Revista Ciência Agronômica, v. 51, n. 1, e20196716, 2020
Centro de Ciências Agrárias - Universidade Federal do Ceará, Fortaleza, CE
www.ccarevista.ufc.br

Scientific Article
ISSN 1806-6690

Statistical model assumptions achieved by linear models: classics and generalized mixed[1]

Pressuposições do modelo estatístico atendidas por modelos lineares: clássicos e generalizados mistos

Rita Carolina de Melo[2]*, Nicole Trevisani[2], Marcio dos Santos[2], Altamir Frederico Guidolin[2] and Jefferson Luís Meirelles Coimbra[2]

ABSTRACT - When an agricultural experiment is completed and the data about the response variable is available, it is necessary to perform an analysis of variance. However, the hypothesis testing of this analysis shows validity only if the assumptions of the statistical model are ensured. When such assumptions are violated, procedures must be applied to remedy the problem. The present study aimed to compare and investigate how the assumptions of the statistical model can be achieved by classical linear model and generalized linear mixed model, as well as their impact on the hypothesis test of the analysis of variance. The data used in this study was obtained from a genetic breeding program on the cooking time of segregating populations. The following solutions were proposed: *i)* Classical linear model with data transformation and *ii)* Generalized linear mixed models. The assumptions of normality and homogeneity were tested by Shapiro-Wilk and Levene, respectively. Both models were able to achieve the assumptions of the statistical model with direct impact on the hypothesis testing. The data transformations were effective in stabilizing the variance. However, several inappropriate transformations can be misapplied and meet the assumptions, which would distort the hypothesis test. The generalized linear mixed models may require more knowledge about the identification of lines of programming, compared to the classical method. However, besides the separation of fixed from random effects, they allow for the specification of the type of distribution of the response variable and the structuring of the residues.

Key words: Analysis of variance. Homogeneity of variance. Normality of errors. Crop breeding. Generalized linear mixed models.

RESUMO - Naturalmente quando concluído a condução de um experimento agrícola e estando disponível o dado coletado referente a uma variável resposta, deve ser procedida a análise de variância. Entretanto, os testes de hipóteses desta análise revelam validade somente se as pressuposições do modelo estatístico forem asseguradas. Quando tais pressuposições são violadas devem ser aplicados procedimentos com o propósito de remediar este problema. O objetivo deste trabalho foi comparar como as pressuposições do modelo estatístico podem ser logradas por métodos "clássico" e "contemporâneo" e seus reflexos sobre o teste de hipótese da análise de variância. Dados oriundos de um programa de melhoramento genético, referentes ao tempo de cocção de populações segregantes foram utilizados. Foram propostas como soluções: *i)* Modelos lineares clássicos com transformação de dados e *ii)* Modelos lineares generalizados mistos. As pressuposições de normalidade e homogeneidade foram testadas por Shapiro-Wilk e Levene, respectivamente. Ambos os métodos foram capazes de lograr as pressuposições do modelo estatístico com impacto direto nos testes de hipóteses. As transformações de dados foram eficazes em estabilizar a variância. Porém, inúmeras transformações não apropriadas, podem ser aplicadas indevidamente e atender as pressuposições, causando distorções no teste de hipótese. Os modelos lineares generalizados mistos podem exigir maior conhecimento na identificação de linhas de programação, se comparado ao método clássico, mas permitem além da separação dos efeitos fixos e aleatórios, a especificação do tipo de distribuição da variável resposta e a estruturação dos resíduos.

Palavras-chave: Análise de variância. Homogeneidade de variância. Normalidade dos erros. Melhoramento de plantas. Modelos lineares generalizados mistos.

DOI: 10.5935/1806-6690.20200015
*Author for correspondence
Received for publication 01/05/2019; approved on 08/10/2019
[1]Parte da Dissertação de Mestrado da primeira autora apresentada na Universidade do Estado de Santa Catarina/UDESC
[2]Programa de Pós-Graduação em Produção Vegetal, Instituto de Melhoramento e Genética Molecular/IMEGEM, Universidade do Estado de Santa Catarina/UDESC, Lages-SC, Brasil, rita_carol_mel@hotmail.com (ORCID ID 0000-0002-5710-7621), nicoletrevisani88@gmail.com (ORCID ID 0000-0003-4583-8125), mdsantos182@hotmail.com (ORCID ID 0000-0001-9282-1057), altamirguidolin@gmail.com (ORCID ID 0000-0003-3028-0958), coimbrajefferson@gmail.com (ORCID ID 0000-0001-9492-6055)

INTRODUCTION

One of the goals of agricultural experiments is obtaining inferences using hypothesis tests. The statistical models present assumptions that need to be met in order ensure the validity of the inferences (BOX; COX, 1964; JUPITER, 2017). Among these assumptions, the homogeneity of variances is considered the most critical one, since the violation of any of the other assumptions of the analysis of variance can affect this assumption (STEEL; TORRIE; DICKEY, 1997). The normality of the residues plays also an important role (ZANARDO *et al.*, 2010). However, in many practical situations, the response variable is obtained from counting data (CUSTÓDIO; BARBIN, 2009). Thus, approximate distributions are used in most cases (SILVA, 2003).

If the assumptions of normality and homogeneity are not met, the conclusions obtained by the statistical analyses may lead to serious mistakes (LÚCIO *et al.*, 2012; XU; LI; SONG, 2013). Under non-normality and heterogeneity conditions, the levels of significance and the sensitivity of the *F* test may be affected. In general, it has been argued that inferences derived from variance analysis, particularly F tests, are relatively robust to small deviations from normality. These statements are based on data analysis studies generated from simulated experiments. Although these tests tolerate slight deviation from the assumptions, important discrepant deviations should be corrected (HOEKSTRA; KIERS; JOHNSON, 2012).

Several methods have been described in order to meet the assumptions of the statistical model. In the classical linear model: *i)* exploratory data analysis for visualization and removal of discrepant values (BUSTOS, 1988); *ii)* non-parametric tests (SPRENT; SMEETON, 2007) and *iii)* data transformation, the most used methods in research (STEEL; TORRIE; DICKEY, 1997; XU, LI; SONG, 2013) and will be covered in this study. The Generalized Linear Mixed Models allows the option to choose non-normally distributed response variable and the structure of residual variances and covariances, which might improve the fitness of the model (MAIA *et al*., 2013; WOLFINGER; O'CONNELL, 1993).

Comparatively, the impacts of the classical linear model and generalized linear mixed model on hypothesis testing have been little discussed. There are almost no guidelines about which model may be the most appropriate for a particular situation. The present work aims to compare and investigate how the assumptions of the statistical model can be achieved by classical linear model and generalized linear mixed model and their impacts on the results obtained in the hypothesis test of the analysis of variance.

MATERIALS AND METHODS

Example applied: Cooking time in fixed and segregating common bean populations

To exemplify the models proposed in this article, we used data referring to genetic constitutions from a complete diallel between the genotypes of *Phaseolus vugaris* L. BAF07, BAF09, BAF50 and IPR Uirapuru and their reciprocals. The characteristics of the parents involved in the hybridization are described in Table 1.

Fixed F_1 populations from hybridization gave rise to segregating populations in the generations F_2, F_3, F_4, F_5, F_6, F_7, F_8 and F_9, through successive self-fertilizations. From these genetic constitutions, 12 populations were conducted in a field trial:

i) BAF50 x BAF07 and BAF09 x IPR Uirapuru in the generation $F_{1\text{-}2}$;

ii) BAF50 x BAF07 and BAF09 x IPR Uirapuru in the generation $F_{2\text{-}3}$;

iii) BAF50 x BAF07 and BAF09 x IPR Uirapuru in the generation $F_{7\text{-}8}$;

iv) BAF50 x BAF07 and BAF09 x IPR Uirapuru in the generation $F_{8\text{-}9}$;

v) Parents BAF07, BAF09, BAF50 and IPR Uirapuru.

The segregant populations were conducted in Lages, State of Santa Catarina, Brazil (27º48' S, and 50º19' W, altitude 930 m asl). According to Koppen classification, climate was temperate cfb (moist mesothermal and mild summer). The experiment was conducted in a randomized block design with two replicates per treatment, which resulted in 24 experimental units. Each experimental unit was composed of four one-meter lines, with density of 10 seeds per linear meter.

The assessment of the cooking time was performed after the assay was harvested. The grains were dried in an oven until they reached 12% moisture. The Mattson cooker, which consists of 25 vertical stems of 90 g each, and a 2 mm diameter tip, was used to evaluate the cooking time, according to the method adapted by Proctor and Watts (1987). Since the methodology used shows an intrinsic variation significative, two samples were collected within each experimental unit, which resulted in 48 observations for the cooking time (ALMEIDA *et al.*, 2011).

Therefore, the following experimental statistical model was proposed:

$$y_{ijk} = \mu + b_i + pop_j + pop(block)_{ij} + pop(block*rep)_{ijk}$$

Table 1 - Characteristics of the parents involved in the hybridization

Parents	Group	Growth habit	Cooking time (minutes)
BAF07	Black	Type III	26.17
BAF09	Black	Type III	23.85
BAF50	Carioca	Type III	23.66
IPR Uirapuru	Black	Type II	20.00

Where: y_{ij} refers to the variable cooking time; μ, to the effect associated with the general mean; b_i, to the effect associated with the *i*-th block level; pop_j, to the effect associated with the *j*-th population level; $pop(block)_{ij}$ to the effect of the experimental error and $pop(block*rep)_{ijk}$, to the effect associated with the sampling error (information obtained from the evaluation of the replications within the experimental units).

Statistical analysis

The data were submitted to the analysis of variance, considering the classical linear models, using the GLM procedure. Initially, the data were submitted to the analysis of variance (as experimentally obtained). The assumption of normality of the errors was verified by the Kolmogorov-Smirnov test (α=0.05) since the experiment has ≥ 30 observations (n=48). While the homogeneity of variance was verified by the Levene test (α=0.05) since it is a more indicated test, in case of violation of the normality assumption of the errors. When the assumptions were violated, the following remedies were independently applied:

i) Classical linear model: the data on the cooking time were transformed by four empirical formulas (SOKAL; ROHLF, 1995; STEEL; TORRIE; DICKEY, 1997): Square root = $\sqrt{y}$; Angular = $sim^{-1}\sqrt{y^{+3/8}/n^{+3/4}}$; Logarithmic = ln y; Logitic = $1og_{10}$ (y/1+y). Where *y* is the response variable (cooking time).

After transformation, the normality assumptions of the errors were assessed by the 'Kolmogorov-Smirnov test' (α=0.05) and the homogeneity of variance, by the Levene test (α=0.05), to verify if the transformation was efficient in adapting the response variable to the violated assumption. The afore mentioned analyses were performed in the PROC GLM.

ii) Generalized linear mixed model: The data were submitted to the analysis of variance with the use of the Generalized Linear Mixed Models. The GLIMMIX procedure was used for this purpose, since, in addition to separating the fixed from the random effects (MODEL and RANDOM option) and specifying the distributions of the response variable most used (DIST option), it allows structuring residual variances and co-variances (TYPE option). In this analysis, two categories of models were considered:

i) Linear model (C_1 e C_2): C_1 - consider the statistical procedure change (PROC GLM para PROC GLIMMIX) for the separation of fixed and random effects and C_2 - a structure of residual covariance and variance was inserted in G of autoregressive type of order 1 (AR(1));

ii) Logarithmic models (C_3 e C_4): C_3 - specification of the response variable distribution and C_4 - were inserted variance and covariance structure in G (random effects) and the distribution specification.

In the four proposed models, the appropriate residue (error between) due to nested effect "population (block)" was specified in the RANDOM command. After the construction of the models for the representation of the observations, it was used the minimization of information criteria of the restricted maximum likelihood, with the Akaike criterion for the selection of the most appropriate model: AIC = - 2 log (maximum likelihood) + 2 (number of independently adjusted parameters). The homogeneity of variances was verified by the specification of the COVTEST HOMOGENEITY command.

The classical and generalized mixed models were compared by the results obtained in the homogeneity and normality tests, in addition to the hypothesis test of the analysis of variance. In the classical model, the quality estimators of the model were obtained: $CV = s/\bar{x}*100$; R^2 = *SQregressão/SQtotal*. In the contemporary analysis (generalized linear mixed models), it was also observed the selection criterion of the model for the purpose of comparison. When the most appropriate method was identified, tests of multiple comparisons of means were obtained by Scheffe at 5% probability of error. All the analyses and statistical procedures described were performed using the SAS (Statistical Analysis System).

RESULTS AND DISCUSSION

Common Analysis of variance

The analysis of variance performed with the original observations of the response variable showed no

significant effect for the controlled factors (Table 2). The population effect did not show significant differences for the cooking time. The ratio found for the errors between and within showed a significant effect, which demonstrates that the variation within the plots - or between the replicates of the method - is a significant portion of the mean squared error (Table 2). Thus, using the total error (error between plus error within) inferences not related to the population effect can be made. In other words, the cooking method has intrinsic variation and, as already explained, its residues must be partitioned in the analysis of variance (ALMEIDA *et al.*, 2011).

The results obtained in the analysis of variance should be carefully analyzed. It is worth to observe two important points: *i*) in the breeding of autogamous plants, populations with a high level of heterozygosis are expected to reveal significant differences, which helps breeders selecting (GINKEL; ORTIZ, 2018; MELO *et al.*, 2017). Another issue, contrary to the above, is that *ii)* if differences between treatments were observed, they would certainly be related to the experimental error, since analyses with specifications of inappropriate models provide less reliable results and cause impacts on the inferences derived from the test (ALMEIDA *et al.*, 2011). This is due to the fact that the results of an assay are affected by both the activity from the treatments and the variations that experimenters do not control, which tend to mask the effects of the treatments (COCHRAN, 1947).

The *F* test should be sensitive or powerful, which means that it should detect the presence of real differences as frequently as possible (COCHRAN, 1947). In plant breeding, this is due to the detection of genetic differences to the detriment of uncontrolled variation. However, the inferences obtained from this test can only be valid if they come from a linear model whose premises of normality and homogeneity are met (ZANARDO *et al.*, 2010). According to Table 2, the normality test was highly significant (p=0.0100). It indicates that the errors do not follow a normal distribution. Besides, the homogeneity test of variances indicated the lack of common residue between the populations (p=0.0269).

Out of the four assumptions of the analysis of variance, normality is the least likely to be valid (GHASEMI; ZAHEDIASL, 2012). Discrete response variables, for example, do not probably follow this assumption. Besides, continuous variables that express weight or height of individuals are restricted to positive values. Thus, they do not represent the expected symmetrical distribution for normality. Therefore, such an assumption can only be approximately verified. Furthermore, assumptions of homogeneity of variance and normal distribution of errors are often simultaneously violated. This means that if the distribution is not normal, the variance is not homogeneous. The reciprocal is true (SILVA, 2003).

The effects of non-normality may decrease the efficiency in the estimation of the treatment effects. Similar effect occurs for heterogeneity of variances, where different rates of error can be detected among the treatments (COCHRAN, 1947). Under such conditions, when the assumptions of the statistical model have been violated, it is necessary to adopt some procedure that can at least reasonably achieve these premises.

Classical linear model with the use of transformations

Data transformation may be indicated for cases with some heterogeneity of variances arising from a relationship between mean and variance (STEEL; TORRIE; DICKEY, 1997). Thus, an appropriate transformation - determined on the basis of this relation - can lead to the stabilization of variance and consequently an approximation of the normal distribution (RIBEIRO-OLIVEIRA *et al.*, 2018). The intriguing question is "which transformation should be employed?". For the

Table 2 - Analysis of variance for the trait cooking time from fixed and segregant populations of beans, considering the total error and its decomposition into error between and error within

Sources of variation		Degrees of freedom	Mean square
Block		1	2.78^{ns}
Population		11	62.78^{ns}
Error between (e)		11	26.41
Error within (d)		24	10.76
Total Error		35	15.67
e/d Ratio		11	2.45*
a D = 0.1723*	F = 2.34*	$R^2 = 0.55$	CV = 15.69

*Significant at 5% probability of error. ^{ns} Non-significant at 5% probability of error. ^{a} D=Kolmogorov-Smirnov test and F=Levene test. R^2 = coefficient of determination. CV = coefficient of variation

purpose of demonstration and discuss, four distinct transformations were performed according to certain distributions. Their effects on the analysis of variance and its respective hypothesis test were verified, as well as the assessment of the assumptions of normality and variance homogeneity (Table 3).

There are different effects on the hypothesis tests of the main factors when the different types of transformation are considered (Table 3). The value of *F* for the population factor obtained around 36% of variation between the transformations. This variation caused a significant impact on the real probability values. Besides, three transformations presented significant differences between the levels of the population factor. Out of the four transformations, two fully satisfied the assumptions tested (Angular and Logitical) and revealed high coefficient of determination (0.67 and 0.66).

A transformation of the response variable usually achieves the assumptions of the statistical model, since it generally aims to change the measurement scale (SILVA, 2003). When the variance tends to change as the mean of the treatments changes, the variance will only be stabilized by an appropriate change in scale (BARTLETT, 1947; BOX; COX, 1964). It is evident that more than one transformation can reveal equivalent statistical results - such as the logistic and the angular (Table 2). For novice researchers, it may seem like "playing with their data" to obtain the desired response, which leads to distortions in the hypothesis test and compromises the actual estimate of the parameter estimator. Therefore, researchers must purposely use a certain type of transformation to the detriment of the inappropriate ones. For example, logistic transformation is suitable for experiments with population growth. However, angular transformation is suitable for experiments whose response variable are proportions of individuals.

It is known that the criterion for choosing a transformation method that is better known and used by researchers is based on knowledge of the distributive aspects of the response variable, empirically or theoretically determining the relationship between variance and mean (BOX; COX, 1964). For example, square root transformation is appropriate for analyses with counting data, whose data often follow a Poisson distribution (distribution whose response variable displays the counting of individuals with small values, $\sigma^2=m$). In contrast, if the data are transformed by a logarithmic scale (if $x = \ln y$), y is said to reveal a Lognormal distribution (distribution whose response variable shows the counting of individuals with high and much variable values, $\sigma=m$) (STEEL; TORRIE; DICKEY, 1997). Thus, the data used in this work (grain cooking time in minutes) were not obtained from counting measurements. So, the application of an angular transformation is not appropriate, for example, even if it has met the assumption of homogeneity of variances and normality of errors.

The transformation method is considered a classical methodology widely used in the scientific community. Several studies have detected improvements in the fulfillment of assumptions after the use of some data transformation. The comparison of treatments for the dose-response curve of agricultural pesticides, after a logarithmic transformation (MANIKANDAN, 2010) exemplifies it. The researchers Liang *et al*. (2015) found that both normality and homogeneity assumptions were improved by the logarithmic transformation of EC 50 values of fungicides. They revealed significant differences between products after the transformation. Other studies, however, raise a hypothesis: "Has the time for discarding the transformation methods arrived?" (WILSON *et al.*, 2010). These authors performed a review on the use of angular transformation in proportion data and the use of logistic regressions (based on generalized linear models) in several high impact journals and concluded that the use of the latter method as opposed to a general linear model improved waste quality by 50% compared to the effect of angular transformation. In addition, the angular

Table 3 - Analysis of variance for the trait cooking time from fixed and segregant bean populations after the Square, Angular, Logarithmic and Logitic Root transformation and their respective indicated distributions. Probability of the Kolmogorov-Smirnov (D) test for normality and Levene (F) test for homogeneity of variances

Transformation	Relevant Distribution	Analysis of variance				Assumptions	
		Block	Population	R^2	CV	Pr>D	Pr>F
Square Root	Poisson	0.08^{ns}	2.70^{ns}	0.59	6.98	0.0363	0.0287
Angular	Binomial	0.15^{ns}	4.20*	0.67	6.29	0.1500	0.1732
Logarithmic	Lognormal	0.06^{ns}	3.04*	0.62	3.90	0.0806	0.0345
Logitic	Empirical	0.04^{ns}	3.61*	0.66	-10.38	0.1500	0.0851

*Significant at 5% probability of error. ^{ns} Non-significant at 5% probability of error. R^2=coefficient of determination. CV=coefficient of variation

transformation changed the final decision of significance in only 5% of the data, while the logistic regression changed the decision in 33%.

Although data transformation may pose some hindrance to the selection of the most appropriate type, it has proved to be one of the most widely used and accessible methods in scientific research. Generally, if a transformation can be identified with the ability of stabilizing the variance, it brings several practical advantages to the analysis, mainly simplicity and efficiency (PIEPHO, 2009). Our study identified the logarithmic transformation (appropriate transformation) metting the homogeneity of variances, as well as approximating the distribution of these data to a normal distribution.

Generalized linear mixed model

The authors of this article have described the generalized linear mixed models. For this, four distinct models were considered, which presume or not the structuring of the errors (linear models) and the specification of a distribution (logarithmic models). Regardless of the proposed model, the hypothesis of homogeneity of variances for both the block effect and the population effect was accepted. The mere separation of the fixed effects from the random effects allowed purifying the rates of errors between the treatments. In addition, it was possible to verify that the correct specification of the model allowed for the penalization of the number of parameters. Besides, a reduced value was obtained by the Akaike criterion (AIC), which indicates that the assumption of normality of residues (Table 4) has better met the assumption. The generalized linear mixed models provide adjustments to the statistical model that isolate disturbing extraneous characteristics. This allows for the reasonable achievement of the assumptions (PIEPHO, 2009).

The hypothesis tests and the quality of the model adjustment also deserve attention. There are clear differences between the four adjusted models for the hypothesis tests of the fixed effects, and the values obtained for the model selection criteria (Table 4). Considering the linear models, no differences were found between the levels of the block and population factors in the C_1 model, even though the homogeneity were satisfied. This was also found for model C_2, which added the structure of residual variances and covariances of autoregressive type of order 1 - AR (1). It explains the absence of a temporal or spatial dependence on the facts considered, such that the inclusion of this structure did not change the results obtained between the linear models C_1 and C_2.

Conversely, it was possible to verify changes of great magnitude in the logarithmic models (Table 4). The model C_3 has undergone changes in either the hypothesis tests or the relative magnitude of the selection criteria. In this model, differences between the levels of the population factor were detected. The selection criteria also showed lower relative value (AIC=-26.11), a desirable fact for the selection of the most appropriate model. This fact indicates that, unlike the structure of residual variances and covariances, the specification of a known distribution allows for reducing the effect of strange factors on the statistical model.

The distribution specified in models C_3 and C_4 was determined according to distribution tests and the respective verification of the AIC value. The lognormal obtained the lowest AIC value, among all the possible distributions to be specified in the GLIMMIX procedure. Therefore, it was appropriate to explain the effects of the cooking time. This fact agreed with the results obtained in the model. The lognormal distribution is appropriate to express the counting of individuals with large values. This distribution is often used to characterize response variables time-related responses, which is true for the present case (SILVA, 2003).

Similarly, to C_3, the C_4 model revealed significant differences between the levels of the population factor. The selection criteria and *F* values were similar or identical between the logarithmic models, which confirms

Table 4 - Analysis of variance with generalized linear mixed models for fixed effects, considering linear and logarithmic models. Selection of the model based on the Akaike criterion for each adjusted model

Sources of variation	Linear Models		Logarithmic Models	
	C_1	C_2	C_3	C_4
Block	0.39^{ns}	0.39^{ns}	0.78^{ns}	0.78^{ns}
Population	2.37^{ns}	2.37^{ns}	3.21*	3.21*
Akaike Criterion	212.39	236.39	-26.11	-2.11
Homogeneity test ($p > \chi^2$)	0.4356	0.9734	0.6200	0.9932

*Significant at 5% probability of error. ^{ns} Non-significant at 5% probability of error. C_1: Simple model. C_2: Model with structuring of errors. C_3: Model with distribution specification. C_4: Model with error structuring and distribution specification

that the residual structure imposed in the C_4 model does not represent significant changes for this particular test (AIC=-2.11). However, in trials with a residual structure, the structuring of errors may lead to a more appropriate model.

Undoubtedly, an advantage of the generalized mixed linear model is the ability of simultaneously structuring residues and specifying distributions (WOLFINGER; O'CONNELL, 1993), which was not possible in the method previously described (transformation). In agricultural experiments, or more specifically in works aimed at genetic breeding, there may be a relationship between the observations and the errors conditioned to these observations. It is also worth mentioning that, in these very tests, the unbalance of information is often observed, which may change the components of the variance and the hypothesis tests (DUARTE; VENCOVSKY, 2001). The generalized mixed linear model also contributes in this regard, adequately adjusting the degrees of freedom of the residue. However, it is worth mentioning that the use of the generalized mixed linear model requires more time and computer resources from researchers. Analyses performed with mixed models may require hours or days, especially when matrices of residual variances and covariance are structured. Besides, the method requires users to identify the correct specification of each command, according to the condition of the experiment. Otherwise, poorly identified models may lead to wrong conclusions.

According to model C_3, multiple comparisons of means were performed by Scheffe (Table 5). The means test shows significant differences between population 6 (BAF50 x BAF07 F_2) and the other populations under study. In terms of genetic breeding, these differences can be capitalized for the selection of superior genotypes, since the F_2 generation is regarded to present the greatest genetic variability. This variability was amplified due to hybridization between parents with distinct cooking times (BAF50 with 23.66' x BAF07 with 26.17'). It is worth mentioning that these differences are reliable, since they were revealed by a model with high fit quality. Thus, it is important to verify the assumptions of the statistical model and use the appropriate remedies to meet them before making any inference based on parametric tests.

The advantages of the mixed model have already been corroborated by other works on different knowledge areas, including biological sciences (WILSON *et al.*, 2010), soil sciences (LARK; CORSTANJE, 2009) or even social sciences (LO; ANDREWS, 2015). Other works, however, still use traditional approaches or classical methods that employ data transformations to the detriment of generalized linear mixed models, arguing that the traditional approach provides for robust statistical tests in a wide range of conditions, unlike the contemporary methodologies, which may lead to erroneous conclusions if the model is poorly specified (IVES, 2015). Piepho (2009) argues that it is worth exploring different data transformations before using a more complex mixed model analysis. Some transformations may stabilize the variance in a simpler way and should be employed, instead of a mixed model.

Data transformation can cause distortions if the change in the measurement scale is not appropriate. However, irrefutably, it proves to be a simple method, with minimal detrimental effects. Therefore, the efforts

Table 5 - Estimation of the difference between the means of the 12 populations from the Scheffe multiple comparisons test. Comparisons of means considering the logarithmic model C_3 (specification of the distribution of the response variable dist=lognormal)

i/j	1	2	3	4	5	6	7	8	9	10	11	12
1		0.09	0.10	0.27	-0.22	-0.16	-0.01	0.04	0.07	0.17	0.22	0.15
2			0.01	0.18	-0.31	-0.25*	-0.10	-0.05	-0.02	0.08	0.12	0.06
3				0.16	-0.32	-0.26*	-0.11	-0.06	-0.03	0.07	0.11	0.05
4					-0.49	-0.42*	-0.28*	-0.23	-0.20	-0.09	-0.05	-0.11
5						0.06	0.21	0.26	0.29	0.39	0.44	0.37
6							0.14	0.20*	0.23*	0.33*	0.37*	0.31*
7								0.05	0.08	0.19	0.23*	0.17*
8									0.03	0.13	0.18*	0.11
9										0.11	0.15	0.08
10											0.04	-0.02
11												-0.06

*Significant at 5% error probability (Pr > |t|)

with the adoption of more contemporary methodologies are not worthwhile. The attributes of the generalized linear mixed models were demonstrated, namely, the structuring of residual matrices and specification of different distributions, simultaneously. The use of one method by breeders to the detriment of another, in heteroscedastic and non-normal models, may be conditioned to the knowledge of the distribution of the response variable and the need to consider other attributes, such as the correlation of the residues in the experiments. Thus, the assumptions of the analysis of variance can be tested in a more efficient and simple form.

CONCLUSIONS

Both models were able to achieve the assumptions of the statistical model. Data transformation is a methodology more accessible than the mixed model but should be used with caution. When researchers have the necessary time and computer ability, they can use mixed models that simultaneously allow for the specification of the type of distribution of the response variable and the structuring of the residues.

REFERENCES

ALMEIDA, C. B. *et al*. Existe variabilidade para o caráter tempo de cocção em feijão? Depende do erro! **Bioscience Journal**, v. 27, n. 6, p. 915-923, 2011.

BARTLETT, M. S. The use of transformations. **Biometrics**, v. 3, n. 1, p. 39-52, 1947.

BOX, G. E. P.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society**, v. 26, p. 211-252, 1964.

BUSTOS, O. Outliers e robustez. **Revista Brasileira de Estatística**, v. 49, p. 7-30, 1988.

COCHRAN, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. **Biometrics**, v. 3, n. 1, p. 22-38, 1947.

CUSTÓDIO, T. N.; BARBIN, D. Modelos de predição para sobrevivência de plantas de *Eucalyptus grandis*. **Ciência e Agrotecnologia**, v. 33, p. 1948-1952, 2009.

DUARTE, J. B.; VENCOVSKY, R. Estimação e predição por modelo linear misto com ênfase na ordenação de médias de tratamentos genéticos. **Scientia Agricola**, v. 58, n. 1, p. 109-117, 2001.

GHASEMI, A.; ZAHEDIASL, S. Normality tests for statistical analysis: a guide for non-statisticians. **International Journal of Endocrinology and Metabolism**, v. 10, n. 2, p. 486-489, 2012.

GINKEL, M.; ORTIZ, R. Cross the best with the best, and select the best: HELP in breeding selfing crops. **Crop Science**, v. 58, n. 1, p. 17-30, 2018.

HOEKSTRA, R.; KIERS, H. A. L.; JOHNSON, A. Are assumptions of well-known statistical techniques checked, and why (not)? **Frontiers in Psychology**, v. 3, p. 1-9, 2012.

IVES, A. R. For testing the significance of regression coefficients, go ahead and log-transform count data. **Methods in Ecology and Evolution**, v. 6, n. 7, p. 828-835, 2015.

JUPITER, D. C. Assumptions of statistical tests: what lies beneath. **The Journal of Foot & Ankle Surgery**, v. 56, p. 910-913, 2017.

LARK, R. M.; CORSTANJE, R. Non-homogeneity of variance components from spatially nested sampling of the soil. **European Journal of Soil Science**, v. 60, n. 3, p. 443-452, 2009.

LIANG, H. J. *et al*. Logarithmic transformation is essential for statistical analysis of fungicide EC50 values. **Journal of Phytopathology**, v. 163, n. 6, p. 456-464, 2015.

LO, S.; ANDREWS, S. To transform or not to transform: using generalized linear mixed models to analyse reaction time data. **Frontiers in Psychology**, v. 6, p. 1171, 2015.

LÚCIO, A. D. *et al*. Violação dos pressupostos do modelo matemático e transformação de dados. **Horticultura Brasileira**, v. 30, n. 3, p. 415-423, 2012.

MAIA, E. *et al*. Aplicação da análise espacial na avaliação de experimentos de seleção de clones de laranjeira Pera. **Ciência rural**, v. 43, n. 1, p. 8-14, 2013.

MANIKANDAN, S. Data transformation. **Journal of Pharmacology & Pharmacotherapeutics**, v. 1, n. 2, p. 126-127, 2010.

MELO, R. C. de *et al*. Heterozygosity level and its relationship with genetic variability mechanisms in beans. **Revista Ciência Agronômica**, v. 48, n. 3, p. 480-486, 2017.

PIEPHO, H. P. Data transformation in statistical analysis of field trials with changing treatment variance. **Agronomy Journal**, v. 101, n. 4, p. 865-869, 2009.

PROCTOR, J. R.; WATTS, B. M. Development of a modified Mattson bean cooker procedure based on sensory panel cookability evaluation. **Food Research International**, v. 20, p. 9-14, 1987.

RIBEIRO-OLIVEIRA, J. P. *et al*. Data transformation: an underestimated tool by inappropriate use. **Acta Scientiarum. Agronomy**, v. 40, p. 1-11, 2018.

SILVA, J. G. C. **Análise estatística de experimentos**. Pelotas: Universidade Federal de Pelotas, 2003.

SOKAL, R. R.; ROHLF, F. J. **Biometry**: the principles and practice of statistics in biological research. 3. ed. New York: W. H. Freeman, 1995. 880 p.

SPRENT, P.; SMEETON, N. C. **Applied non parametric statistical methods**. 4. ed. Boca Raton: Chapman; Hall, 2007. 542 p.

STEEL, R. G. D.; TORRIE, J. H.; DICKEY, D. A. **Principles and procedures of statistics**: a biometrical approach. 3. ed. New York: McGraw-Hill, 1997. 672 p.

WILSON, E. *et al*. **The arcsine transformation**: has the time come for retirement? Canada: Memorial University of Newfoundland, Newfoundland and Labrador, 2010. Disponível em: http://www.mun.ca/biology/dschneider/b7932/B7932Final10Dec2010.pdf. Acesso em: 20 dez. 2018.

WOLFINGER, R.; O'CONNELL, M. Generalized linear mixed models: a pseudo-likelihood approach. **Journal of Statistical Computation and Simulation**, v. 4, p. 233-243, 1993.

XU, W.; LI, W.; SONG, D. Testing normality in mixed models using a transformation method. **Statistical Papers**, v. 54, p. 71-84, 2013.

ZANARDO, B. *et al*. Posições das mudas de alface nas bandejas de poliestireno e efeitos na normalidade e homogeneidade dos erros na produção de plantas. **Revista Ciência Agronômica**, v. 41, n. 2, p. 285-293, 2010.

This is an open-access article distributed under the terms of the Creative Commons Attribution License