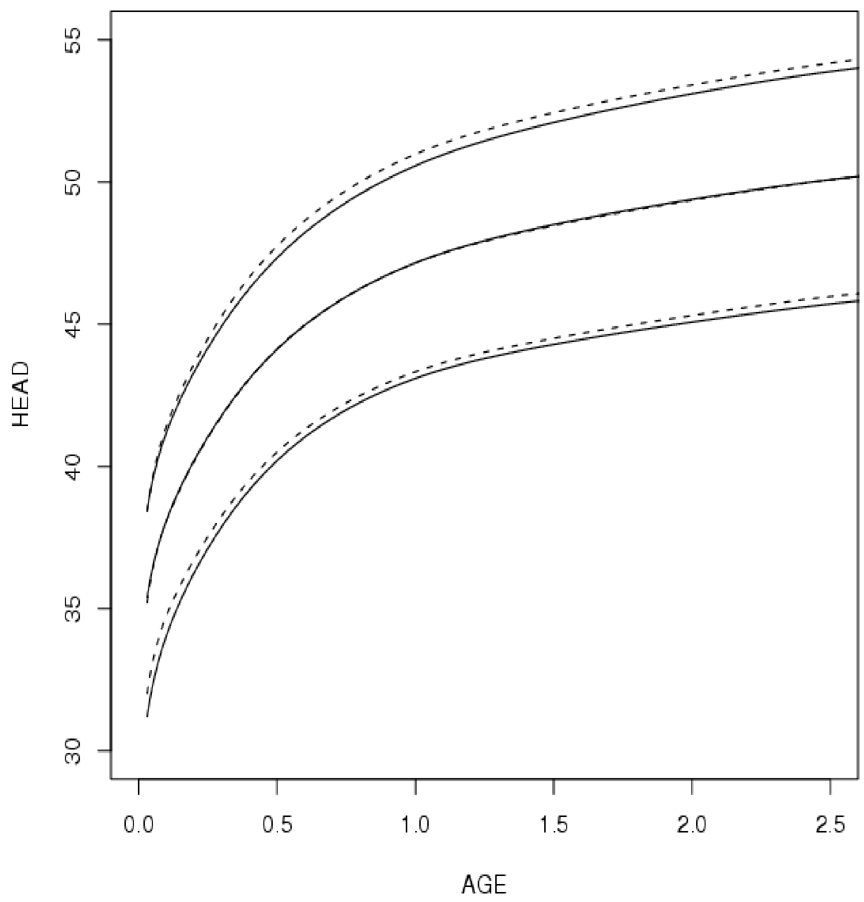


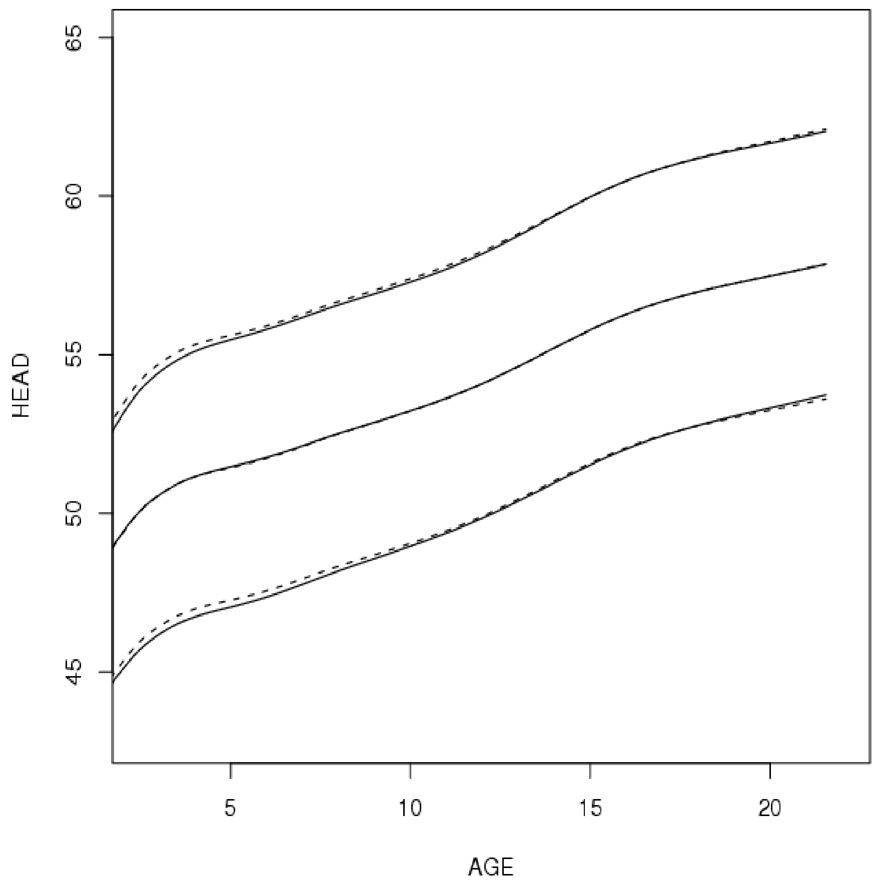
Comparison of centiles, (1, 50, 99)%, for BCT models with # = 2 (---), 3 (—)





Comparison of centiles , (1, 50, 99)%, BCT (—) and TF (---) for # = 3

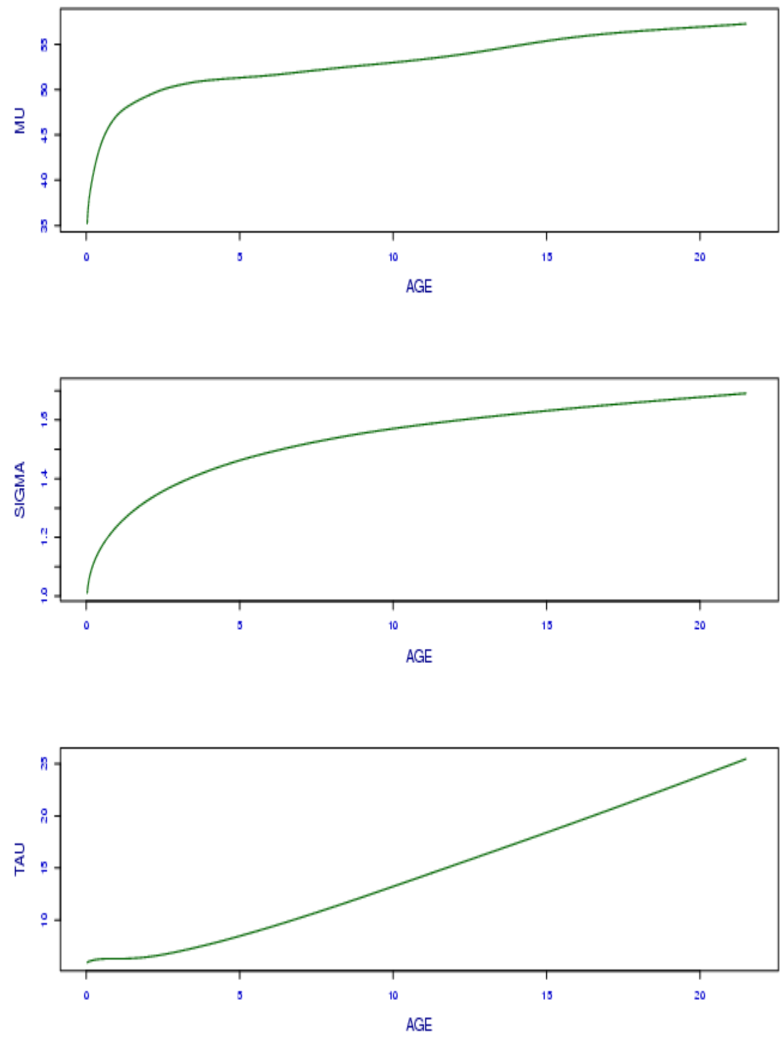




gamlss
for statistical modelling

Fitted parameters μ, σ, τ against AGE

(for TF model chosen with penalty # = 3)



5. Model diagnostics

5.1 Residuals (or 'z-scores')

The residuals from the fitted model are given by:

$$r = \Phi^{-1}\left[\hat{F}_Y(y)\right]$$

where Φ^{-1} is the inverse cdf of an N(0,1) variate, and

$\hat{F}_Y(y)$ is the fitted model cdf of Y,

The residuals from the model have an N(0,1) distribution,

whatever the original correct distribution of Y.

gamlss

for statistical modelling