

Applications of GAMLSS

Mikis Stasinopoulos¹ Bob Rigby¹

¹STORM, London Metropolitan University

28th Annual Conference of the International Society for Clinical Biostatistics 2007, Alexandroupolis

The talk

- What is GAMLSS

The talk

- What is GAMLSS
- GAMLSS: Two discrete data examples

The talk

- What is GAMLSS
- GAMLSS: Two discrete data examples
- GAMLSS: the R implementation

Generalized additive models for location scale and shape

- GAMLSS is a general framework for fitting regression type models

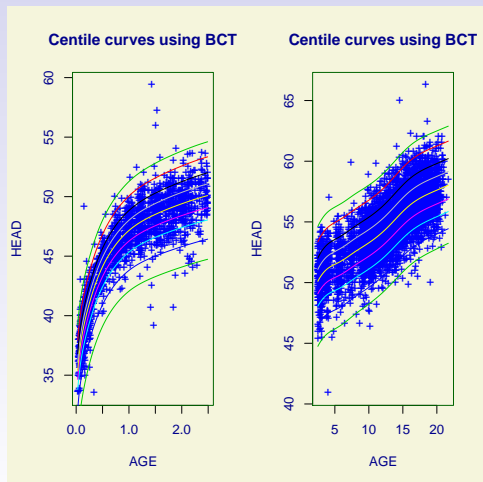
Generalized additive models for location scale and shape

- GAMLSS is a general framework for fitting regression type models
- The response variable $y \sim D(y|\mu, \sigma, \nu, \tau)$ where $D()$ can be any distribution (including highly skew and kurtotic continuous and discrete distributions). There are about 40 different distributions (truncated, censored and finite mixtures).

Generalized additive models for location scale and shape

- GAMLSS is a general framework for fitting regression type models
- The response variable $y \sim D(y|\mu, \sigma, \nu, \tau)$ where $D()$ can be any distribution (including highly skew and kurtotic continuous and discrete distributions). There are about 40 different distributions (truncated, censored and finite mixtures).
- All the parameters of the distribution can be modelled as linear/non-linear parametric functions and/or smoothing functions of the explanatory variables (i.e. cubic splines, penalized splines, loess) and/or random effects.

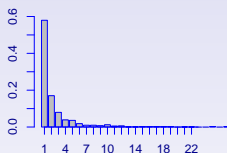
GAMLSS for centile estimation



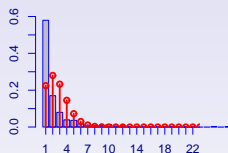
Examples: The stylometric data, Chappas and Corina-Borja (2006)

Number of times a word appears in the text

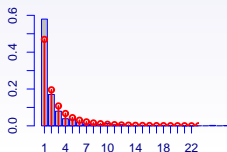
(a) the data



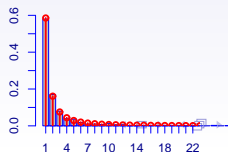
(b) Poisson



(c) negative binomial

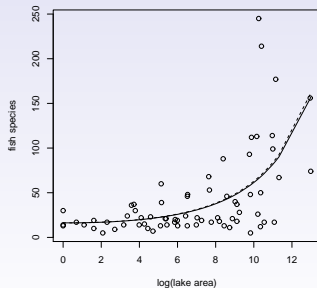


(d) Sichel



Examples: The fish species data, Stein and Juritz (1988)

y: the number of fish species, x: the area of the lake



The fish species data

There are several questions that need to be answered.

- How does the mean of y depend on x ?

The fish species data

There are several questions that need to be answered.

- How does the mean of y depend on x ?
- Is y overdispersed Poisson?

The fish species data

There are several questions that need to be answered.

- How does the mean of y depend on x ?
- Is y overdispersed Poisson?
- How does the variance y depend on its mean?

The fish species data

There are several questions that need to be answered.

- How does the mean of y depend on x ?
- Is y overdispersed Poisson?
- How does the variance y depend on its mean?
- What is the distribution of y given x ?

The fish species data

There are several questions that need to be answered.

- How does the mean of y depend on x ?
- Is y overdispersed Poisson?
- How does the variance y depend on its mean?
- What is the distribution of y given x ?
- Do the scale and shape parameters of the distribution of y depend on x ?

Different (overdispersed) count data approaches

(i) *Ad-hoc* solutions

Different (overdispersed) count data approaches

(i) *Ad-hoc* solutions

(a) quasi-likelihood (QL), Extended QL

Different (overdispersed) count data approaches

(i) *Ad-hoc* solutions

- (a) quasi-likelihood (QL), Extended QL
- (b) Efron's Double Exponential

Different (overdispersed) count data approaches

(i) *Ad-hoc* solutions

- (a) quasi-likelihood (QL), Extended QL
- (b) Efron's Double Exponential
- (c) pseudo-likelihood (PL)

Different (overdispersed) count data approaches

(i) *Ad-hoc* solutions

- (a) quasi-likelihood (QL), Extended QL
- (b) Efron's Double Exponential
- (c) pseudo-likelihood (PL)

(ii) Discretized continuous distributions

for example if $F_W(w)$ is the cdf a continuous random variable W defined in \mathfrak{R}^+ then $f_Y(y) = F_W(y + 1) - F_W(y)$

Different (overdispersed) count data approaches

(i) *Ad-hoc* solutions

- (a) quasi-likelihood (QL), Extended QL
- (b) Efron's Double Exponential
- (c) pseudo-likelihood (PL)

(ii) Discretized continuous distributions

for example if $F_W(w)$ is the cdf a continuous random variable W defined in \mathfrak{R}^+ then $f_Y(y) = F_W(y + 1) - F_W(y)$

(iii) Random effect at the observation level solutions.

$$f_Y(y) = \int f(y|\gamma)f_\gamma(\gamma)d\gamma.$$

Random effect at the observation level

- (a) when an explicit continuous mixture distribution, $f_Y(y)$, exists.

Random effect at the observation level

- (a) when an explicit continuous mixture distribution, $f_Y(y)$, exists.
- (b) when a continuous mixture distribution, $f_Y(y)$, is not explicit but is approximated by integrating out the random effect using approximations, e.g. Gaussian quadrature or Laplace approximation.

Random effect at the observation level

- (a) when an explicit continuous mixture distribution, $f_Y(y)$, exists.
- (b) when a continuous mixture distribution, $f_Y(y)$, is not explicit but is approximated by integrating out the random effect using approximations, e.g. Gaussian quadrature or Laplace approximation.
- (c) when a 'non-parametric' mixture (effectively a finite mixture) is assumed for the response variable.

Table: Some overdispersed count data distributions $f_Y(y)$ and their mixing distributions $f_\gamma(\gamma)$ where $f(y|\gamma)$ has a Poisson distribution.

$f_Y(y)$: marginal	$f_\gamma(\gamma)$: mixing distribution
Negative binomial type I	Gamma
Poisson-inverse Gaussian	inverse Gaussian
Sichel	generalized inverse Gaussian
Delaporte	shifted gamma
Poisson-Tweedie	Tweedie family
Zero inflated Poisson	binary

Families modelling the variance-mean relationship

$V[Y] = \mu + \mu^2 V[\gamma]$ where $V[\gamma] = v(\sigma, \nu, \tau)$ is a function of the parameters of the mixing distribution $f_\gamma(\gamma)$.

Alternative variance-mean relationship can be obtained by reparametrization.

i.e NB type I $V[Y] = \mu + \sigma\mu^2$.

If $\sigma = \sigma_1/\mu$ then

$V[Y] = (1 + \sigma_1)\mu$ (negative binomial type II)

$\sigma = \sigma_1\mu$ then $V[Y] = \mu + \sigma_1\mu^3$.

More generally $\sigma = \sigma_1\mu^{2-\nu}$ giving $V(Y) = \mu + \sigma_1\mu^\nu$

Overdispersed count data approaches

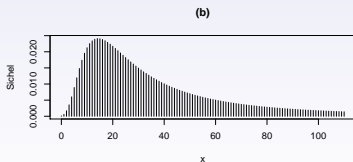
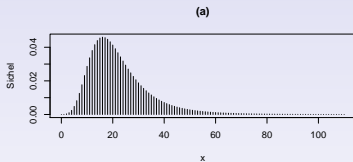
Table: Comparison of models for the fish species data

Model	$f_Y(y)$	μ	σ	ν	DEV	df	AIC	SBC
1	PO	$x < 2 >$	-	-	1849.3	3	1855.3	1862.0
2	NBI	x	1	-	619.8	3	625.8	632.6
3	NBI	$x < 2 >$	1	-	614.3	4	622.3	631.3
4	NBI	$cs(x, 3)$	1	-	611.9	6	623.9	637.4
5	NBI	$x < 2 >$	x	-	605.0	5	615.0	626.2
6	NBI-fam	$x < 2 >$	1	1	606.0	5	616.0	627.3
7	NBI-fam	$x < 2 >$	x	1	604.9	6	616.9	630.4

Overdispersed count data approaches

Model	$f_Y(y)$	μ	σ	ν	DEV	df	AIC	SBC
8	PIG	$x < 2 >$	1	-	613.3	4	621.3	630.3
9	SI	$x < 2 >$	1	x	597.7	6	609.7	623.2
10	DEL	$x < 2 >$	1	x	600.6	6	612.6	626.1
11	DEL	$x < 2 >$	-	x	600.6	5	610.6	621.9
12	PO-Normal	$x < 2 >$	1	-	615.2	4	623.2	632.2
13	NBI-Normal	$x < 2 >$	x	1	603.7	6	615.7	629.2
14	PO-NPFM(5)	$x < 2 >$	-	-	601.9	13	627.9	657.2
15	NB-NPFM(2)	$x < 2 >$	1	-	611.9	6	623.9	637.4
16	doublePO	$x < 2 >$	x	-	616.4	5	626.4	637.6
17	IGdisc	$x < 2 >$	1	-	603.3	4	611.3	620.3

Fitted Sichel distributions for observations (a) 40 and (b) 67



The GAMLSS software implementation in R

CRAN: six different packages:

- 1 the original `gamlss` package for fitting GAMLSS

The GAMLSS software implementation in R

CRAN: six different packages:

- 1 the original `gamlss` package for fitting GAMLSS
- 2 the `gamlss.cens` package for fitting censored (interval) response variables.

The GAMLSS software implementation in R

CRAN: six different packages:

- 1 the original `gamlss` package for fitting GAMLSS
- 2 the `gamlss.cens` package for fitting censored (interval) response variables.
- 3 the `gamlss.dist` package for additional new distributions

The GAMLSS software implementation in R

CRAN: six different packages:

- 1 the original `gamlss` package for fitting GAMLSS
- 2 the `gamlss.cens` package for fitting censored (interval) response variables.
- 3 the `gamlss.dist` package for additional new distributions
- 4 the `gamlss.mx` package for fitting finite mixture distributions.

The GAMLSS software implementation in R

CRAN: six different packages:

- 1 the original `gamlss` package for fitting GAMLSS
- 2 the `gamlss.cens` package for fitting censored (interval) response variables.
- 3 the `gamlss.dist` package for additional new distributions
- 4 the `gamlss.mx` package for fitting finite mixture distributions.
- 5 the `gamlss.nl` package for fitting nonlinear models

The GAMLSS software implementation in R

CRAN: six different packages:

- 1 the original `gamlss` package for fitting GAMLSS
- 2 the `gamlss.cens` package for fitting censored (interval) response variables.
- 3 the `gamlss.dist` package for additional new distributions
- 4 the `gamlss.mx` package for fitting finite mixture distributions.
- 5 the `gamlss.nl` package for fitting nonlinear models
- 6 the `gamlss.tr` package for fitting truncated distributions.

References

- 1 Rigby RA, Stasinopoulos DM (2004). Smooth centile curves for skew and kurtotic data modelled using the Box-Cox Power Exponential distribution. *Statistics in Medicine*, **23**, 3053-3076.

References

- 1 Rigby RA, Stasinopoulos DM (2004). Smooth centile curves for skew and kurtotic data modelled using the Box-Cox Power Exponential distribution. *Statistics in Medicine*, **23**, 3053-3076.
- 2 Rigby RA, Stasinopoulos DM (2005). Generalized additive models for location, scale and shape, (with discussion). *Appl. Statist.*, **54**, 507-554.

References

- 1 Rigby RA, Stasinopoulos DM (2004). Smooth centile curves for skew and kurtotic data modelled using the Box-Cox Power Exponential distribution. *Statistics in Medicine*, **23**, 3053-3076.
- 2 Rigby RA, Stasinopoulos DM (2005). Generalized additive models for location, scale and shape, (with discussion). *Appl. Statist.*, **54**, 507-554.
- 3 Rigby RA, Stasinopoulos DM (2006). Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling*, **6**, 209-229.

References

- 1 Rigby RA, Stasinopoulos DM (2004). Smooth centile curves for skew and kurtotic data modelled using the Box-Cox Power Exponential distribution. *Statistics in Medicine*, **23**, 3053-3076.
- 2 Rigby RA, Stasinopoulos DM (2005). Generalized additive models for location, scale and shape, (with discussion). *Appl. Statist.*, **54**, 507-554.
- 3 Rigby RA, Stasinopoulos DM (2006). Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling*, **6**, 209-229.
- 4 Stasinopoulos D. M., Rigby R.A. and Akantziliotou C. (2006) Instructions on how to use the GAMLSS package in R. (see also <http://www.londonmet.ac.uk/gamlss/>).