1. The statistical modelling philosophy
2. Motivating examples
   - BMI data
   - Lung function data
   - Stylometric data
   - The number of physician office visit
3. The past: GAMLSS Model definition
   - What is GAMLSS?
   - Distributions and Additive terms
   - R implementation
4. The present and future
   - GAMLSS components
   - Problems, solutions, and future research
5. Conclusions
Statistical modelling is the art of using statistical reasoning to build a parsimonious models for a better understanding of the phenomena of interest.

- get data
- build a model
- interpretate/predict
The statistical modelling philosophy

The statistical modelling principals

- Any **model** is a simplification of reality therefore **no model is correct** but some of them are useful.

- **Occam’s Razor** which states ‘**entities should not be multiplied beyond necessity**’ or **KISS** (Keep It Simple Stupid)

- Far better an **approximate** answer to the **right** question, which is often vague, than an exact answer to the wrong question, which can always be made precise. – John W. Tukey

- ”**no matter how beautiful your theory/model, no matter how clever you are or what your name is, if it disagrees with experiment”/data, ”it’s wrong”** (Richard Feynman)

- **Test** all the time your assumptions (there is no free meal)

- Try **different** models and choose the most appropriate for the data (have a data scientist attitude).
The Dutch boys data

**BMI**: the BMI of 7294 boys

**age**: the age in years

Source: van Buuren and Fredriks (2001)
The Dutch boys data: statistical challenges
The Dutch boys data: Histograms by age
The Dutch boys data: Conditional histograms by age
The Dutch boys data: centile estimation
The Dutch boys data: centiles
World Health Organisation Child Growth Standards: Girls

BMI-for-age GIRLS
5 to 19 years (z-scores)

![BMI-for-age Girls Graph](image-url)
3164 male observations of lung function data
The lung function data

\[ Y = \frac{FEV_1}{FVC} \] : the Spirometric lung function an established index for diagnosing airway obstruction (3164 male)

\[ \text{height} \] : the height in cm

Source: Stanojevic et al. 2009
The lung function data: fitted centile curves

(a) LMS

(b) BEINF1

(c) Inf. logitSST

(d) Gen. Tobit
A stylometric application

64 observations

- **word**: is the number of times a word appears in a single text
- **freq**: the number of different words which occur exactly word times in the text

Source: Prof. Mario Cortina-Borja
The stylometric data
Motivating examples

The number of physician office visit

- visits: number of physician office visits,
- hospital: number of hospital stays,
- health: health status: a factor indicating whether self-perceived health is poor, average (reference category) or excellent,
- chronic: number of chronic conditions,
- gender: a factor indicating gender,
- school: number of years of education,
- insurance: a factor indicating whether the individual is covered by private insurance.

Data in AER package in R
Motivating examples

The number of physician office visit

The number of physician office visit

Motivating examples

The number of physician office visit

Mikis Stasinopoulos, Rob Rigby, Gillian Heller, Vlasios Voudouris and Fernanda De Bastiani

Generalised Additive Models for Location Scale and Shape

2015 18 / 50
What we need for modelling the above data?

We need

- **flexible distributions** for the response variable
- to be able to deal with **heterogeneity** in the data
- to be able to model **skewness** and **kurtosis**
- to be able to model **overdispersion, excess of zeros** and **long tails** in count data
- We need modelling **all** the parameters of the distributions
- **flexible functions** to model the relationship between the parameter of the distribution and the explanatory variables
Important events in the creation of the GAMLSS models

**Linear model** (Gauss, 1809)

**1972** Generalised Linear Models (Nelder and Wedderburn)

**1990** Generalised Additive Models (Hastie and Tibshirani)

**2005** Generalised Additive Models for Location Scale and Shape (GAMLSS) (Rigby and Stasinopoulos).
Generalised Additive Model for Location Scale and Shape

Rigby and Stasinopoulos (2005)

\[ y \sim D(\mu, \sigma, \nu, \tau) \]

\[
g_\mu(\mu) = X_\mu \beta_\mu + h_{1,\mu}(x_{1,\mu}) + \ldots + h_{k,\mu}(x_{k,\mu})
\]

\[
g_\sigma(\sigma) = X_\sigma \beta_\sigma + h_{1,\sigma}(x_{1,\sigma}) + \ldots + h_{k,\sigma}(x_{k,\sigma})
\]

\[
g_\nu(\nu) = X_\nu \beta_\nu + h_{1,\nu}(x_{1,\nu}) + \ldots + h_{k,\nu}(x_{k,\nu})
\]

\[
g_\tau(\tau) = X_\tau \beta_\tau + h_{1,\tau}(x_{1,\tau}) + \ldots + h_{k,\tau}(x_{k,\tau})
\]

where \( D(\mu, \sigma, \nu, \tau) \) can be any distribution and where \( h_j(x_j) \) are smooth functions of the \( X \)'s.
GAMLSS assumptions
What is GAMLSS?

GAMLSS: are semi-parametric regression type models.

- **regression type**: we have many explanatory variables $\mathbf{X}$ and one response variable $\mathbf{y}$ and we believe that $\mathbf{X} \rightarrow \mathbf{y}$
- **parametric**: a parametric distribution assumption for the response variable,
- **semi**: the parameters of the distribution, as functions of explanatory variables, may involve non-parametric smoothing functions
- **GAMLSS philosophy**: try different models

GAMLSS is a generalisation of GLM and GAM models.
There are more than 100 explicit discrete, continuous, and mixed distributions, implemented as `gamlss.family` in the R, including highly skew and kurtotic distributions.

- Creating a new distribution is relatively easy.
- Truncating an existing distribution.
- Using a censored version of an existing distribution.
- Mixing different distributions to create a new finite mixture distribution.
- Discretising continuous distributions.
- Log or logit any continuous distribution in $(-\infty, \infty)$.
- Any distribution in $(0, \infty)$ can be zero adjusted to $[0, \infty)$.
- Any distribution in $(0, 1)$ can be inflated to $[0, 1]$. 
Additive Terms
GAMLSS: R implementation

GAMLSS is implemented in series of packages in R

- **gamlss** the original package
- **gamlss.dist** for distributions
- **gamlss.data** for distributions
- **gamlss.demo** for demos
  - **gamlss.nl** for non-linear terms
  - **gamlss.tr** for truncated distributions
- **gamlss.cens** for censored (left, right or interval) response variables
- **gamlss.mx** for finite mixtures and random effects
- **gamlss.spatial** for Gaussian Markov Random Fields
- **gamlss.inf** for zero adjusted and inflated mixed distributions

The GAMLSS packages can be downloaded from CRAN, the R library at

http://www.r-project.org/
Let \( M = \{D, G, T, \lambda\} \) represent the GAMLSS model

- \( D \): distribution
- \( G \): the link function for distributional parameters
- \( T \): predictor terms for \( (\eta\text{'s}) \) i.e. \( \eta = X\beta + \sum_j h_j(x_j) \)
- \( \lambda \): the hyper-parameters
Problems, solutions and future research

- which distribution $\mathcal{D}$?
  - a book on distribution is prepared
  - a new function chooseDist()
  - robustify distributions
    - before fitting
    - after fitting

- which additive term for $\mu$, $\sigma$, $\nu$ and $\tau$?
  - all step-GAIC's are now parallel
  - possible connection of ChooseDist() and stepGAIC()
  - Machine learning techniques
    - GAMLSS boosting is well developed
    - connection to glmnet

- choosing the smoothing hyper parameters for terms
  - connection to caret package
The present and future Problems, solutions, and future research

Problems, solutions and future research

- selection between different (GAMLSS or not) models
  - GAIC and diagnostics exist but more work is needed to see where the benefits of using GAMLSS are coming from
  - influential observations

- Which inferential procedure?
  - penalised likelihood
  - Bayesian, see package BAMLSS
  - boosting, see package gamboostLSS

- Forcasting
  - developing time series modelling within GAMLSS
  - distributional forecast
  - automation
The present and future

Problems, solutions, and future research

The Books

- **Flexible Regression and Smoothing: Using GAMLSS in R** (out in April 2017)
- **Distributions for Location Scale and Shape: Using GAMLSS in R** (expected in six to eight months)
- **Generalized Additive Models for Location Scale and Shape: A Distributional Regression Approach.** (starts in September)
Conclusions

- **GAMLSS** is a very flexible statistical model
- It is a **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 **penalised additive** terms in the models for the distribution parameters
- The fitted algorithm is **modular**, where different components can be added easily
- It can easily introduced to **students** since it relies on known concepts
- It deals with **overdispersion**, **skewness** and **kurtosis**
This is a collaborative work

<table>
<thead>
<tr>
<th>co-authors</th>
<th>current collaborators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vlasios Voudouris</td>
<td>Paul Eleirs</td>
</tr>
<tr>
<td>Gillian Heller</td>
<td>Marco Enea</td>
</tr>
<tr>
<td>Andreas Mayr</td>
<td>Daniil Kiose</td>
</tr>
<tr>
<td>Fernanda De Bastiani</td>
<td>Majid Djennad</td>
</tr>
<tr>
<td>Thomas Kneib</td>
<td>Luiz Nakamura</td>
</tr>
<tr>
<td>Nadja Klein</td>
<td>Abu Hossain</td>
</tr>
<tr>
<td></td>
<td>past collaborators</td>
</tr>
<tr>
<td></td>
<td>Popi Akantziliotou</td>
</tr>
<tr>
<td></td>
<td>Fiona McEllduff</td>
</tr>
<tr>
<td></td>
<td>Raydonal Ospina</td>
</tr>
<tr>
<td></td>
<td>Konstantinos Pateras</td>
</tr>
<tr>
<td></td>
<td>Nicoleta Mortan</td>
</tr>
</tbody>
</table>
For more GAMLSS

the END

for more information see

www.gamlss.org
Example of mixed distribution distributions
Continuous distributions: different shapes

- **Negative skewness**
  - Distribution with a tail on the left side
  - Skewed to the left

- **Positive skewness**
  - Distribution with a tail on the right side
  - Skewed to the right

- **Platy-kurtosis**
  - Distribution with a flatter peak than a normal distribution

- **Lepto-kurtosis**
  - Distribution with a sharper peak than a normal distribution

---

Mikis Stasinopoulos, Rob Rigby, Gillian Heller, Vlasios Voudouris and Fernanda De Bastiani

Generalised Additive Models for Location Scale and Shape

2015 36 / 50
Continuous distributions: different types
Continuous distributions: different types

Gamma truncated distribution

pdf

0.00
0.05
0.10
0.15
0
5
10
15
x
Continuous distributions: different types
Continuous distributions: different types
The stylometric data,
Choose Distribution

mf <- chooseDist(m1, type="count")

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3.84</th>
<th>8.39</th>
</tr>
</thead>
<tbody>
<tr>
<td>PO</td>
<td>35959.23</td>
<td>35973.95</td>
<td>36010.35</td>
</tr>
<tr>
<td>GEOM</td>
<td>24402.77</td>
<td>24417.49</td>
<td>24453.89</td>
</tr>
<tr>
<td>GEOMo</td>
<td>24402.77</td>
<td>24417.49</td>
<td>24453.89</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>ZASICHEL</td>
<td>24469.07</td>
<td>24502.19</td>
<td>24584.09</td>
</tr>
<tr>
<td>ZINBF.1</td>
<td>24207.21</td>
<td>24240.33</td>
<td>24322.23</td>
</tr>
<tr>
<td>ZIBNB</td>
<td>24107.94</td>
<td>24141.06</td>
<td>24222.96</td>
</tr>
<tr>
<td>ZISICHEL</td>
<td>24196.25</td>
<td>24229.37</td>
<td>24311.27</td>
</tr>
</tbody>
</table>

getOrder(mf)
The linear model

Linear Model, Gauss

\[ y = X\beta + \epsilon \text{ where } \epsilon \sim NO(0, \sigma^2 I) \]

The model can be also written as:

\[ y \sim NO(\mu, \sigma^2 I) \text{ where } \mu = X\beta \]
The linear model assumptions
The weighted linear model assumptions
The linear model: comments

- estimation is achieved by Least Squares or Weighted Least Squares (WLS)
- the normal distribution is important for inference
- we only modelling the mean as linear function of the explanatory variables
- One of the top ten reasons to become statistician (according to Friedman, Friedman & Amoo, 2002, Journal of Statistics Education):

  “Statisticians are mean lovers”.
The generalised linear model

**Generalised Linear Model**, Nelder and Wedderburn (1972)

\[ g(\mu) = X\beta \text{ where } y \sim \text{ExpFamily}(\mu, \phi) \]

where \( g() \) is the link function

The exponential family

1. normal
2. Gamma
3. inverse Gaussian
4. Poisson
5. binomial
The generalised linear model
The generalised linear model

- estimation is achieved by Iterative Re-weighted Least Squares (IRLS)
- we can model discrete response variables
- we are still “mean lovers”.
The generalised additive model

Generalised additive model Hastie and Tibshirani (1990)

\[ y \sim \text{ExpFamily}(\mu, \phi) \]

\[ g(\mu) = X\beta + h_1(x_1) + \ldots + h_k(x_k) \]

where \( h_j(x_j) \) are smooth functions of the \( X \)'s.