Statistical modelling and GAMLSS

Modellazione statistica attraverso i GAMLSS

Mikis D. Stasinopoulos and Robert A. Rigby and Fernanda De Bastiani

Abstract This paper reflects on the impact and future development of the generalized additive models for location scale and shape, GAMLSS. GAMLSS application is illustrated with an analysis of a real discrete data set.
Abstract *Questo lavoro discute le potenzialit attuali e future dei modelli GAMLSS (generalized additive models for location scale and shape) attraverso l'analisi di un dataset con risposta discreta e modellata attraverso un modello 'zero inflated beta negative binomial'.*

Key words: Beta negative binomial, Flexible.regression, Zero inflated beta negative binomial

1 Introduction

The generalized additive models for location scale and shape (GAMLSS) was introduced by [17]. It has been applied to a variety of different scientific fields including: actuarial science, [7], biology, [6], economics, [26], environment, [24], genomics, [10], finance, [9], [3], fisheries, food consumption, management science, [1], marine research, medicine, [18], meteorology, and vaccines. GAMLSS have also become standard for centile estimation, e.g. [25], [23], [15].

WHO, [27] and [28], use GAMLSS for centile estimation to produce growth charts for children. Their charts are used by more than 140 countries as the stan-

Mikis D. Stasinopoulos
London Metropolitan University, London, UK, e-mail: dmh.stasinopoulos@gmail.com

Robert A. Rigby
London Metropolitan University, London, UK

Fernanda De Bastiani
Universidade Federal de Pernambuco, Recife, PE, Brazil

dard charts monitoring the growth of children. The Global Lung Function Initiative (GLFI), [http://www.lungfunction.org, [16]] use GAMLSS to provide a unified worldwide approach to monitoring lung function, by obtaining centiles for lung function based on age and height.

Section 2 discussed GAMLSS within the general framework of statical modelling. Section 3 defines GAMLSS and show one of its application. In the conclusions we discuss the future of GAMLSS.

2 What is GAMLSS

GAMLSS was built around the basic principals of statistical modelling which can be summarized as: i) all models are *wrong* but some are useful (attributed to Gorge Box); ii) statistical modelling is an *iterative* process where, after fitting an initial model, assumptions are checked, followed by refitting models until an appropriate model is found; iii) a simple model is preferable to a more complicated one if both explain the data adequately (*Occam's Razor*) and iv) "no matter how beautiful your theory, no matter how clever you are or what your name is, if the model disagrees with the data, it is wrong"[1].

GAMLSS, is a general framework for *univariate* regression where we assume that the response variable depends on many explanatory variables. This dependance can be linear, non-linear or smooth non-parametric. For example, in the classical linear regression model (LM) the mean of the response variable is a linear function of the explanatory variables. In the generalized linear models (GLM), [14], a monotonic function of the mean, called the linear predictor, is a linear function of the explanatory variables. Non linear relationships between the response variable and the explanatory variables, within both LM and GLM, are dealt with by using nonparametric smoothing functions, giving additive models (AM) and generalized additive models (GAM) respectively. The generalized additive models (GAM) introduced by [5] and popularized by [29], have made the smoothing techniques within a regression framework available to a wide range of practitioners.

GAMLSS is an extension of the LM, GLM and GAM and has two main features. Firstly, in GAMLSS the assumed distribution can be any parametric distribution. Secondly, all the parameters (not only the location e.g. the mean) of the distribution can be modelled as linear or smooth functions of the explanatory variables. As a result the shape of the distribution of the response variable is allowed to vary according to the values of the explanatory variables. The GAMLSS models are an example of a "Beyond Mean Regression" model, [11]. and because of their explicit distributional assumption the response variable they also also part of the "distributional regression" modelling approach, [2].

GAMLSS allows a variety of smooth functions of explanatory variables including the ones which employ a quadratic penalty in the likelihood. The basic ideas of

[1] paraphrasing Richard Feynman famous quote

GAMLSS, have been implemented in **R** in a series of packages, [19], where residual based diagnostics facilities are also provided.

3 The GAMLSS framework

The response variable observations $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ are assumed independent with

$$\mathbf{Y} \sim D(\mu, \sigma, \nu, \tau) \tag{1}$$

where D is any (up to four distribution) distribution and where usually μ is the location parameter (e.g. the mean), σ is the scale parameter (e.g. the variance), ν and τ are the shape parameters (e.g. skewness and kurtosis). A GAMLSS model allows the modelling of all the parameters of the distribution as linear i.e. $\mathbf{X}_k\beta_k$ or smooth term functions $s_{kj}(x_{kj})$, for example:

$$\begin{aligned} g_1(\mu) &= \eta_1 = \mathbf{X}_1\beta_1 + \sum_{j=1}^{J_1} s_{1j}(\mathbf{x}_{1j}) \\ g_2(\sigma) &= \eta_2 = \mathbf{X}_2\beta_2 + \sum_{j=1}^{J_2} s_{2j}(\mathbf{x}_{2j}) \\ g_3(\sigma) &= \eta_3 = \mathbf{X}_3\beta_3 + \sum_{j=1}^{J_3} s_{3j}(\mathbf{x}_{3j}) \\ g_4(\sigma) &= \eta_4 = \mathbf{X}_4\beta_4 + \sum_{j=1}^{J_4} s_{4j}(\mathbf{x}_{4j}) \end{aligned} \tag{2}$$

where $\mathbf{X}_k$ is a known design matrix, $\beta_k = (\beta_{k1}, \ldots, \beta_{kJ_k'})^\top$ is a parameter vector of length $J_k^{'}$, s_{kj} is a smooth nonparametric function of variable X_{kj} the $\mathbf{x}_{kj}$'s are vectors of length n, and $g_k(.)$ known monotonic link function relating a distribution parameter to a predictor η_k, for $k = 1, 2, 3, 4$ and $j = 1, \ldots, J_k$.

Stasinopoulos et al. [21] provide a variety of examples using GAMLSS. Here we use the example given by [20] pages 12-22. The data consists of 4406 observations, on the following variables: `visits`, number of physician office visits (the response variable), `hospital`, number of hospital stays, `health`, a factor indicating health status, `chronic`, number of chronic conditions, `gender`, a factor, `school`, number of years of education, `insurance`, a factor indicating whether the individual is covered by private insurance. The data are available from the **AER** package under the name NMES1988. There are more than 30 available discrete count distributions in the **gamlss** package. After a stepwise selection procedure the following model using a zero inflated beta negative binomial, *ZIBNB*, distribution was chosen:

$Y \sim \texttt{ZIBNB}(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau}),$

$$
\begin{aligned}
\log(\hat{\mu}) = {} & 0.980 + 0.382\sqrt{\text{hospital}} + 0.332\sqrt{\text{chronic}} \\
& +0.025\text{school} + 0.255(\text{if health=poor}) \\
& -0.313(\text{if health=excellent}) - 0.112(\text{if gender=male}) \\
& +0.123(\text{if insurance=yes}) \\
\log(\hat{\sigma}) = {} & -1.7026 - 0.208\sqrt{\text{chronic}} + 0.394(\text{if health=poor}) \\
& -0.345(\text{if health=excellent}) + 0.197(\text{if gender=male}) \\
\log(\hat{\nu}) = {} & -2.679 + 0.966\sqrt{\text{hospital}} \\
\log[\hat{\tau}/(1-\hat{\tau})] = {} & -1.077 - 0.744\sqrt{\text{chronic}} - 1.546(\text{if insurance=yes}),
\end{aligned}
\qquad (3)
$$

Figure 1 shows the worm plot and the rootogram of the fitted final model both indicating the the fit is adequate except for the extreme right tail.

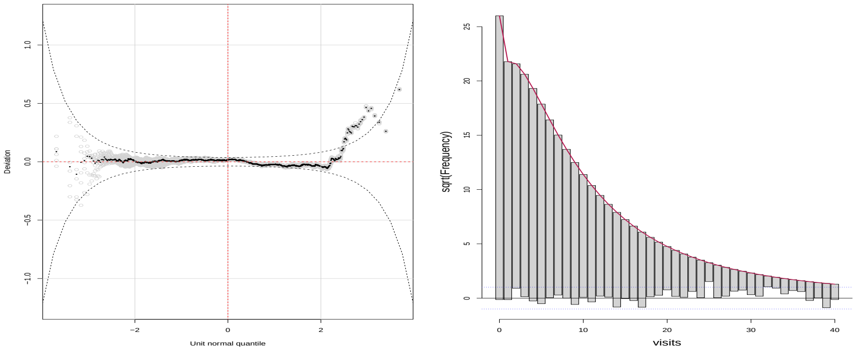

Fig. 1 Worm plot (a) and rootogram of the randomised quantile residuals for the final *ZIBNB* models. .

4 Conclusions

The GAMLSS models are especially useful for continuous response variables if they are negatively skew, highly positively skew, platykurtic, or leptokurtic. For discrete responses, if there exist over-dispersion or under-dispersion, excess or shortage of zero values and long tails. The GAMLSS models allows **any** parametric distribution for the response variable. The current implementation in the **gamlss** package in `R` allows the user to choose from more than 100 distributions with up to four parameters, allowing changes in modelling the location, the scale and shape of the distribution. A boosting and a Bayesian versions of GAMLSS exist in `R` packages , see [8, 13] and [22] respectively. There are alternative approaches to GAMLSS for quantile or centile estimation: for continuous response variable quantile regression, [12], can be used. In quantile regression there are less assumption than GAMLSS and for this reason more difficult to check the model. For mean (and variance) estimation the generalized estimation equation (GEE) [4] also can be used.

At this moment of time the following work is under way for the development and enhancement of GAMLSS:

- A second book on GAMLSS with the title "Distributions for Modelling Location, Scale and Shape: Using GAMLSS in R" is in its final draft.
- Robust methods of GAMLSS model are developed.
- Alternative model selection techniques are explored.
- Time series modelling techniques within GAMLSS are investigated.

The GAMLSS model provide a very general framework for regression type of modelling but it flexibility it is also its burden. More automated procedures would help its spread and its popularity

References

1. Budge, S., Ingolfsson, A., and Zerom, D. Empirical analysis of ambulance travel times: The case of calgary emergency medical services. Management Sci- ence, **56**(4):716-723. (2010).
2. Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. Regression: Models, Methods and Applications. Springer-Verlag, New York. (2013).
3. Giraud, G. and Kockerols, T. Making the european banking union macro- economically resilient: Cost of non-europe report. Report to the European Parliament. (2015).
4. Hardin, J. W. and Hilbe, J. M. Generalized Estimating Equations. Chapman and Hall/CRC. (2003).
5. Hastie, T. J. and Tibshirani, R. J. Generalized additive models. Chapman and Hall, London. (1990).
6. Hawkins, E., Fricker, T. E., Challinor, A. J., Ferro, C. A., Ho, C. K., and Osborne, T. M. Increasing influence of heat stress on french maize yields from the 1960s to the 2030s. Global change biology, **19**(3):937-947. (2013).
7. Heller, G., Stasinopoulos, D., Rigby, R., and De Jong, P. Mean and dis- persion modelling for policy claims costs. Scandinavian Actuarial Journal, 2007(4):281-292. (2007).
8. Hofner, B., Mayr, A., Fenske, N., and Schmid, M. gamboostLSS: Boosting Methods for GAMLSS Models. R package version 2.0-0. (2017).
9. International Monetary Fund Stress Testing, Technical Note. Country Report No. 15/173. (2015).
10. Khondoker, M. R., Glasbey, C., and Worton, B. A comparison of parametric and nonparametric methods for normalising cdna microarray data. Biometrical Journal, **49**(6):815-823. (2007).
11. Kneib, T. Beyond mean regression. Statistical Modelling, **13**:275-303. (2013).
12. Koenker, R. Quantile regression: 40 years on. Annual Review of Economics, **9**:155-176. (2017).
13. Mayr, A., Fenske, N., Hofner, B., Kneib, T., and Schmid, M. Generalized additive models for location, scale and shape for high dimensional data, a flexible approach based on boosting. J. R. Statist. Soc. Series C, **61**:403-427. (2012).
14. Nelder, J. A. and Wedderburn, R. W. M. Generalized linear models. Journal of the Royal Statistical Society, Series A, **135**:370-384. (1972).

15. Neuhauser, H. K., Thamm, M., Ellert, U., Hense, H. W., and Rosario, A. S. Blood pressure percentiles by age and height from nonoverweight children and adolescents in germany. Pediatrics, pages peds-2010. (2011).
16. Quanjer, P. H., Stanojevic, S., Cole, T. J., Baur, X., Hall, G. L., Culver, B. H., Enright, P. L., Hankinson, J. L., Ip, M. S. M., Zheng, J., et al. Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung func- tion 2012 equations. European Respiratory Journal, **40**(6):1324-1343. (2012).
17. Rigby, R. A. and Stasinopoulos, D. M. Generalized additive models for location, scale and shape, (with discussion). Applied Statistics, **54**:507-554. (2005).
18. Rodrigues, J., de Castro, M., Cancho, V., and Balakrishnan, N. Com-poisson cure rate survival models and an application to a cutaneous melanoma data. Journal of Statistical Planning and Inference, **139**(10):3605-3611. (2009).
19. Stasinopoulos, D. M. and Rigby, R. A. Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software, **23**(7):1-46. (2007).
20. Stasinopoulos, D. M., Rigby, R. A., and F., D. B. Gamlss: A distributional regression approach. Statistical Modelling, **18**:1-26. (2018).
21. Stasinopoulos, D. M., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. Flexible Regression and Smoothing: Using GAMLSS in R. Chapman and Hall, Boca Raton. (2017).
22. Umlauf, N., Klein, N., and Zeileis, A. BAMLSS: Bayesian additive models for location, scale and shape (and beyond). Journal of Computational and Graphical Statistics. (2017).
23. Villar, J., Cheikh, I. L., Victora, C. G., Ohuma, E. O., Bertino, E., Altman, D., Lambert, A., Papageorghiou, A. T., Carvalho, M., Jaffer, Y. A., Gravett, M. G., Purwar, M., Frederick, I., Noble, A. J., Pang, R. Barros, F. C., Chumlea, C. Bhutta, Z. A., and Kennedy, S. H. International standards for newborn weight, length, and head circumference by gestational age and sex: The newborn cross-sectional study of the intergrowth-21st project. The Lancet., **384**(9946):857-868. (2014).
24. Villarini, G., Smith, J., Serinaldi, F., Bales, J., Bates, P., and Krajewski, W. Flood frequency analysis for nonstationary annual peak records in an urban drainage basin. Advances in Water Resources, **32**(8):1255-1266. (2009).
25. Visser, G. H., Eilers, P. H. C., Elferink-Stinkens, P. M., Merkus, H. M., and Wit, J. M. New dutch reference curves for birthweight by gestational age. Early human development, **85**(12):737-744. (2009).
26. Voudouris, V., Ayres, R., Serrenho, A. C., and Kiose, D. The economic growth enigma revisited: The EU-15 since the 1970s. Energy Policy. (2015).
27. WHO, M. G. R. S. G. WHO Child Growth Standards: Head circumference- for-age, arm circumference-for-age, triceps circumference-for-age and subscapular skinford-for-age: Methods and development. Geneva: World Health Organization. (2007).
28. WHO, M. G. R. S. G. WHO Child Growth Standards: Growth velocity based on weight, length and head circumference: Methods and development. Geneva: World Health Organization. (2009).
29. Wood, S. N. Generalized additive models. An introduction with R. Chapman and Hall. (2017).