Flexible Regression and Smoothing

Discrete Distributions

Bob Rigby   Mikis Stasinopoulos

Graz University of Technology, Austria, November 2016

gamlss

Count distributions

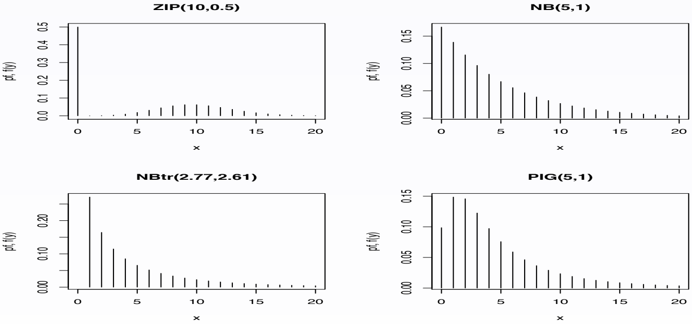
The three major problems encounter when modelling count data using the Poisson distribution.

- overdispersion
- excess (or shortage) of zero values
- long tails (rare events)

Discrete distribution modelling

Par.	Modelling	Distributions
1	Location	PO
2	Location and scale	NBI, NBII, PIG
2	Location and zero probability	ZALG, ZAP, ZIP, ZIP2
3	Location, scale and skewness	DEL, SI, SICHEL
3	Location, scale and zero probability	ZANBI, ZINBI, ZIPIG

Different count data distributions



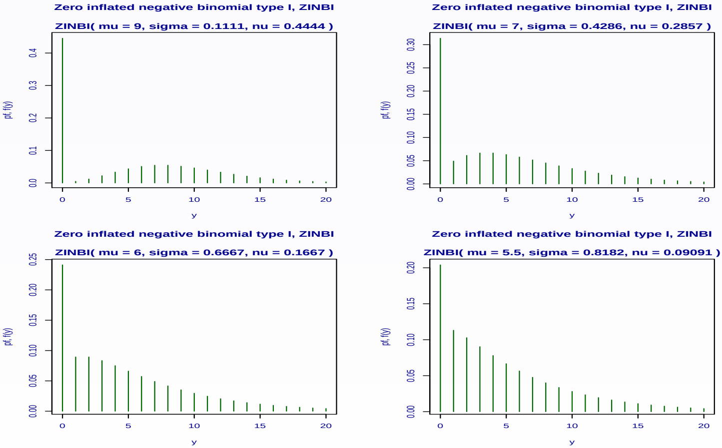
Zero inflated distributions

Zero inflated distribution, $Y \sim \mathbf{ZID}$ is given by
$Y = 0$ with probability p
$Y \sim \mathbf{D}$ with probability $1 - p$.
Hence

$$P(Y = y) = \begin{cases} p + (1-p)P(Y_1 = 0) & \text{if } y = 0 \\ (1-p)P(Y_1 = y) & \text{if } y = 1, 2, 3, \ldots \end{cases}$$

where $Y_1 \sim \mathbf{D}$.

ZINBI distribution plots



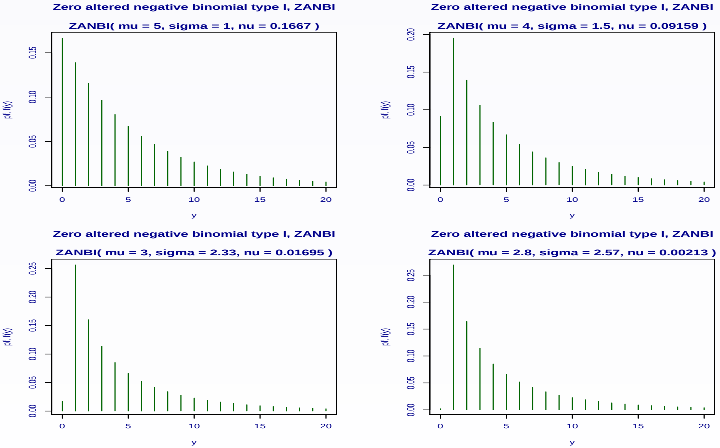
Zero adjusted distributions

Zero adjusted distribution, $Y \sim \textbf{ZAD}$ is given by
$Y = 0$ with probability p
$Y \sim \textbf{Dtr}$ with probability $1 - p$,
where **Dtr** is a truncated distribution, **D** truncated at zero.
Hence

$$P(Y = y) = \begin{cases} p & \text{if } y = 0 \\ (1-p)\frac{P(Y_1 = y)}{1 - P(Y_1 = 0)} & \text{if } y = 1, 2, 3, \ldots \end{cases} \tag{1}$$

where $Y_1 \sim \textbf{D}$.

ZANBI distribution plots



Different (overdispersed) count data approaches

(a) *Ad-hoc* solutions
  (i) quasi-likelihood (QL), Extended QL
  (ii) Efron's Double Exponential
  (iii) pseudo-likelihood (PL)

(b) Discretized continuous distributions
for example if $F_W(w)$ is the cdf a continuous random variable W defined in $\Re^+$ then $f_Y(y) = F_W(y+1) - F_W(y)$

(c) Random effect at the observation level solutions.
$f_Y(y) = \int f(y|\gamma) f_\gamma(\gamma) d\gamma$.

(c) Random effect at the observation level

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

(ii) 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.

(iii) when a 'non-parametric' mixture (effectively a finite mixture) is assumed for the response variable.

gamlss

Random effect at the observation level case (i)

(i) Explicit continuous mixture distribution

$$\underbrace{f_Y(y)}_{\text{discrete}} = \int \underbrace{f(y|\gamma)}_{\text{discrete}} \underbrace{f_\gamma(\gamma)}_{\text{continuous}} \, d\gamma$$

- $Y \sim NBI(\mu, \sigma)$
- $Y|\gamma \sim PO(\gamma\mu)$
- $\gamma \sim GA(1, \sigma^{1/2})$

gamlss

Random effect at the observation level case (ii)

(ii) Non-explicit continuous mixture distribution

$$\underbrace{f_Y(y)}_{\text{discrete}} = \int \underbrace{f(y|\gamma)}_{\text{discrete}} \underbrace{f_\gamma(\gamma)}_{\text{continuous}} \, d\gamma$$

- $Y \sim PO - Normal(\mu, \sigma)$
- $Y|\gamma \sim PO(\gamma\mu)$
- $\log(\gamma) \sim NO(1, \sigma)$

Random effect at the observation level case (iii)

(iii) Non-parametric mixture distribution

$$\underbrace{f_Y(y)}_{\text{discrete}} = \sum_{k=1}^{K} \underbrace{f(y|\gamma_k)}_{\text{discrete}} \underbrace{p(\gamma = \gamma_k)}_{\text{continuous}}$$

- $Y \sim PO - NPFM(\mu, \sigma)$
- $Y|\gamma \sim PO(\gamma\mu)$
- $\log(\gamma) \sim NPFM(2)$

where NPFM(2) equals Non-Parametric Finite Mixture with 2 point probabilities

gamlss

Explicit continuous mixture distribution

Distributions	R Name	mixing distribution for γ
Poisson	$\texttt{PO}(\mu)$	-
Neg. bin. I	$\texttt{NBI}(\mu, \sigma)$	$\texttt{GA}(1, \sigma^{\frac{1}{2}})$
Neg. bin. II	$\texttt{NBII}(\mu, \sigma)$	$\texttt{GA}(1, \sigma^{\frac{1}{2}}/\mu)$
Poisson IG	$\texttt{PIG}(\mu, \sigma)$	$\texttt{IG}(1, \sigma^{\frac{1}{2}})$
Sichel	$\texttt{SICHEL}(\mu, \sigma, \nu)$	$\texttt{GIG}(1, \sigma^{\frac{1}{2}}, \nu)$
Delaporte	$\texttt{DEL}(\mu, \sigma, \nu)$	$\texttt{SG}(1, \sigma^{\frac{1}{2}}, \nu)$
Zero inflated Poisson	$\texttt{ZIP}(\mu, \sigma)$	$\texttt{BI}(1, 1-\sigma)$
Zero inflated Poisson 2	$\texttt{ZIP2}(\mu, \sigma)$	$(1-\sigma)^{-1}\texttt{BI}(1, 1-\sigma)$
Zero inflated neg. bin.	$\texttt{ZINBI}(\mu, \sigma, \nu)$	zero inflated gamma
Poisson-Tweedie	-	Tweedie family

Table: Discrete gamlss family distributions for count data

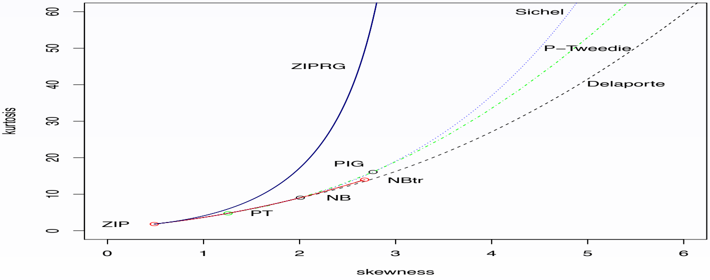
R Name	params	mean	variance
$\texttt{PO}(\mu)$	1	μ	μ
$\texttt{NBI}(\mu, \sigma)$	2	μ	$\mu + \sigma\mu^2$
$\texttt{NBII}(\mu, \sigma)$	2	μ	$\mu + \sigma\mu$
$\texttt{PIG}(\mu, \sigma)$	2	μ	$\mu + \sigma\mu^2$
$\texttt{SICHEL}(\mu, \sigma, \nu)$	3	μ	$\mu + h(\sigma, \nu)\mu^2$
$\texttt{DEL}(\mu, \sigma, \nu)$	3	μ	$\mu + \sigma(1-\nu)^2\mu^2$
$\texttt{ZIP}(\mu, \sigma)$	2	$(1-\sigma)\mu$	$(1-\sigma)\mu + \sigma(1-\sigma)\mu^2$
$\texttt{ZIP2}(\mu, \sigma)$	2	μ	$\mu + \frac{\sigma}{(1-\sigma)}\mu^2$

Families modelling the variance-mean relationship

$V[Y] = \mu + \mu^2 V[\gamma]$ where $V[\gamma] = \upsilon(\sigma, \nu, \tau)$ 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 $V(Y) = \mu + \sigma_1\mu^\nu$

gamlss

Comparison of the marginal distributions using a (ratio moment) diagram of their skewness and kurtosis



A stylometric application

Data summary:

R data file: `stylo` in package **gamlss.data** of dimensions 64×2

source: Dr Mario Corina-Borja

variables

`word` : is the number of times a word appears in a single text

`freq` : the frequency of the number of times a word appears in a text

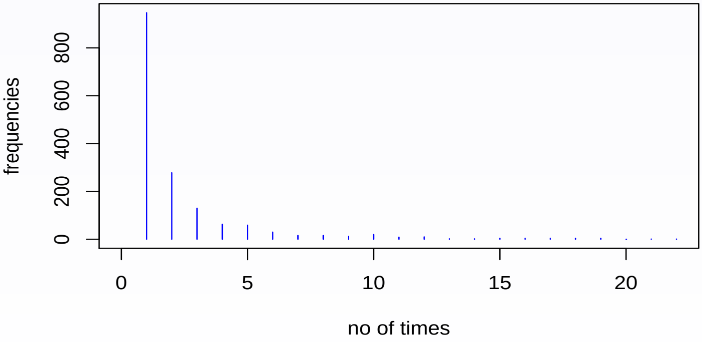
purpose: to demonstrate the fitting of a truncated discrete dist.

conclusion the truncated SICHEL distributions fits best

A stylometric application

```
 library(gamlss.tr)
 data(stylo)
 plot(freq ~ word, data = stylo, type = "h", xlim =
+  c(0, 22), xlab = "no of times", ylab =
+ "frequencies", col = "blue")
```

The stylometric data



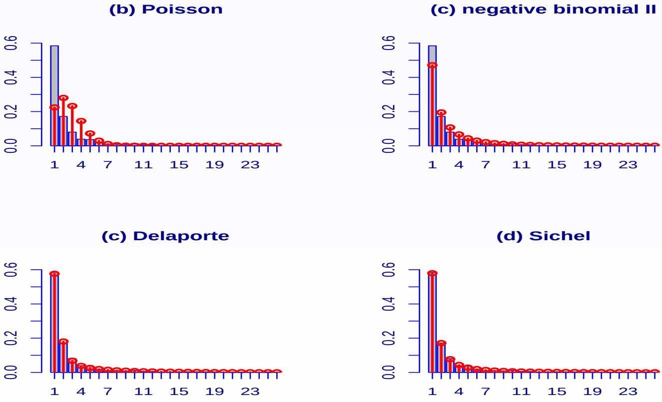
A stylometric application

```
> library(gamlss.tr)
> gen.trun(par = 0, family = PO, type = "left")
A truncated family of distributions from PO has been generated
 and saved under the names:
 dPOtr pPOtr qPOtr rPOtr POtr
The type of truncation is left and the truncation parameter is
> gen.trun(par = 0, family = NBII, type = "left")
...
> gen.trun(par = 0, family = DEL, type = "left")
...
> gen.trun(par = 0, family = SICHEL, type = "left",
+       delta = 0.001)
...
```

A stylometric application

```
> mPO <- gamlss(word ~ 1, weights = freq, data = stylo,
+ family = POtr, trace = FALSE)
> mNBII <- gamlss(word ~ 1, weights = freq, data = stylo,
+ family = NBIItr, n.cyc = 50, trace = FALSE)
> mDEL <- gamlss(word ~ 1, weights = freq, data = stylo,
+ family = DELtr, n.cyc = 50, trace = FALSE)
> mSI <- gamlss(word ~ 1, weights = freq, data = stylo,
+ family = SICHELtr, n.cyc = 50, trace = FALSE)
> GAIC(mPO, mNBII, mDEL, mSI)
      df      AIC
mSI    3 5148.454
mDEL   3 5160.581
mNBII  2 5311.627
mPO    1 9207.459
```

The stylometric data



The fish species data

Data summary: the fish species data

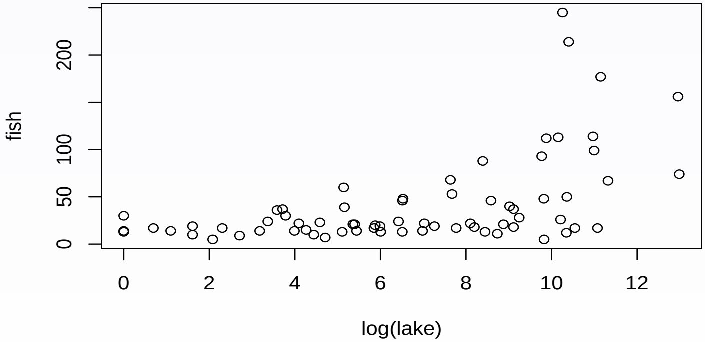
R data file: `species` in package gamlss.data of dimensions 70×2

variables

`fish` : the number of different species in 70 lakes in the world

`lake` : the lake area

The fish species data



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?

Overdispersed count data approaches

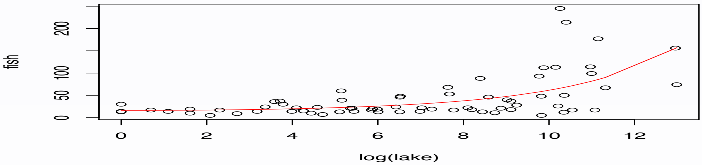
Table: Comparison of models for the fish species data

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

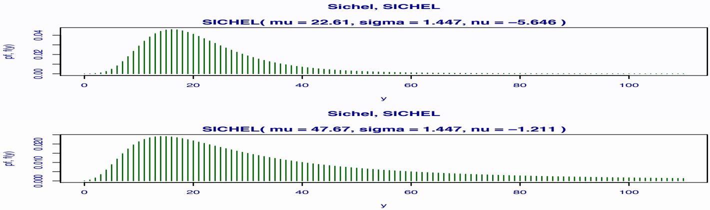
Overdispersed count data approaches

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 mean of the Sichel distribution



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



Binomial response variables

There are only two distributions here

- binomial
- beta binomial

The hospital stay data

Data summary:

R data file: aep in package **gamlss** of dimensions 1383×8

source: Gange *et al.* (1996)

variables

`los` : total number of days

`noinap` : number of inappropriate days patient stay in hospital

`loglos` : the log of `los/10`

`sex` : the gender of patient

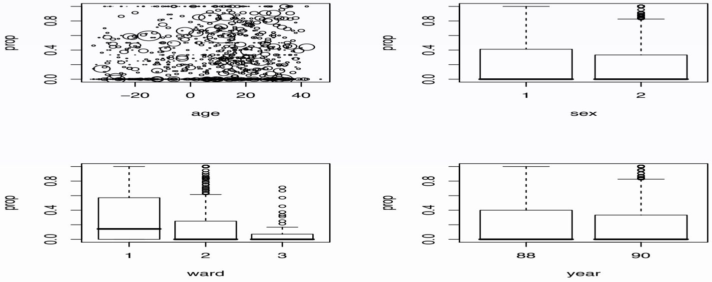
`ward` : type of ward in the hospital (medical, surgical or

`year` : 1988 or 1990

`age` : age of the patient subtracted from 55

`y` : the response variable, a matrix with columns (`noinap`, `los-noinap`)

The hospital stay data



The hospital stay data

```
> mI <- gamlss(y ~ ward + year + loglos, sigma.fo = ~year,
+ family = BB,data = aep)
> mII <- gamlss(y ~ ward + year + loglos, sigma.fo = ~year +
+ ward, family = BB, data = aep)
> mIII <- gamlss(y ~ ward + year + cs(loglos, 1),
+ sigma.fo = ~year+ward, family = BB, data = aep)
> mIV <- gamlss(y ~ ward + year + cs(loglos, 1) + cs(age, 1),
+ sigma.fo = ~year + ward, family = BB, data = aep)
> GAIC(mI, mII, mIII, mIV, k = 0)
           df      AIC
mIV  12.00010 4454.362
mIII 10.00045 4459.427
mII   9.00000 4483.020
mI    7.00000 4519.441
```

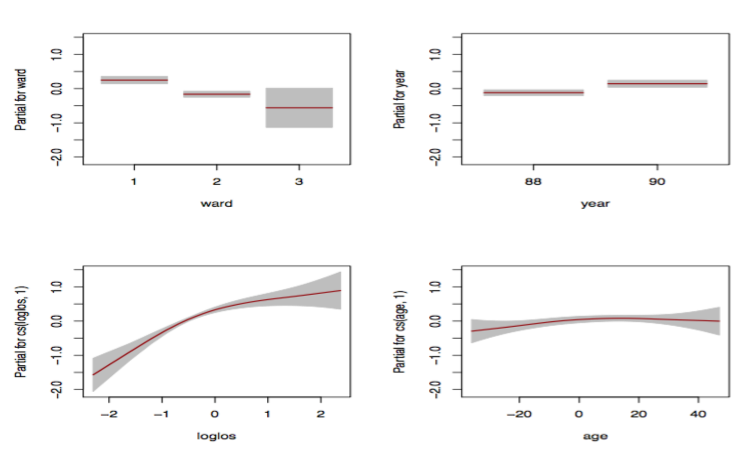
The hospital stay data

Models	Links	Terms	GD (AIC) [SBC]
I	$\text{logit}(\mu)$ $\log(\sigma)$	1+ward+loglos+year 1+year	4519.4 (4533.4) [4570.1]
II	$\text{logit}(\mu)$ $\log(\sigma)$	1+ward+loglos+year 1+year+ward	4483.0 (4501.0) [4548.1]
III	$\text{logit}(\mu)$ $\log(\sigma)$	1+ward+cs(loglos,1)+year 1+year+ward	4459.4 (4479.4) [4531.8]
IV	$\text{logit}(\mu)$ $\log(\sigma)$	1+ward+cs(loglos,1)+year+cs(age,1) 1+year+ward	4454.4 (4478.4) [4541.2]

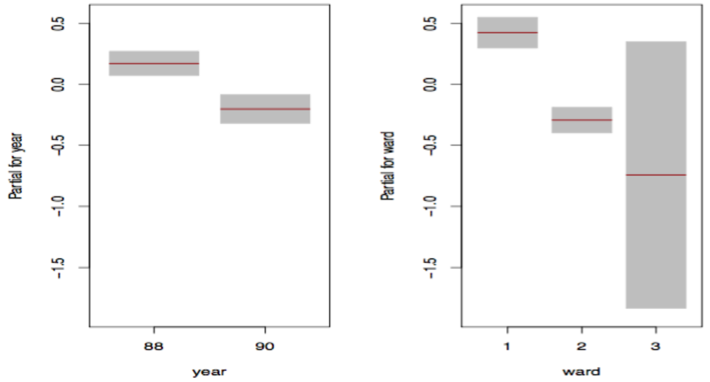
The hospital stay data

```
> op <- par(mfrow = c(2, 2))
> term.plot(mIV, se = T)
> par(op)
> op <- par(mfrow = c(2, 1))
> term.plot(mIV, "sigma", se = T)
> par(op)
> rqres.plot(mIV)
```

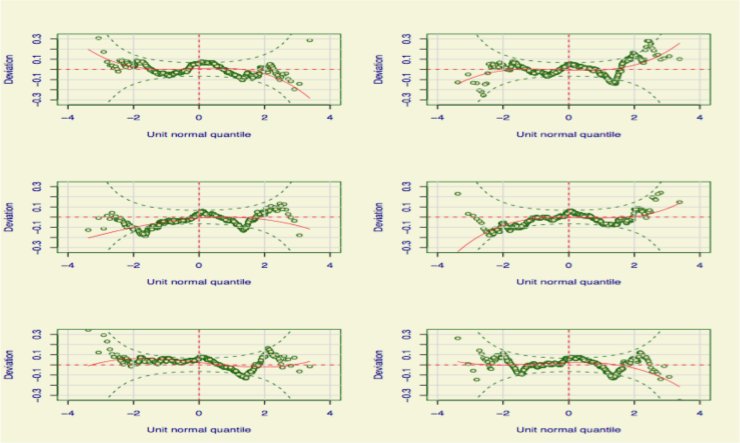
The hospital stay data: fitted model fop μ



The hospital stay data: fitted model for σ



The hospital stay data: normalised randomised quantile residuals



END

for more information see

www.gamlss.org