Ciência Rural, Santa Maria, v.54:07, e20230056, 2024
http://doi.org/10.1590/0103-8478cr20230056
ISSNe 1678-4596
CROP PRODUCTION

Comparison of Normal, Logistic, Laplace, and Student's t distributions for experimental error in the Bayesian description of dry matter accumulation in *Allium sativum*

George Lucas Santana de Moura[1*] Felipe Guzzo[1] Paulo Roberto Cecon[1] Sebastião Martins Filho[1] Antônio Policarpo Souza Carneiro[1] Moysés Nascimento[1]

[1]Departamento de Estatística Aplicada e Biometria, Universidade Federal de Viçosa (UFV), 36570-900, Viçosa, MG, Brasil. E-mail: gelucas.moura.2016@gmail.com. *Corresponding author.

ABSTRACT: This study assessed distributions associated with Bayesian nonlinear modeling error in the description of total plant dry matter accumulation (TDMA) of *Allium sativum* as a function of days after planting (DAP). According to the DIC criterion, Logistic and Gompertz models that use student's t distribution error exhibited the highest DIC with logistic error distribution. In general, the difference of DIC in all the scenarios was not more than 5. The Bayes factor (BF) criterion showed no difference in the Logistic and Gompertz model fit when four distributions are used for the errors, where BF values do not exceed 2. Posterior distributions and the usual estimators of Logistic and Gompertz model parameters were similar even for varied error distribution. In summary, there was no difference in the use of 4 distributions associated with the modeling error of garlic plant growth by the Bayes factor, whereby the results showed that alternating between error distributions significantly changes the number of Markov Chain Monte Carlo (MCMC) iterations.
Key words: Bayesian regression, Nonlinear regression, MCMC, Symmetrical location-scale family, Empirical Bayes.

Comparação das distribuições Normal, Logística, Laplace e t de Student para o erro experimental na descrição bayesiana do acúmulo de matéria seca de *Allium sativum*

RESUMO: O objetivo deste trabalho foi avaliar algumas distribuições associadas ao erro na modelagem não linear bayesiana na descrição do acúmulo de matéria seca total da planta (MSTP) de *Allium sativum* em função dos dias após o plantio (DAP). Pelo critério DIC os modelos Logístico e Gompertz que utilizam a distribuição do erro t de Student apresentaram a melhor qualidade de ajuste, sendo que o modelo Logístico apresentou o maior DIC com a distribuição de erro Logística. No geral, a diferença de DIC em todos os cenários não apresentou valores superiores a cinco. Pelo critério do Fator de Bayes (FB), não houve diferença no ajuste do modelo Logístico e Gompertz quando se utilizam as quatro distribuições para os erros, sendo que os valores de FB não superaram 2. As distribuições a posteriori e os estimadores usuais dos parâmetros dos modelos Logístico e Gompertz apresentaram semelhanças mesmo variando a distribuição do erro. Em suma não houve diferença na utilização das quatro distribuições associadas ao erro na modelagem do crescimento planta de alho pelo fator de Bayes, sendo que os resultados mostram que alternar entre as distribuições dos erros altera de forma significativa o número de iterações de MCMC.
Palavras-chave: Regressão bayesiana, Regressão não linear, MCMC, Família simétrica de locação-escala, Bayes empírico.

INTRODUCTION

Sigmoid models are widely used in agricultural science to describe animal and plant growth. Most of the models, including Logistic and Gompertz, are analytical solutions of ordinary differential equations (STEWART, 2016).

According to CORDEIRO & DEMÉTRIO (2008), since the 1970s, nonlinear regression theory has restricted the use of nonlinear models (NLM) to the assumption of normality for the residual and consequently, the response variable. With the introduction of generalized linear models (GLM), NELDER & WEDDERBURN in 1972 defined the response variables that belong to the exponential family. Likewise, CORDEIRO & PAULA (1989) defined the exponential family for a normal nonlinear model in which the systematic component is not a linear combination of the parameters.

Nonlinear regression with a normal error is susceptible to extreme observations. However, the assumption of normal error can be relaxed using different symmetrical and asymmetrical distributions, in both linear and nonlinear models under Bayesian estimation (DE LA CRUZ & BRANCO, 2009; ROSSI & SANTOS, 2014).

In terms of dimensionality, a data base to be analyzed should contain at least 4 observations,

Received 02.01.23 Approved 08.14.23 Returned by the author 01.31.24
CR-2023-0056.R2
Editors: Leandro Souza da Silva Alessandro Dal'Col Lúcio

so that the number of parameters is smaller than the dataset ($n > p$). With maximum likelihood estimation, in addition to the 3 usual parameters of the Gompertz and Logistic models (α, β, γ) contained in the location parameter of a symmetrical nonlinear model, it is necessary to estimate the scale (s) and (v) parameters in the case of the student's t error. Dimensionality increases from 3 to 5, which is impossible to estimate in a 4-point regression. In these cases, a necessary option is Bayesian regression.

In Bayesian inference, a priori selection is made before accessing the data, using methodologies such as improper priors, conjugated distributions or meta-analysis assessment. In FIRAT et al. (2016) and MACEDO et al. (2017), the Logistic and Gompertz parameters (α, β, γ) follow normal prior distribution.

When the prior distribution obtained originates in the data, the methodology is empirical. Inference that uses prior empirical distribution is known as an empirical Bayesian approach, which is not necessarily a Bayesian inference, since the data are used twice, once in the likelihood function and once in prior distribution. However, this methodology is a good approximation for Bayesian inference (CARLIN & LOUIS, 2008).

In terms of parametric inference, GUJARATI & PORTER (2012) and SOUZA (1998) report that significance tests and confidence intervals of normal nonlinear models are asymptomatically valid, while F and t-tests, confidence intervals and regions depend on the asymptotic normality of parametric estimators. In Bayesian theory, the results of credible intervals and significance tests are valid, irrespective of sample size.

This study compared the symmetrical class of normal scale-location, student's t-test, Laplace and logistic distributions for experimental error using the Bayesian methodology to describe MSTP accumulation of garlic as a function of days after planting (DAP).

MATERIALS AND METHODS

The data used in this study were obtained from the UFV Germplasm Bank (BHG/UFV), which contains 89 *Allium sativum* fruit accessions. The experiment was conducted in the Zona da Mata region of Minas Gerais State, Brazil, (20º45'S, 42º51'W, at 650m of altitude) at the Universidade Federal de Viçosa, in randomized blocks with 8 repetitions. Accumulated plant dry matter (g) was calculated 60, 90, 120 and 150 DAP (independent variable with n =4). Each DAP includes 8 repetitions, and the mean of each DAP was used.

The normal distribution is a member of the exponential family, location-scale family and symmetrical location-scale family. Distributions other than normal also belong to the symmetrical location-scale family and are generally denoted by the letter S. In this regression, the error follows a $\varepsilon_i \sim S\ (0, \sigma^2)$ distribution and its density is $f_{\varepsilon_i}(e_i, \mu, s) = s^{-1} \cdot g[s^{-1}(e_i - \mu)]^2$, where $\mu \in R$ is the location parameter and $s > 0$ the scale parameter (CORDEIRO et al., 2000). Figure 1 represents distribution density S, these being the normal -

$$f_Y(y, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(y-\mu)^2\}$$

Logistic - $$f_Y(y, \mu, s) = \frac{\exp\left(-\frac{y-\mu}{s}\right)}{s\{1+\exp\left(-\frac{y-\mu}{s}\right)\}^2};$$

Laplace – $$f_Y(y, \mu, s) = \frac{1}{2s} \exp\left(-\frac{|y-\mu|}{s}\right)$$

and student's t models –

$$f_Y(y, \mu, s, v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \frac{1}{s\sqrt{v\pi}} \{1 + \frac{\left(\frac{y-\mu}{s}\right)^2}{v}\}^{\frac{v+1}{2}}$$

The nonlinear regression model is defined by $y_i = f(x_i, \theta) + \varepsilon_i$, and the sample distribution of the response variable also has an independent S distribution, but not identically distributed ($Y_i \sim S\ (f(x_i, \theta), s)$. For the description of garlic plant growth, the Logistic $(f(x_i; \alpha, \beta, y) = \alpha\ [1 + \beta\ exp\ (-yx_i)]^{-1}$ and Gompertz $f(x_i; \alpha, \beta, y) = \alpha\ exp\ [-\beta\ exp\ (-yx)]$ models were used, where α is asymptomatic plant growth, β a value with no biological interpretation and y the growth rate (RATKOWSKY, 1983).

In order to implement Bayesian regression, only accession 63 was used to fit the Logistic and Gompertz models. The definition of a nonlinear regression considers the Bayes theorem, where the joint posterior $\pi(\alpha, \beta, \gamma, \tau, v \mid \mathbf{y})$ is proportional to the product of the likelihood function of the response variable $L(\alpha, \beta, \gamma, \tau, v \mid \mathbf{y})$ with joint distribution of prior $\pi(\alpha, \beta, \gamma, \tau, v)$, where $\mathbf{y} = (y_1, y_2, \ldots, y_n)$.

$$\pi(\alpha, \beta, \gamma, \tau, v|\mathbf{y}) \propto L(\alpha, \beta, \gamma, \tau, v|\mathbf{y}) \cdot \pi(\alpha, \beta, \gamma, \tau, v) \quad (1)$$

The distribution of each parameter was obtained by the MCMC (Markov Chain Monte Carlo) method using the Openbugs interface, where

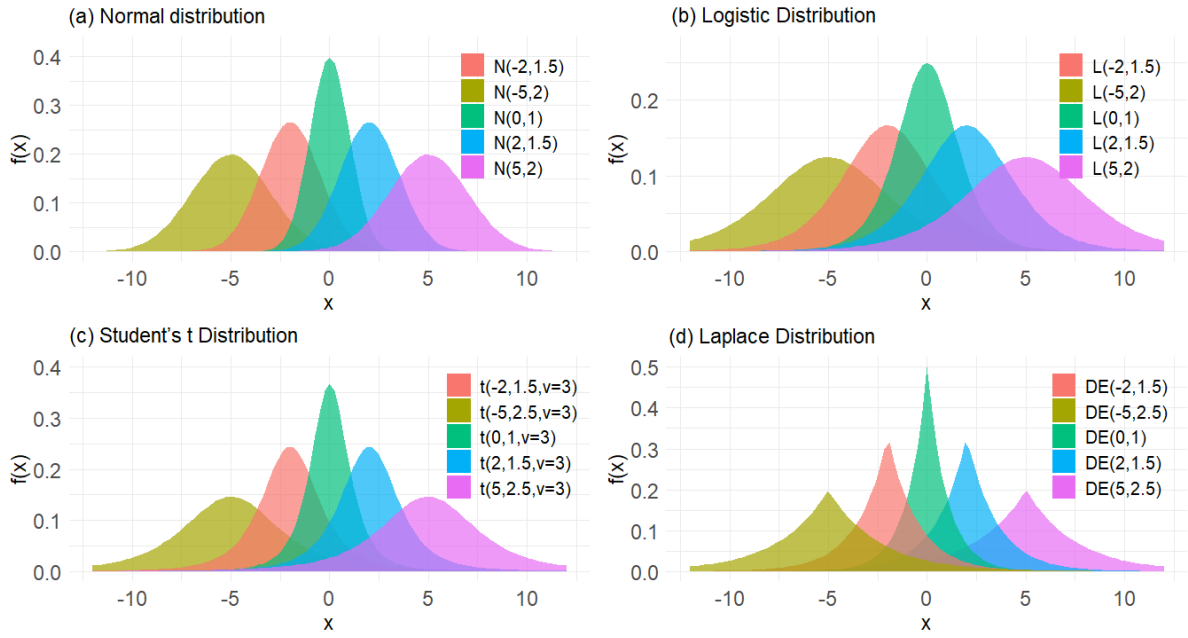

Figure 1 - Normal, Logistic, Laplace and Student's t model densities and their respective probability density functions, varying the location (μ), scale (*s*) and degrees of freedom (*v*) parameters for the t error.

the PFCD (Posterior Full Conditional Distribution) of each parameter depends on the individual posterior distribution.

The likelihood function is the product of the PDF (probability density function) of Y_i that has S distribution:

$$L(y_i,\alpha,\beta,\gamma,s,v) = \prod_{i=1}^{n} \frac{1}{s} g\left(\frac{y_i - f(\alpha,\beta,\gamma)}{s}\right)^2 \quad (2)$$

With regard to the determination of prior distributions, the precision parameter is defined as $\tau = \frac{1}{s} \sim U(10^{-1},10^2)$ for models (b), (c), (d), and $v \sim U(2,10)^2$ for case t. All the choices involving uniform priors are based on Laplace's principle of insufficient reason. The PFCD of σ^2, for the normal S case is known and has inverse Gamma distribution:

$$\sigma^2|y_i \sim GI\left(\frac{n}{2}+a, \frac{1}{2}\sum\left(y_i - f(x_i,\alpha,\beta,\gamma,\tau,v)\right)^2\right) \quad (3)$$

Where $\tau = \frac{1}{\sigma^2} \sim Gamma(a,b)$ and $n = 4$.

In order to implement empirical prior distribution for the vector (α, β, γ) and σ^2 for the normal case, Gauss-Newton method estimates were obtained for the Logistic and Gompertz models of 50 accessions from the database (only 50 of the 89 accessions are sigmoid), which are used to obtain the histograms and densities of each parameter estimate, and thereby determine the adequate candidate distribution. After probability density distributions were determined, the hyper parameters were obtained by equaling the mean and variance of these distributions to the mean and sample variance, respectively.

$$\begin{cases} E(X) = n^{-1}\sum_{i=1}^{n} \widehat{\theta}_i \\ V(X) = S_{\widehat{\theta}}^2 \end{cases} \quad (4)$$

Figure 2 shows the results of this methodology for the 50 database accessions analyzed in the present study and the results computed were used as empirical prior.

Heindenberg-Welch and Geweke criteria were used to analyze the convergence of MCMC chains, selecting nIter (number of iterations), nBurnin (values disregarded in the initial iterations of the chain) and nThin (jump values) values are selected to meet the two criteria simultaneously.

After the nThin and nBurnin values are obtained, the Openbugs interface computes the DIC (Deviance Information Criterion) calculations to compare the models fit by the Bayesian methodology. The Bayes factor is a measure of plausibility,

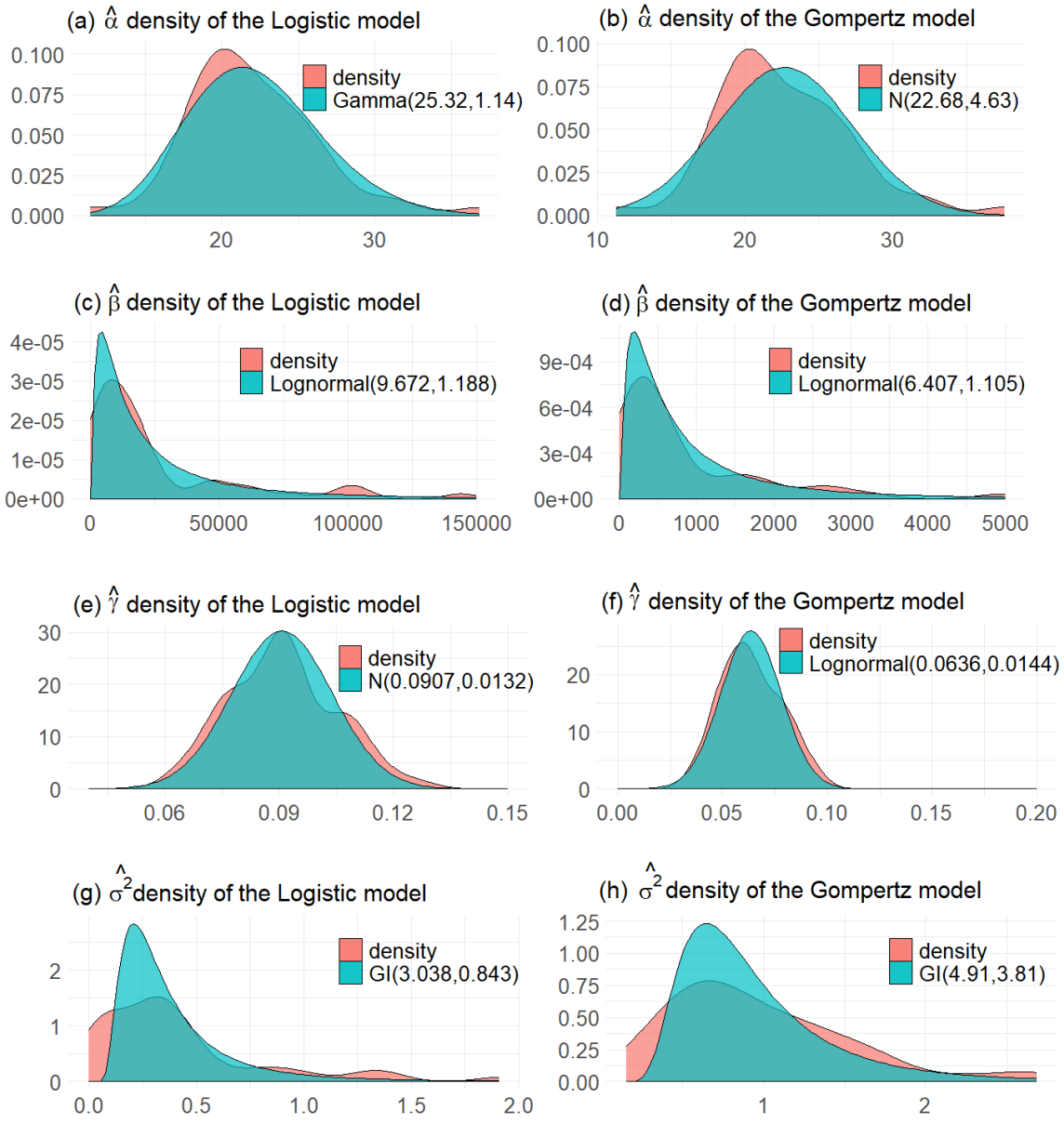

Figure 2 - Gauss-Newton estimate density - computed by the density function *()* of ggplot2 – of the parameters (a, β, γ, σ²) of the Logistic and Gompertz models and their empirical prior determined by equation 4.

given that DIC may be affected by posteriors that are bimodal or asymmetrical. The Bayes factor is defined as the ratio of marginal posteriors BF = p (y, M_1) p^{-1} (y, M 2), with M_i being the nth model to be compared, and the value of each p (y) is calculated by the harmonic mean:

$$\hat{p}(\boldsymbol{y}) = \left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{f(\boldsymbol{y}|\boldsymbol{\theta}^{(g)})}\right]^{-1} \quad (5)$$

Where G are the prior values generated, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$.

RESULTS AND DISCUSSION

Table 1 shows the usual estimators of the parameter posteriors. According to the DIC comparison criterion, the Logistic (6.94) and Gompertz (4.94) models exhibited the lowest

Table 1 - Estimate of parameters (α, β, γ, τ, *ν*) followed by the usual mean estimators, HPD credible intervals with lower (LL) and upper limit (UL), Bayes factor (BF) and Deviance Information Criterion (DIC).

Error	---Model---	------θ------	---Média---	-----LL-----	-----UL-----	----DIC----	-----BF-----
Normal Distribution	Logistic	α	22.7	21.47	24.14	8.21	1.2454
		β	39390	4184	100780		
		γ	0.1	0.09	0.11		
		τ	2.85	0.52	5.75		
	Gompertz	α	22.74	21.97	23.54	8.51	
		β	1462	192.99	3540		
		γ	0.07	0.06	0.09		
		τ	1.34	0.42	2.41		
Logistic Distribution	Logistic	α	22.7	20.91	24.65	9.84	0.9607
		β	38460	1620	97766		
		γ	0.1	0.08	0.12		
		τ	2.76	0.54	5.71		
	Gompertz	α	22.76	22.03	23.54	7.71	
		β	1513	104.89	3258		
		γ	0.07	0.06	0.09		
		τ	2.78	0.53	5.46		
Laplace Distribution	Logistic	α	22.7	21.07	24.75	9.38	1.7727
		β	39530	2296	97022		
		γ	0.1	0.08	0.11		
		τ	2.28	0.21	5.02		
	Gompertz	α	22.74	22.08	23.54	7.23	
		β	1603	193.41	3377		
		γ	0.07	0.06	0.09		
		τ	2.4	0.35	5.11		
Student's t Distribution	Logistic	α	22.72	21.34	24.27	6.94	1.2374
		β	41580	1890	100625		
		γ	0.1	0.09	0.11		
		τ	4.91	0.1	13.31		
		ν	48.71	2.05	94.82		
	Gompertz	α	22.75	22.03	23.47	4.94	
		β	1573	213.39	3399		
		γ	0.07	0.06	0.09		
		τ	5.16	0.1	14.54		
		ν	46.16	2	94.4		

values when the student's t-test is considered. In line with Bayesian modeling of the Cordona growth curve created by ROSSI & SANTOS (2014), the student's t-test has a smaller DIC when compared to the normal error. In all the scenarios, except for the normal error, the Gompertz model obtained lower DICs than those of its Logistic counterpart. According to the criterion of the Bayes factor in all the error distribution scenarios, there was no evidence that either the Logistic or Gompertz model is more plausible, and the values obtained were less than 2. The high β estimates obtained in the Logistic and Gompertz models in the present study were also reported by MACEDO et al. (2017), who analyzed the dry matter accumulation of garlic using frequentist and Bayesian regression.

Table 1 and figure 3 show the (α, β, γ) posteriors from the Logistic and Gompertz models, exhibiting similar graphs and values when the normal - graphs (a), (b) and (c); Logistic – graphs (e), (f) and (g); Laplace – graphs (i), (j) and (k); and student's t – graphs (m), (n) and (o) errors are considered. This demonstrated that alternating error distribution had little influence on obtaining the usual estimators and posterior distributions.

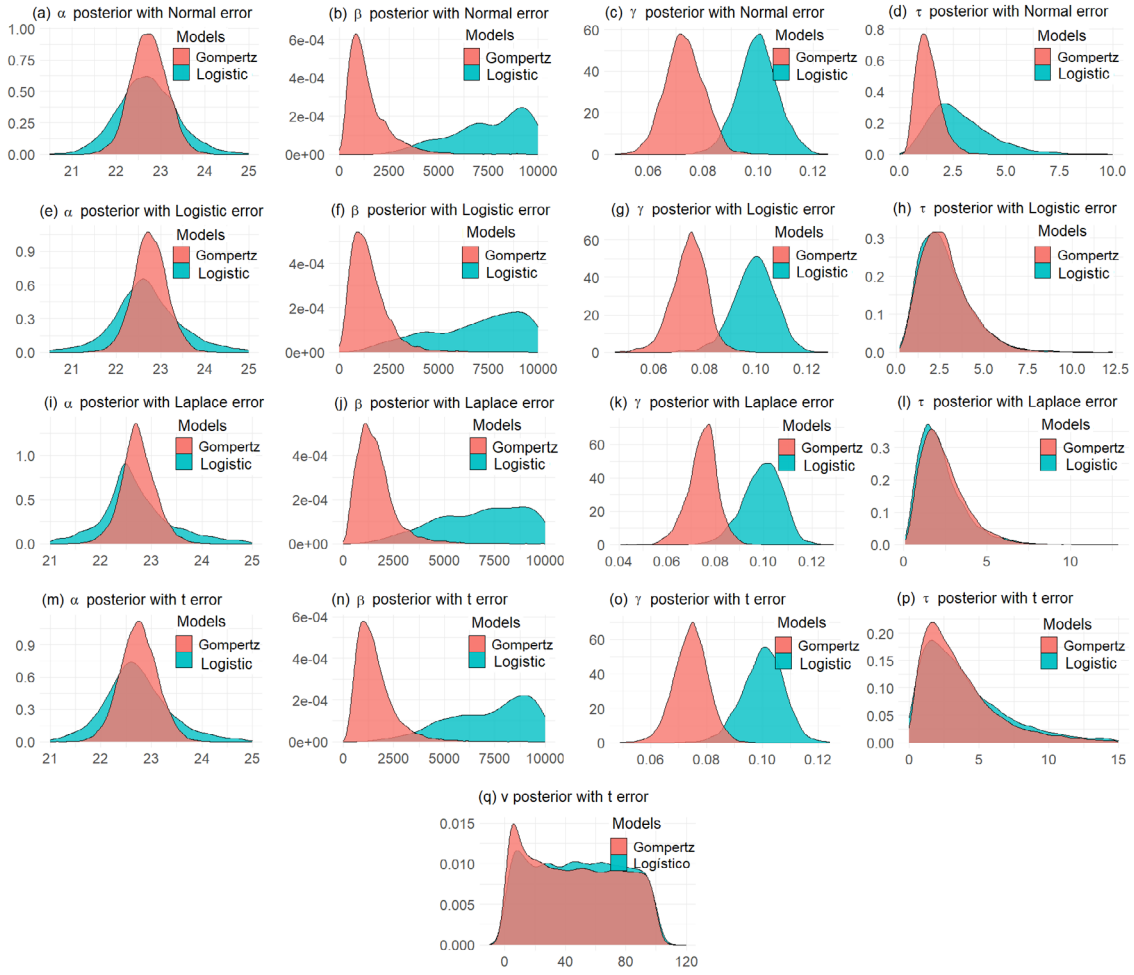

Figure 3 - Posterior parameter distribution (a, β, γ, τ) of Logistic and Gompertz models fit to access 63 considering an error with normal, Logistic, Laplace and student's t distribution.

In the graph (q) of Figure 3, the posterior density of *v,* in the case of the student's t error, exhibited a uniform trend, indicating prior dominance as a function of likelihood, which does not occur in (h), (l) and (p), whose graphs are more informative. Similar to the study of MARTINS FILHO et al. (2008), the (α, β, γ) posteriors showed a uniform and more informative trend respectively, in the Bayesian growth modeling of the "neguinho" and "carioca" bean cultivars when these consider a uniform prior.

In computational terms, the MCMC iterative process and convergence analysis required some computational time, as explained by PEREIRA et al. (2022). Figure 4 shows the nIter values of Markov chains that each model needed in the 4 error scenarios. In all the scenarios, except the Gompertz model with normal error, nThin was less than 20.

The Gompertz model with normal error needed an nThin of 1100 to control the high self-correlation of its chains, which contributed to the nIter of $4 \cdot 10^6$. When considering the Logistic error, the same Gompertz model needed an nThin and nIter of 20 and 40,000, respectively, which reveals computational economy. All the chains created in this process passed the Geweke and Heidenberg-Welch tests.

The results of Bayes factor, DIC and fitted graphs (a) and (b) of Figure 5 show no difference in the use of 4 garlic plant growth modeling errors. It was

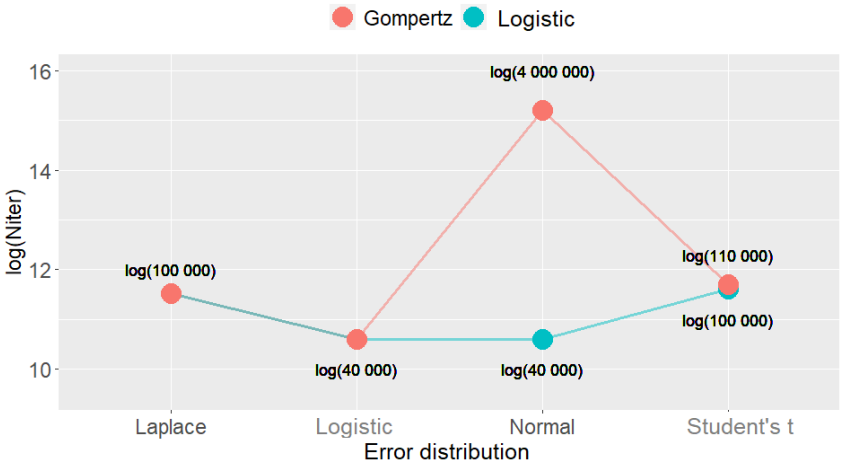

Figure 4 - Log graph of the number of iterations (niter) of the Logistic and Gompertz models considering the Laplace, Logistic, normal and student's t distributions.

concluded that alternating between the 4 symmetrical distributions for the error significantly alters the nIter values and, as such, it is up to the researcher to select the error with the highest computational economy.

CONCLUSION

There was no difference in the use of the normal, Logistic, Laplace and student's t errors for

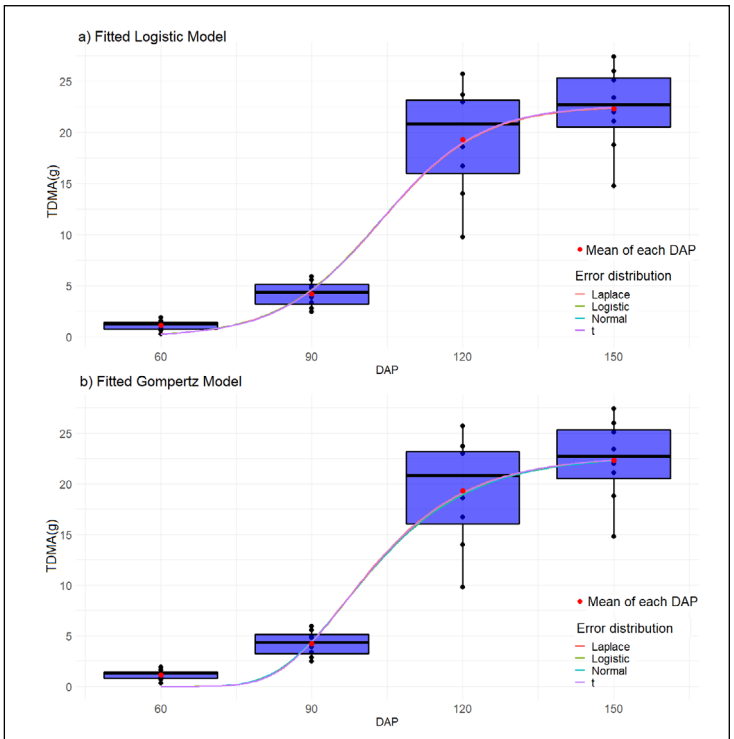

Figure 5 - Logistic and Gompertz models fitted to the mean of each DAP, considering the error with normal, Logistic, Laplace and Student's t distribution.

the experimental error in the Bayesian nonlinear modeling of garlic using the Logistic and Gompertz models. There are significant differences in the size of MCMC iterations for each error distribution.

ACKNOWLEDGEMENTS

The present study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil - Finance code 001.

DECLARATION OF CONFLICT OF INTERESTS

We have no conflict of interest to declare.

AUTHORS' CONTRIBUTIONS

The authors contributed equally to the manuscript.

REFERENCES

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. **Modelos lineares generalizados e extensões**. Pernambuco, 2008.

CORDEIRO, G. M.; PAULA, G. A. Improved likelihood ratio statistics for exponential family nonlinear models**. Biometrika**, v.76, n.1, p.93-100, 1989. Available from: <https://doi.org/10.1093/biomet/76.1.93>. Accessed: Feb. 10, 2022. doi: 10.1093/biomet/76.1.93.

CARLIN, B. P.; LOUIS, T. A. **Bayesian Methods for Data Analysis**. 2 ed. United States: CRC Press, 2008, 246p.

CORDEIRO, G. M. et al. Corrected maximum-likelihood estimation in a class of symmetric nonlinear regression models. **Statistics & Probability Letters**, v.46, p.317-328, 2000. Available from: <https://doi.org/10.1016/S0167-7152(99)00118-2>. Accessed: Feb. 10, 2022. doi: 10.1016/S0167-7152(99)00118-2.

DE LA CRUZ, R.; BRANCO, M. Bayesian analysis for nonlinear regression model under skewed errors, with application in growth curves. **Biometrical journal**, v.51, p.588-609, 2009. Available from: <http://dx.doi.org/10.1002/bimj.200800154>. Accessed: Feb. 10, 2022. doi: 10.1002/bimj.200800154.

FIRAT, MZ. et al. Bayesian analysis for the comparison of Nonlinear Regression Model Parameters: an Application to the Growth of Japanese Quail. **Revista Brasileira de Ciência Avícola**, v.18, p.19-26, 2016. Available from: <https://doi.org/10.1590/1806-9061-2015-0066>. Accessed: Nov. 10, 2022. doi: 10.1590/1806-9061-2015-0066.

GUJARATI, D. M.; PORTER D. C. **Econometria básica**. 5ed. Rio de Janeiro: AMGH ltda, 2012, 924p.

MACEDO, L. R. et al. Bayesian inference for the fitting of dry matter accumulation curves in garlic plants. **Pesquisa Agropecuária Brasileira**, v.52, n.8, p.572-581, 2017. Available from: <https://doi.org/10.1590/S0100-204X2017000800002>. Accessed: Nov. 10, 2022. doi: 10.1590/S0100-204X2017000800002.

MARTINS FILHO, S. et al. Abordagem Bayesiana das curvas de crescimento de duas cultivares de feijoeiro. **Ciência Rural**, Santa Maria, v.38, n.6, p.1516-1521, 2008. Available from: <https://doi.org/10.1590/S0103-84782008000600004>. Accessed: Nov. 10, 2022. doi: 10.1590/S0103-84782008000600004.

NELDER, J. A.; WEDDERBURN. R. W. M. Generalized linear Models. **Journal of the Royal Statistical Society. Series A (General)**. v.135, n.3, p.370–84, 1972. Available from: <https://doi.org/10.2307/2344614>. Accessed: Nov. 10, 2022. doi: 10.2307/2344614.

PEREIRA, A. A. et al. Bayesian modeling of the coffee tree growth curve. **Ciência Rural**, v.52, n.9, 2022. Available from: <https://doi.org/10.1590/0103-8478cr20210275>. Accessed: Nov. 10, 2022. doi: 10.1590/0103-8478cr20210275.

RATKOWSKY, D. A. **Nonlinear regression modeling**. New York: Dekker, 1983.

ROSSI, R. M.; SANTOS, L. A. Bayesian modeling growth curves for quail assuming skewness in errors. **Semina: Ciências Agrárias**, v.35. p.1637, 2014. Accessed: Nov. 10, 2022. Available from: <https://doi.org/10.5433/1679-0359.2014v35n3p1637>. doi: 10.5433/1679-0359.2014v35n3p1637.

SOUZA, G. S. **Introdução aos modelos de Regressão Linear e não linear**. Brasília: Embrapa, 1998. 480p.

STEWART, J. **Cálculo volume 2**. São Paulo: Cengage Learning, 2016. 672p.