Environment, Resource and Ecology Journal (2021) 5: 19-26
Clausius Scientific Press, Canada

DOI: 10.23977/erej.2021.050404
ISSN 2616-3756

Application of GAMLSS Model in the Analysis of Higher Mathematics Achievements in Our University

Xinhong Liu, Xingyun Duan*, Xiangyu Zhou, Yuan Feng, Chunxia Wu

Beijing Institute of Petrochemical Technology, Beijing, 102617, China

*Corresponding author: 2018311578@bipt.edu.cn

Keywords: data mining, GAMLSS, EGB2 distribution, JSUo distribution, regression model

Abstract: With the help of statistical software R, EGB2 distribution and JSUo distribution are used to fit the test results of higher mathematics, two GAMLSS models are established, and the rationality and coefficients of the models are tested. GAMLSS model not only quantifies the impact of ordinary learning on final test scores, but also provides timely feedback and control for students'learning, and provides a basis for teachers to reflect on the teaching process and improve teaching methods in time.

1. Introduction

The progress of human society is inseparable from the wide application of mathematics. Under the background of global digitization, the emergence and popularization of computers have widened the application field of mathematics. Modern mathematics has become a powerful driving force for the development of science and technology, and has penetrated into various fields of natural science and social science widely and deeply. As a part of modern mathematics, higher mathematics is a basic course in Colleges and universities, and its learning effect has been highly valued by all levels of society. The learning effect of higher mathematics is not only affected by the learning effect of primary mathematics, but also by the learning effect of higher mathematics at ordinary times. Many scholars have studied the learning effect, such as Xu Yongmei[1], Dong Qinghua and Wang Suyan[2], Chen Wei and Yang Yue[3], who have used analytic hierarchy process, principal component analysis, diversified intelligence theory, learning harvest theory model and other methods to evaluate and study the learning effect from different angles, but due to different research objects, different methods, The analysis results are also different. Therefore, in order to accurately grasp the learning effect, this paper applies the GAMLSS model introduced by Rigby and Stasinopoulos in 2005[4] to study the learning effect of Higher Mathematics A (I) of a professional student of grade 2018 in our university. GAMLSS model is widely used in medicine, hydrology, climate, non life insurance actuarial and other fields. For example, Liu Xinhong, Feng Yuan and Mi Haijie[5] studied the pricing of auto insurance by using GAMLSS model. Zhang Dongdong, Lu fan, Zhou Xiangnan and others[6] studied the inconsistency of extreme precipitation in Dadu River Basin Based on GAMLSS model. In this paper, the GAMLSS model between the final test scores and the influencing factors is established, and the rationality test of the model is realized by using the software package GAMLSS[7] in R software. GAMLSS model quantifies the impact of usual learning on final test scores. The results can not only provide timely feedback and control for students' learning, but also provide a basis for teachers to reflect on the

teaching process and improve teaching methods in time.

2. GAMLSS model

In 2005, Rigby and Stasinopoulos introduced the model generalized additive models for location, scale and shape (GAMLSS). This model is more flexible and complex than multiple regression model. Firstly, the response variables are no longer limited to the exponential distribution family. Secondly, the system not only considers the relationship between location parameters and explanatory variables, but also includes the relationship between scale parameters, shape parameters and explanatory variables. These improvements enable GAMLSS model to better explain the relationship between data.

2.1 GAMLSS model definition

Assume that the probability density function of the response variable is $f\left(y \mid \theta\right)$, where, θ is a vector of m parameters $\theta = \left(\theta_1, \theta_2, \cdots, \theta_m\right)^T$. Given $\theta = \theta^i = \left(\theta_{i1}, \theta_{i2}, \cdots, \theta_{im}\right), y_i \left(i = 1, 2, \cdots, n\right)$ are independent, their probability density function are $f\left(y_i \mid \theta = \theta^i\right)$, Stasinopoulos and Rigby wrote GAMLSS package in 2007, which contains more than 60 different distributions. This paper adopts exponential generalized beta 2 distribution and JSUo distribution, which can describe the left and right deviation of data.

Generalized beta 2 distribution with $\mathrm{EGB2}(\mu, \rho, \upsilon, \tau)$ is

$$f_Y(y \left| \mu, \rho, \nu, \tau \right.) = e^{\upsilon z} \{ \left| \sigma \right| B(v, \tau)[1 + e^z]^{\nu+\tau} \}^{-1}, \quad -\infty < y < \infty$$

Where $-\infty < \mu < \infty, \sigma > 0, \nu > 0, \tau > 0$, $z = (y - \mu) / \sigma$, Their expectations and variances are shown below.

$$E(Y) = \mu + \sigma[\Psi(\nu) - \Psi(\tau)], \quad Var(Y) = \sigma^2[\Psi^{(1)}(\nu) - \Psi^{(1)}(\tau)]$$

JSUo distribution with $JSUo(\mu, \sigma, \nu, \tau)$ is

$$f_Y(y \left| \mu, \rho, \nu, \tau \right.) = \frac{\tau}{\sigma} \frac{1}{(r^2 + 1)^{\frac{1}{2}}} \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2} z^2], \quad -\infty < y < \infty$$

Where $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \nu < \infty$, $\tau > 0$, $z = \nu + \tau \sinh^{-1}(r) = \nu + \tau \log[r + (r^2 + 1)^{\frac{1}{2}}]$, $r = (y - \mu) / \sigma$.

The parameters ν in the JSUo distribution determine the skewness of the distribution, which $\nu > 0$ is negative skewness and $\nu < 0$ positive skewness. The parameter τ determines the peak value of the distribution, which should be positive and most likely in the region above 1. When τ tends to infinity, the distribution is close to the normal density function. The mathematical expectation and variance of the distribution are:

$$E(Y) = \mu - \sigma \omega^{1/2} \sinh(\nu / \tau), \quad Var(Y) = \sigma^2 \frac{1}{2}(\omega - 1)[\omega \cosh(2\nu / \tau) + 1], \omega = \exp(1 / \tau^2)$$

The systematic part of GAMLSS model can establish a regression model of θ^i and explanatory

variables. $\mathbf{y}^T = \left(y_1, y_2, \cdots, y_n\right)$ is a vector composed of the observed values of response variables, $g_k\left(\cdot\right)\left(k = 1,2,\cdots,p\right)$ is a known monotone connection function. The form of connected regression model is:

$$g_k\left(\theta_k\right) = \eta_k = \mathbf{X}_k\beta_k$$

Where θ_k and η_k are n-dimensional vector, $\mathbf{X}_k$ is a known $n \times J_k'$-dimensional design matrix, $\beta_k^T = \left(\beta_{1k}, \beta_{2k}, \cdots, \beta_{J_k'k}\right)$ is a J_k'-dimensional vector. This is a simplified form of GAMLSS model.

2.2 Parameter estimation of GAMLSS model

The parameters are estimated by empirical Bayesian method. Firstly, assuming that β_k obeys the prior distribution of uniform distribution, the posterior mode of parameter vector can be estimated by Rigby Stasinopoulos (RS) algorithm or Cole green (CG) algorithm or a hybrid algorithm of the two methods. These algorithms are relatively mature and can be implemented in R software package GAMLSS.

3. Empirical analysis

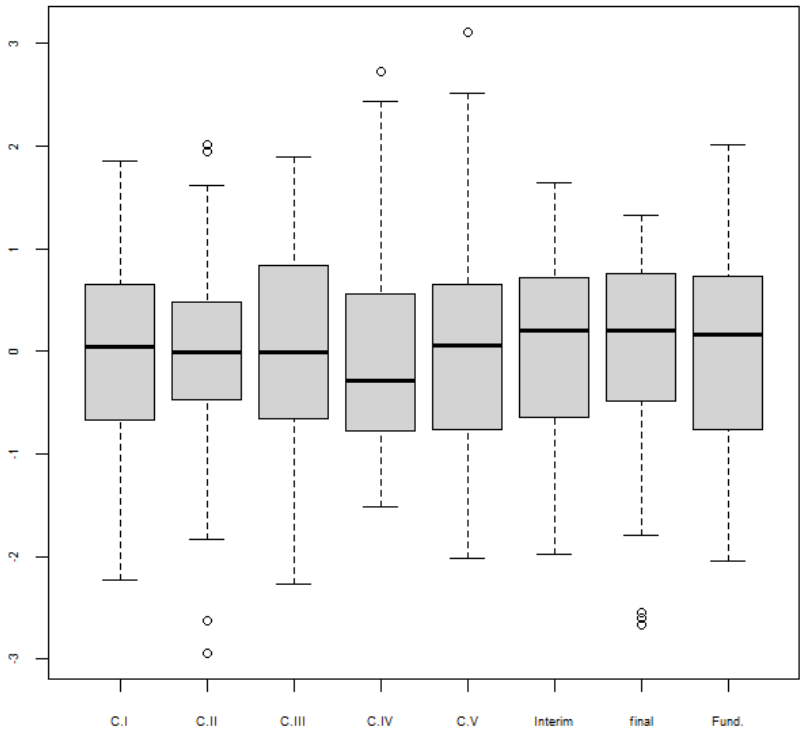

Figure 1: Distribution of mathematics scores

Combined with the actual learning situation of 56 students in a major in Higher Mathematics A (I), this paper summarizes the main factors evaluating the learning effect as follows: Mathematics in college entrance examination (the basis of Higher Mathematics), five chapters, and 8 tests at the

middle and end of the period. The collected score box line diagram is shown in Figure 1. As can be seen from Figure 1, with the continuous progress of teaching, the difficulty of chapter test is continuously improved, and the test scores in Chapter 4 and Chapter 5 are relatively low. However, after a planned review, the students' midterm test and final scores have been greatly improved.

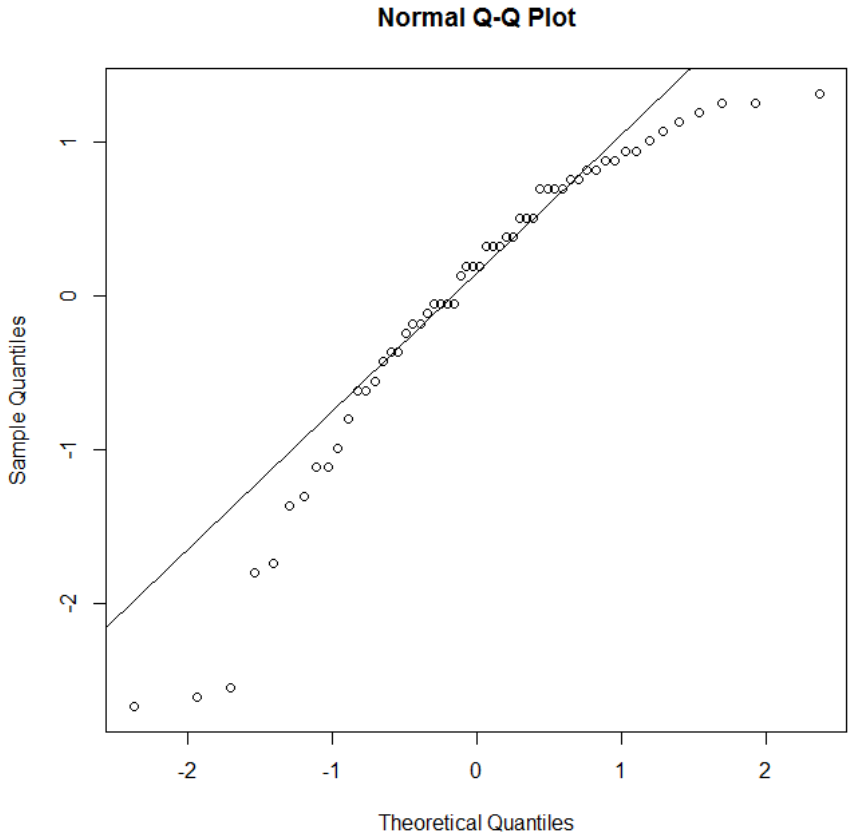

Figure 2: QQ chart of final exam results

As can be seen from Figure 2, the quantile of final grade is not in line with the theoretical quantile of normal distribution, and the final grade does not conform to the law of normal distribution.

3.1 Relationship between 8 test scores

Because higher mathematics is based on elementary mathematics, the mathematics scores of college entrance examination have a certain impact on the learning of higher mathematics, and there is a certain correlation between the knowledge of each chapter of higher mathematics, so the scores of 8 tests have a certain relationship. The correlation analysis of 8 test scores is carried out through R software. The results are shown in Figure 3 and table 1. It can be seen that there is a certain correlation between them.

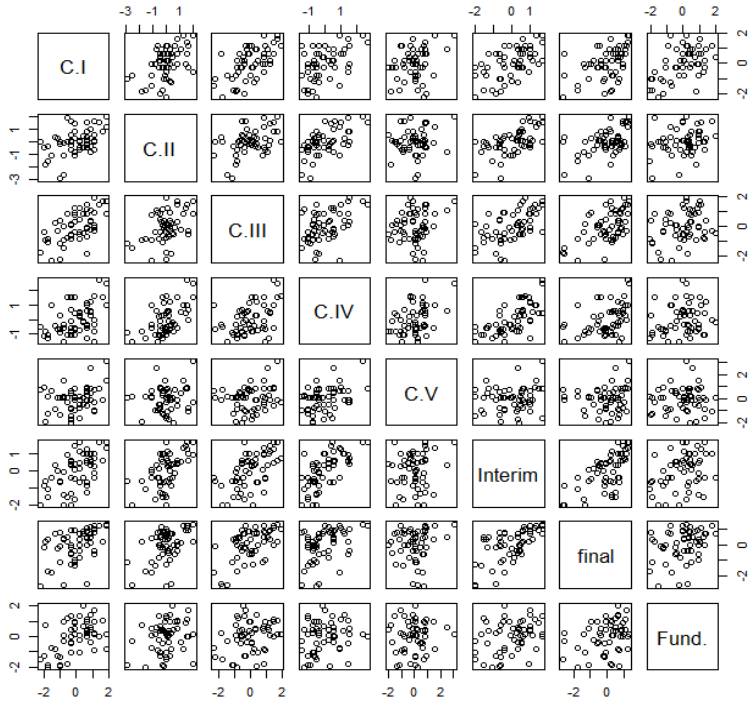

Figure 3: scatter diagram of the relationship between the scores of each chapter

Table 1: correlation coefficient

	Fund.	C.I	C.II	C.III	C.IV	C.V	Interim
Final	0.3235	0.4717	0.4846	0.5721	0.4171	0.1202	0.6334

3.2 Establishment of model

GAMLSS is an effective mathematical method to deal with the interdependence between multiple variables. Using this method, with the help of statistical software R, the GAMLSS model is established between the final grade, the test scores of each chapter and the mathematics scores of the college entrance examination, and the rationality and coefficient of the model are tested.

The GAMLSS model is established,

GAMLSS model I: $f\left(y_i \middle| \beta_1, \beta_2, \beta_3, \beta_4\right) \sim EGB2,$

$$\mu_i = x_i\beta_1, \quad \sigma_i = x_i\beta_2, \quad \log(\nu_i) = x_i\beta_3, \quad \log(\tau_i) = x_i\beta_4$$

GAMLSS model II: $f\left(y_i \middle| \beta_1, \beta_2, \beta_3, \beta_4\right) \sim JUSo,$

$$\mu_i = x_i\beta_1, \quad \log(\sigma_i) = x_i\beta_2, \quad \nu_i = x_i\beta_3, \quad \log(\tau_i) = x_i\beta_4$$

Where, $x_i (i = 1, \ldots, 56)$ is a 56×8 design matrix composed of $(1, 1, \ldots, 1)^T$ and explanatory variables, which β_1, β_2, β_3, β_4 are 8×1-Dimensional coefficient vectors respectively.

3.3 Analysis of model results

Using the GAMLSS package in R, the parameter estimation of the mathematical model between the final examination scores and the scores of each chapter is obtained. The estimation results are shown in Table 2.

Table 2: parameter estimation results

			model I			model II	
		Estimated value	error	Pr(>|t|)	Estimated value	error	Pr(>|t|)
	Intercept	3.3347	0.0584	< 2e-16	0.8637	0.0680	< 2e-16
	C.I	0.3999	0.0560	0.0000	0.5226	0.1276	0.0002
	C.II	-0.3495	0.0490	0.0000	0.2807	0.1060	0.0109
μ	C.III	-1.6587	0.0515	< 2e-16	-1.6384	0.0990	< 2e-16
	C.IV	0.2454	0.0498	0.0000	0.7460	0.1354	0.0000
	C.V	-0.1048	0.0430	0.0186	1.1485	0.1129	0.0000
	Interim	0.9652	0.0686	< 2e-16	-0.2340	0.1166	0.0505
	Fund	0.2048	0.0470	0.0001	0.4029	0.1086	0.0005
	Intercept	-0.9429	0.0162	< 2e-16	1.7393	0.0552	< 2e-16
	C.I	0.5746	0.0203	< 2e-16	-1.0953	0.1043	0.0000
	C.II	-0.3409	0.0149	< 2e-16	-0.7322	0.0793	0.0000
σ	C.III	-0.3310	0.0178	< 2e-16	0.8812	0.0768	0.0000
	C.IV	0.3610	0.0211	< 2e-16	0.7693	0.0907	0.0000
	C.V	-0.2899	0.0171	< 2e-16	-0.2468	0.0815	0.0039
	Interim	-0.2597	0.0237	1.60E-15	-0.2030	0.1022	0.0527
	Fund	-0.3131	0.0162	< 2e-16	-0.3490	0.0720	0.0000
	Intercept	0.2204	0.0964	0.0267	2.1508	0.1487	< 2e-16
	C.I	1.3780	0.1291	0.0000	1.2554	0.2799	0.0000
	C.II	-0.2680	0.1130	0.0217	0.3616	0.2343	0.1293
υ	C.III	-0.5250	0.1104	0.0000	-3.7950	0.2221	< 2e-16
	C.IV	0.3408	0.1233	0.0081	2.0540	0.3058	0.0000
	C.V	-0.8681	0.0983	0.0000	2.3281	0.2258	0.0000
	Interim	0.3978	0.1390	0.0062	-1.6777	0.2546	0.0000
	Fund	-0.4643	0.1012	0.0000	0.7266	0.2302	0.0028
	Intercept	7.8667	0.1468	< 2e-16	2.6692	0.0530	< 2e-16
	C.I	-0.9781	0.1787	1.57E-06	-0.9295	0.0934	0.0000
	C.II	1.0899	0.1352	1.77E-10	-0.4924	0.0720	0.0000
τ	C.III	-2.1603	0.1353	< 2e-16	0.6973	0.0714	0.0000
	C.IV	-1.3451	0.1477	4.98E-12	0.6101	0.0865	0.0000
	C.V	-0.2800	0.1307	0.0373	-0.1165	0.0747	0.1254
	Interim	3.1439	0.1633	< 2e-16	-0.1077	0.0963	0.2689
	Fund	1.9067	0.1272	< 2e-16	-0.1734	0.0636	0.0089
	AIC		129.7381			141.0628	

The standard for evaluating the fitting of distribution is AIC = - 2 *log likelihood function value + 2 * number of parameters. It can be known that the smaller the AIC value, the better the fitting degree. Here, the AIC values of model 1 and model 2 are 129.7381 and 141.0628 respectively, so model 1 fits best. The final grade of model I follows the residual diagram, residual density diagram

and residual QQ diagram under the distribution, as shown in Figure 4. It can be seen that the randomized residual fluctuates in the [- 3,3] interval and there is no obvious trend. The distribution shape of the residual is similar to the standard normal distribution, and the QQ diagram of the residual is almost 45 degrees straight line, indicating that the residual is in the standard normal distribution, All these show that it is reasonable to choose distribution to fit the final grade data.

It can be seen from table 3 that in the estimation of model I parameter μ, the size of regression coefficient can be explained as the contribution of each chapter's grade to μ in the final grade distribution. The regression coefficient corresponding to Chapter 1 is 0.3999, indicating that chapter 1 contributes 39.99% to μ in the final grade distribution, and the regression coefficients estimated in Chapter 2, Chapter 3 and Chapter 5 are -0.3495, -1.6587 and -0.1048 respectively, which are negative numbers, indicating that there are many knowledge points in Chapter 2, Chapter 3 and Chapter 5, which are not involved in high school mathematics, and these chapters are more difficult, These knowledge points in the final exam become the students' points of loss, and make a negative contribution to μ in the final exam score distribution, thus affecting the average score of the final exam. Other parameters are interpreted similarly.

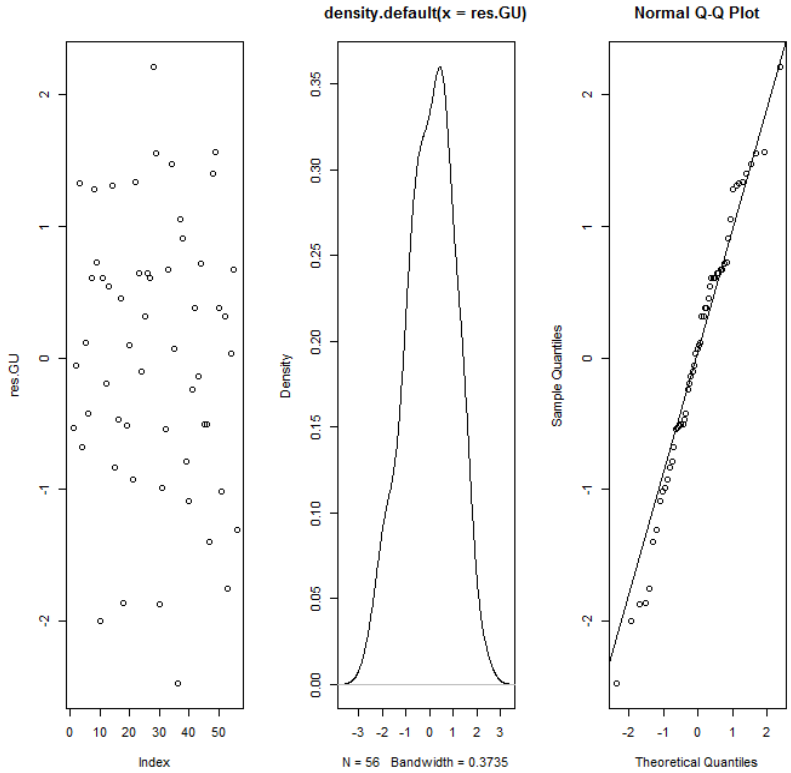

Figure 4: residual distribution of final grade

The GAMLSS model established by the above two models is in line with the reality and has a certain enlightenment for teachers' teaching and students' learning.

4. Conclusion

Using the statistical software R, this paper analyzes the correlation between the standardized data based on the higher mathematics test scores of a professional student of grade 2018, and confirms the correlation between the final scores and the test scores of each chapter and the mathematics scores of the college entrance examination.

The final test scores are fitted by EGB2 distribution and JSUo distribution. With the help of GAMLSS, a GAMLSS model between the final test scores, the test scores of each chapter and the mathematics scores of the college entrance examination is established, which is an innovative application. Through this model, the contribution of each chapter to the parameters in the final score distribution is obtained, and the influence of the test scores of each chapter on the final test scores is quantified. The model can be applied to the prediction of students' final scores of higher mathematics. Through the establishment of GAMLSS model and the analysis of final grades, it plays a certain reference role for students to plan their own learning and teachers' teaching guidance.

Acknowledgements

The authors gratefully acknowledge the financial support from the municipal URT project 2021J00118, 2021J00063 and 2020J00023, the general teaching reform project of Beijing Institute of Petrochemical Technology “Student centered, hierarchical teaching reform and research of higher Mathematics course”.

References

[1] Xu Yongmei. Analysis of teaching effect evaluation model of inquiry teaching model in higher mathematics [J]. Value engineering,2016(16):188-190.
[2] Dong Qinghua, Wang Suyan. Research on the evaluation system of "higher mathematics" learning effect based on multiple intelligences theory [J]. Textile and garment education,2016, 31(3): 251-253.
[3] Chen Wei, Yang Yue. Impact of information technology on educational ecology and educational effect: model analysis based on learning harvest [J]. University education management, 2018, 12(3): 80-86.
[4] Rigby R. A., Stasinopoulos D. M.. Generalized Additive Models for Location, Scale and Shape (with Discussion) [J]. Applied Statistics, 2005, 54(3): 507-554.
[5] Liu Xinhong, Feng Yuan, MI Haijie. Application of GAMLSS model in auto insurance pricing[J]. Mathematical practice and understanding, 2017, 47(11): 1-8.
[6] Zhang Dongdong, Lu fan, Zhou Xiangnan, et al. Inconsistency analysis of extreme precipitation in Dadu River Basin Based on GAMLSS model[J]. Water resources and hydropower technology, 2016, 47(5): 12-15.
[7] Stasinopoulos M., Rigby B.. Generalized Additive Models for Location Scale and Shape (GAMLSS) in R [J]. Journal of Statistical Software,2007, 23(7): 1-46.