Introduction: Why use gamlss?

- Unified framework for univariate regression type of models
- 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 (and then use your favourite mode of inference)
- It deals with
  1. Skewness
  2. Kurtosis
  3. Overdispersion

Example 1: BMI against AGE for Dutch girls

BMI

AGE
Example 2: The fish species data, Stein and Juritz (1988)

Example 3: Visual analog scale (VAS) data

2. Model definition

Univariate Regression type model

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

\[
\begin{align*}
g_1(\mu) &= \eta_1 = X_1\beta_1 + \sum_{j=1}^{J_1} Z_{j1}\gamma_{j1} \\
g_2(\sigma) &= \eta_2 = X_2\beta_2 + \sum_{j=1}^{J_2} Z_{j2}\gamma_{j2} \\
g_3(\nu) &= \eta_3 = X_3\beta_3 + \sum_{j=1}^{J_3} Z_{j3}\gamma_{j3} \\
g_4(\tau) &= \eta_4 = X_4\beta_4 + \sum_{j=1}^{J_4} Z_{j4}\gamma_{j4}.
\end{align*}
\]

GAMLSS philosophy: The response should have a distribution and all the parameters of the distribution could be modelled as functions of explanatory variables.

The GAMLSS model

Here \( \gamma_{jk} \sim N_{q_{jk}}(0, G_{jk}^{-1}) \) and \( G_{jk} = G_{jk}(\lambda) \).
MAP estimation of \((\beta, \gamma)\) given \(\lambda\)

Hence given \(\lambda\),
posterior mode (or MAP) estimation of \((\beta, \gamma)\)

1. maximising \(l_h\), hierarchical log likelihood
2. maximising \(l_p\), penalised log likelihood

with respect to \((\beta, \gamma)\)

3. Parametric Additive terms

- Linear and interaction terms for variates and factors.
- Polynomials, inverse polynomials, piecewise polynomials (with fixed knots), fractional polynomials (Royston and Altman, 1994)
- Non-linear parametric terms

Additive terms

- Additive smoothing terms
  - cubic splines (Green and Silverman, 1994)
  - P-splines (Eilers and Marx, 1996)
  - varying coefficient models (Hastie and Tibshirani, 1993)
  - loess (Cleveland et al., 1993)
- Random effects (overdispersion, simple random effects, random coefficients)
- Parameter driven Time Series (random walks)

1. Population distributions for \(Y\)

4.1 General comments

2. A wide range of discrete, continuous and mixed distributions implemented, including highly skew and kurtotic distributions
3. Easy implementation of new distributions
4. Different parameterisations of a distribution can be implemented
5. Truncated distributions and censored data easily implemented
6. Finite mixture distributions easy to implement (new)
4.2 Discrete distributions for \( Y \)

Two parameter distributions

- **BB**  Beta-Binomial
- **NBI**  Negative Binomial type I
- **NBII**  Negative Binomial type II
- **PIG**  Poisson-Inverse Gaussian
- **ZIP**  Zero inflated Poisson

Three parameter distributions

- **SICHEL**  Sichel
- **DEL**  Delaport

4.3 Continuous distributions for \( Y \)

Four parameters

- \( \mu \)  location
- \( \sigma \)  scale
- \( \nu \)  skewness
- \( \tau \)  kurtosis

Skewness and kurtosis

- Negative skewness
- Positive skewness
- Platykurtosis
- Leptokurtosis

Distributions in real line

<table>
<thead>
<tr>
<th>Distributions</th>
<th>family</th>
<th>no parameters</th>
<th>skewness</th>
<th>kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gumbel</td>
<td>GU()</td>
<td>2</td>
<td>(negative)</td>
<td>-</td>
</tr>
<tr>
<td>Johnson's SU</td>
<td>JSU()</td>
<td>4</td>
<td>both</td>
<td>lepto</td>
</tr>
<tr>
<td>Johnson's original SU</td>
<td>JSU0()</td>
<td>4</td>
<td>both</td>
<td>lepto</td>
</tr>
<tr>
<td>Logistic</td>
<td>LO()</td>
<td>2</td>
<td>(symmetric)</td>
<td>lepto</td>
</tr>
<tr>
<td>NET</td>
<td>NET()</td>
<td>2 + (2 fixed)</td>
<td>(symmetric)</td>
<td>lepto</td>
</tr>
<tr>
<td>Normal</td>
<td>NO()</td>
<td>2</td>
<td>(symmetric)</td>
<td>meso</td>
</tr>
<tr>
<td>Power Exponential</td>
<td>PE()</td>
<td>3</td>
<td>(symmetric)</td>
<td>both</td>
</tr>
<tr>
<td>Reverse Gumbel</td>
<td>RG()</td>
<td>2</td>
<td>positive</td>
<td>-</td>
</tr>
<tr>
<td>Sinh Arc Sinh</td>
<td>SHASH()</td>
<td>4</td>
<td>both</td>
<td>lepto</td>
</tr>
<tr>
<td>Skew Exponential Power</td>
<td>SEP()</td>
<td>4</td>
<td>both</td>
<td>both</td>
</tr>
<tr>
<td>Skew t</td>
<td>ST()</td>
<td>4</td>
<td>both</td>
<td>lepto</td>
</tr>
<tr>
<td>t Family</td>
<td>TF()</td>
<td>3</td>
<td>(symmetric)</td>
<td>lepto</td>
</tr>
</tbody>
</table>
Distributions in positive real line

5. The R packages

- `gamlss`
- `gamlss.nl`
- `gamlss.tr`
- `gamlss.dist` (to be released)
- `gamlss.mx` (new)
- `gamlss.cen` (future)

The `gamlss` package

- The `gamlss()` function creates a `gamlss` object
- Methods for a `gamlss` object:
  - AIC(), addterm(), coef(), deviance(), fitted(), formula(), lot(), print(), predict(), residuals(), update(),
  - Others functions:
    - centiles(), fitted.plot(), GAIC(), gamlss.scope(),
    - par.plot(), lperd(), pdf.plot(), prof.plot(), prof.term(),
    - Q.stats(), refit(), rqres.plot(), stepGAIC(), term.plot()
6. Examples: Modelling body mass index (BMI) against AGE

Variables
- Body mass index ($Y = BMI$) against AGE,
  for 20243 Dutch girls aged under 20

Study
- cross sectional data,
- Dutch population study,
- Cole and Roede (1999)

Model for BMI

$$Y \sim BCT (\mu, \sigma, \nu, \tau)$$

where $BCT$ is the Box-Cox $t$ distribution, where $Y = BMI$ and $x = AGE^\xi$

- $\mu = cs(x, df_\mu)$
- $\log(\sigma) = cs(x, df_\sigma)$
- $\nu = cs(x, df_\nu)$
- $\log(\tau) = cs(x, df_\tau)$

We need to select the five values $df_\mu, df_\sigma, df_\nu, df_\tau, \xi$
Centile curves of BMI against AGE (# = 2.4)
(0.4, 2.3, 10, 25, 50, 75, 90, 97.7, 99.6) %

Fish species data: Stein and Juritz (1988)

\[
\begin{align*}
\log(\mu) &= h_1(x) \\
\log(\sigma) &= h_2(x) \\
\nu &= h_3(x) \quad \text{for} \quad SI(\mu, \sigma, \nu) \\
\logit(\nu) &= h_3(x) \quad \text{for} \quad DEL(\mu, \sigma, \nu)
\end{align*}
\]
Visual analog scale (VAS) data

368 patients, measured at 18 times with 7 treatments

Model for all data

\[ Y \sim BEINF(\mu, \sigma, \nu, \tau) \text{ where } BEINF \text{ is the Beta inflated distribution} \]

\[ \mu = cs(time, df=10) + treat + random(patient, 250) \]

\[ \log|\sigma| = cs(time, df=10) + treat \]

\[ \log|\nu| = cs(time, df=5) + treat \]

\[ \log|t| = cs(time, df=5) + treat \]

Model for mu

\[ \mu = cs(time, df=10) + treat + random(patient, 250) \]
Model for 50 and 550 participant

Fitting distributions to data

Fitting discrete distributions to computer failure data
tensile strength data: plots
6. Conclusions

GAMLSS allows flexible modelling of both:

i) the distribution of $Y$, including models for skewness and kurtosis

ii) the dependence of the distribution parameters, e.g. $\mu$, $\sigma$, $\nu$, $\tau$,
on explanatory variables and random effect additive terms.

GAMLSS papers, manual and related publications, available from website

http://www.londonmet.ac.uk/gamlss/

END