Generalized Lineal Models for the Analysis of Binary Data from Propagation Experiments of Brazilian Orchids

Freddy Mora¹, Letícia de Menezes Gonçalves², Carlos Alberto Scapim², Elias Nunes Martins³ and Maria de Fátima Pires da Silva Machado⁴
¹Universidade Estadual de Maringá; Av. Colombo, 5790; Bloco 05; fmora@universiabrasil.net; 87020-900; Maringá - Paraná - Brasil. ²Departamento de Agronomia, Universidade Estadual de Maringá; Maringá - PR - Brasil. ³Departamento de Zootecnia; Universidade Estadual de Maringá; Maringá - PR - Brasil. ⁴Departamento de Biologia Celular e Genética; Universidade Estadual de Maringá; Maringá - PR - Brasil

ABSTRACT

This study aimed at applying the generalized linear models (GLM) for the analysis of a germination experiment of Cattleya bicolor in which the response variable was binary. The purpose of this experiment was to assess the effects of the storage temperatures and culture mediums on the seed viability. The analyses of variance was also carried out either with or without the data transformation. All the statistical approaches indicated the importance of the storage temperature on the seed viability. But, the culture media and interaction effects were significant only by the GLM. Based on the GLM, the seeds stored at 10°C increased viability, in which the coconut medium achieved the best performance. The results emphasized the importance of adopting the GLM to improve the reliability in many situations where the response variable followed a non-normal distribution.

Key words: Cattleya bicolor, deviance, in vitro culture, logit model, maximum-likelihood

INTRODUCTION

Orchid plants are generally cultivated by the growers for the commercial ornamental aims in different parts of the world; they are one of the most important and appreciated ornamental plants, which are reaching a high commercial value (Faria et al., 2001; Oliveira and Sajo, 1999). Cattleya bicolor L. is a Brazilian native tropical orchid growing naturally in the central eastern Brazil. This orchid is considered vulnerable species due to several factors, i.e. over-collection and habitat destruction. In vitro seed propagation system used in many tropical orchid species is widely considered an effective technique for attending ex situ conservation programs (Stenberg and Kane, 1998; Gangaprasad et al., 1999; Buyun et al., 2004). In vitro non-symbiotic seed propagation of the orchids represents also an ecological and commercial relevant procedure. The plants produced by the non-symbiotic procedures are very useful for re-introducing the native orchid species in the preservation areas (Martini et al., 2001). Several experiments have shown how the culture conditions are specific for the genus and sometimes for the species within Orchidaceae family (Arditti and Ernst, 1993).

According to Raghavan (2003), the first application of the embryo culture technique came from the work of Knudson, in 1922, which
succeeded in germinating the orchid embryos with the non-symbiotic medium and growing them on a nutrient agar medium containing sucrose. Statistically, discrete data sets are commonly recorded in the studies of in vitro germination (Shiau et al., 2002; Bhadra and Hossian 2003; Buyun et al., 2004; Damon et al., 2005); proportion of viable seeds is one typical example of this. The studies where the response variable is either “success” or “failure” (i.e., 0 or 1) are fairly common in nearly all areas of the science (Myers et al., 2002). The problem with the analyses of the variables representing count, proportion or binary data, is the possibility of the violation of one or more assumptions of the analysis of the variance, and then, affecting the result of the study (Sokal and Rohlf, 2003).

In the germination studies where the variable response follows a non-normal distribution, three inference procedures are frequently assumed: first, in vitro seed analysis by using a non-parametric procedure (Droste et al., 2005), second, the analysis of the germination by employing the generalized linear models (Clauss and Venable, 2000; Prati and Bossdorf, 2004; Willenborg et al., 2005), and third, germination assessment by transforming of the response variable, making the distribution of the response closer to the normal distribution (McKendrick et al., 2000; Reddy, 2000; Moravcová et al., 2002; Walck et al., 2002; Hawkes, 2004). Generally, the transformations are also used for stabilizing the response variance and improving the fit of the model to the data (Myers et al., 2002).

The generalized linear model methodology allows that the response probability distribution be any member of an exponential family of distributions using the methods closely analogous to the normal linear methods for the normal data (Myers et al., 2002; Nelder and Wedderburn, 1972). The generalized linear model may be viewed as a unification of the linear and non-linear regression models that incorporate a rich family of the normal and non-normal response distributions (Myers et al., 2002).

The objectives of this study were to apply the generalized linear models (GLM) for the analysis of a germination experiment of Cattleya bicolor in which the response variable was binary; to assess the effects of the storage temperatures and culture mediums on in vitro germination of C. bicolor; to confirm that this statistical procedure could contribute to analyze the experiments either involving binary data or any distribution that was a member of the exponential family.

MATERIAL AND METHODS

In vitro non-symbiotic germination procedures

The seed lots of C. bicolor were stored at 10±2 and 25±2 °C over the silica gel for approximately two years. In order to determine and verify the seeds viability before the storage, 2, 3, 5-triphenyltetrazolium chloride (TTC) histochemical procedure was used. This method was carried out according to Deswal and Chand (1997). The TTC reduction is frequently applied as a quantitative method for evaluating the tissue viability. The intensity and extent of the TTC staining were successfully employed to predict the germination percentage of a lot of seeds (Chang et al., 1999). Subsequently, the seed lots were transferred to different culture medium:

1) Nutritive Knudson or C medium: This medium is used for germinating the seeds of the other orchid species and usually employed as the basal medium (Martini et al., 2001; Droste et al., 2005).

2) Coconut water (150 ml L⁻¹) containing 20 g L⁻¹ of sucrose, and 6.5 g L⁻¹ agar (pH adjusted to 5.3). The cultures were maintained in the germination chambers for 20 days under continuous fluorescent at 25±2 °C. Coconut water was from the Cocos nucifera L. palm trees growing in the botanical garden of the university. This is energy rich and contains a variety of substances such as vitamins, hormones, amino acids and lipids (Arditti and Ghani, 2000). Twenty days after the seed inoculations the germoplasms were transferred for the TTC analysis. The experiment had a completely randomized design considering the factorial scheme, with two factors: culture medium and storage procedure (factorial 2x2), and four replications, with a variable number of seeds (34 to 243) per plot.

Conventional statistical analysis outlook

Two analyses were carried out in order to revise the conventional statistical view. First, the analysis of the variance considering the percentage of the viable seeds, and second, the analysis of variance with the transformed percentage by the formula:

\[ Y_{tr} = ar \sin e^{\frac{Y}{100}} - 1 \]
Where \( Y \) was the original response variable (in percentage) and \( Y_{tr} \) was the transformed variable. In both ways, Shapiro-Wilk statistics (W) and plot examination were used in order to assess the data normality and residual distribution performance, by the means of PROC UNIVARIATE, NORMALTEST and PLOT options, in SAS (SAS, Institute 1996; Carrão-Panizzi et al., 2002 and 2004; Rodríguez et al., 2006). Each analysis of the variance was carried out by the Analysis of Variance (PROC ANOVA) of SAS (SAS, Institute 1996), regarding the following linear model:

\[ y = X\beta + \epsilon \]

Where \( y \) was the observations vector (either percentage or transformed data); \( X \) was the incidence matrix of fixed effects; \( \beta \) was the fixed effect vector due to the culture medium, temperature of the storage and their interaction; \( \epsilon \) was to residual effect. The inference procedures for the conventional analysis assumed that the response variable \( y \) had a univariate normal distribution, with means \( X\beta \), and variance:

\[ V = \text{Var}(y) = \text{Var}(X\beta) + \text{Var}(\epsilon) = R = \sigma^2 I \]

A generalized linear models approach

Each seed from each plot was considered as the experimental unit and the response, \( Y \), took only one of two possible values, recorded as either 0 or 1. This was given by the TTC histochemical procedure for the viability of the seeds, where a value of 1 was recorded if hydrogenation was achieved, i.e. altering the original color and a value of 0 otherwise. Then, the probabilities of failure (0) and success (1) were given by:

\[ P(Y_i = 0) = 1 - \pi_i \]
\[ P(Y_i = 1) = \pi_i \]

Thus, the random variable \( Y \) followed a binomial distribution, \( Y \sim B(m, \pi) \), with probability density function given by:

\[ f(y; \pi) = \binom{m}{y} \pi^y (1-\pi)^{m-y} \quad y = 0,1,...,m \]

Consequently, the log likelihood function could be represented as:

\[ l(\pi; y) = \sum_{i=1}^{n} y_i \log \left( \frac{\pi_i}{1-\pi_i} \right) + n_i \log(1-\pi_i) \]

Where \( n \) was the number of the observed values of the independent random variable.

A generalized linear model (GLM) was used for the data set that followed the binomial distribution. The method of maximum likelihood was applied to the GLM as the theoretical basis for the parameter estimation. Maximum likelihood estimates was obtained by solving the systems of score equations for the parameters (Myers et al., 2002). The logit link function was used for connecting the linear predictor to the natural mean of the response variable (Myers et al., 2002) by the following function:

\[ g(\pi_i) = \eta_i = \log(\pi_i / 1-\pi_i) \]

GENMOD procedure of the SAS software (SAS-Institute, 1996) was used to fit the generalized linear model by specifying the following general statements in the editor of the SAS package:

PROC GENMOD data=orchid order=data descending;
CLASS Medium Storage;
MODEL response=Storage Medium Storage*Medium / dist=binomial link=logit type1 type3 covb lrci;
CONTRAST 'Medium differences' Medium -1 1;
CONTRAST 'Storage differences' Storage -1 1;
LSMEANS Medium Storage Storage*Medium / diff;
RUN;

Note that the binomial distribution and the link function were specified with the options “distribution” (i.e. dist=binomial) and “Link” (i.e. link=logit), respectively, and in the model statement, type I and type III statistical options were used.

RESULTS AND DISCUSSION

The analyses of variance from the original data (in percentage) and transformed data (by using the formula 1) are showed in Table 1. At first, in the conventional case, it is important to mention that the residual effect has not more than 12 degrees of
freedom, a fact that should be improved by adding more repetitions into the experiment for reaching better statistical reliability. An important unifying concept underlying the analysis of variance is relative to the probability distribution of the response data in which the normality plays a central role (Myers et al., 2002). In the current study, the original and transformed data sets did not satisfy to this supposition which was judged by visually inspecting a normal-plot and using Shapiro-Wilk statistic (p < 0.05) (Rodríguez et al., 2006). The data transformation procedure was not sufficiently effective to making the distribution of the response variable closer to the normal distribution. Ignoring the normal hypothesis, it could be observed from the Table 1 that the temperatures of storage showed a significant effect on the seeds viability in both analysis (p < 0.01). On the other hand, these approaches did not detect statistical significance (p > 0.05) for the culture medium and interaction effects. In both the analyses, the coefficients of variation were low, ranging from 5 to 6%, showing absence of over-dispersion.

Table 1 – Analyses of variance from the original and transformed data.

| Source     | DF | Original data | Transformed data | Source     | DF | Original data | Transformed data |
|------------|----|---------------|------------------|------------|----|---------------|------------------|
|            |    | Mean Square   | F-Value          |            |    | Mean Square   | F-Value          |
| Storage (S)| 1  | 514.1556      | 24.61            | 0.0003     | 0.1185 | 22.89        | 0.0004           |
| Medium (M) | 1  | 62.0156       | 2.97             | 0.1106     | 0.0194 | 3.74         | 0.0769           |
| M x S      | 1  | 8.2656        | 0.40             | 0.5412     | 0.0068 | 1.31         | 0.2740           |
| Residual   | 12 | 20.8948       |                  |            |       |              |                  |
| Mean       |    | 85.17         | 1.19             |            |       |              |                  |
| Coefficient of variation (%) | 5.4 | 6.1 |

DF: degrees of freedom

Table 2 showed that the coconut water medium was superior to Knudson medium but, as indicated previously, this difference was not sufficiently large to achieve statistical significance by taking out the normal assumption in the conventional statistical analysis. As a result, the seeds stored at 10°C increased the viability by 11% and the storage effect did not depend on the medium effect, because the medium x storage interaction was not significant.

| Effect            | Medium  | Storage (°C) | Estimate (%) |
|-------------------|---------|--------------|--------------|
| Medium            | Knudson | 25 ± 2       | 79.50        |
| Medium            | Coconut | 10 ± 2       | 90.84        |
| Storage           | 25 ± 2  |              |              |
| Storage           | 10 ± 2  |              |              |
| Medium x Storage  | Knudson | 25 ± 2       | 78.25        |
| Medium x Storage  | Knudson | 10 ± 2       | 88.15        |
| Medium x Storage  | Coconut | 25 ± 2       | 80.75        |
| Medium x Storage  | Coconut | 10 ± 2       | 93.53        |

The statistical analysis by the generalized lineal model approach achieved the convergence for the maximum likelihood estimate. According to Webb et al. (2004), the existence of a maximum likelihood estimate depended on the concavity of the log likelihood function. However, the concavity of the log likelihood function alone did not imply that the maximum likelihood always existed.

Table 2 - Germination success (%) estimate per effect. Just storage effect evidenced significant difference (p<0.01) determined by the analysis of variance from the original and transformed data.
The GENMOD procedure fit a generalized linear model as defined by Nelder and Wedderburn (1972). The class of the generalized linear models was an extension of the traditional linear models that allowed the mean of a population to depend on a linear predictor through a nonlinear link function and allowed the response probability distribution to be any member of an exponential family of the distributions (SAS-Institute, 1996). The binomial, Poisson, negative binomial, normal, geometric, exponential, gamma, and inverse normal distributions are members of this family and there are many possible choices of the link function for each model. In the current study, note that many other useful statistical models can be formulated as the generalized linear models by the selection of another link function, i.e. probit, complementary log-log, or normit. Logit link function is the canonical link for the generalized linear model with binomial distribution; also, it is known as natural link function. Canonical or natural links are usually preferred due to their statistical theoretical properties (Myers et al., 2002); the use of the canonical link simplifies the arithmetic greatly.

In the present study, contrary to the conventional analysis carried out by ignoring the normality assumptions, the evaluation based on the generalized linear models with a response variable that followed binomial distribution indicated the statistical significance for the three effects: storage temperature (p<0.01), culture medium (p<0.01 and p<0.05 for Type III and Type I analysis, respectively) and culture medium x storage temperature interaction (p<0.05) (Table 3). Willenborg et al. (2005) reported that the data transformations sometimes presents the statistical limitations and, traditionally, the main approach used by the agronomists and crop scientists to account for non-normality has been this procedure.

| Source   | Deviance | DF | Type I Chi-Square | p > χ^2   | Type III Chi-Square | p > χ^2   |
|----------|----------|----|-------------------|----------|---------------------|----------|
| Intercept| 1442.2962|     |                   |          |                     |          |
| Storage (S)| 1404.3566| 1  | 37.94             | < 0.0001 | 45.63               | < 0.0001 |
| Medium (M)| 1399.0211| 1  | 5.34              | 0.0209   | 9.50                | 0.0021   |
| M x S    | 1393.7261| 1  | 5.30              | 0.0214   | 5.30                | 0.0214   |

DF: Degree of freedom.

Similar results between Type I and Type III analysis were evidenced (Table 3). Medium in Type I analysis was significant to 2% of probability while Type III to 0.2%. This difference was due to the fact that the Type I analysis fit a sequence of models beginning with null model or intercept term, and then continuing throughout the model with one additional effect on each step: storage, medium and interaction. However, the Type III analysis did not depend on the order in which the terms for the model was specified (Spyrides-Cunha et al., 2000).

Table 4 shows that summarize the fit of the specified model. These data were useful to review the efficiency of the model, considering the culture medium, storage temperature and interaction effects. The mean deviance (D=1393.7261) divided by the degree of freedom (DF=1650) was inferior to 1.0, but not much. According to Myers et al. (2002), this suggested that under or over dispersion was not likely to be a problem, as the ratio was close to 1.0, or similarly a deviance that was approximately equal to its degrees of freedom represented a possible indication of a good model. Consequently, the specified model performed adequately.

The relative means (original scale) of Medium-Storage interaction effect are given in Table 5. The interaction effect was also caused by the magnitudes differences. For example, the seeds stored at 10°C achieved a better performance than 25°C, in both the culture media, but in the coconut medium the difference between the storage temperatures effects was higher. This could also be viewed in Table 2 by the means of germination success estimates. At 25°C, Knudson and coconut media were not different statically (p > 0.05), unlike storage at 10 °C in which the coconut medium was higher (p < 0.01). Although the coconut appeared advantageous medium compared
to Knudson, this generally presented certain variability across the time and/or local conditions that should also be considered in an in vitro seed propagation program.

Table 4 - Results of the criteria for assessing goodness of fit for the generalized linear model.

| Criterion             | DF  | Value       | Value/DF |
|-----------------------|-----|-------------|----------|
| Deviance              | 1650| 1393.7261   | 0.8447   |
| Scaled Deviance       | 1650| 1393.7261   | 0.8447   |
| Pearson Chi-Square    | 1650| 1654.0000   | 1.0024   |
| Scaled Pearson \(\chi^2\) | 1650| 1654.0000   | 1.0024   |
| Log Likelihood        |     | -696.8630   |          |

DF: Degree of Freedom.

Table 5 - Relative means (original scale) of Medium-Storage interaction effect.

| Medium      | Knudson | Coconut |
|-------------|---------|---------|
| Temperature  | 25 ± 2  | 10 ± 2  |
| Means for temperature / Knudson | 0.78 | 0.88 * |
| Means for temperature / Coconut   | 0.80 | 0.95 * |
| Means for medium / 25°C           | 0.78 | 0.80 ns |
| Means for medium / 10°C           | 0.88 | 0.95 * |

*The pair of means within the same row differs significantly (p < 0.01), ns: the pair of means within the same row does not differ significantly (p > 0.05).

The lowest germination success showed by the medium C of Knudson had also been reported by Martini et al. (2001) studying the propagation in another native orchid (Gongora quinquenervis), which after 15 days of seed inoculations showed a rapid die of the seeds, characterized by the total necrosis of embryos. It was very important to note that the differences between the culture medium could only be seen by using a generalized linear models approach. The conventional analysis of the variance ignoring the normality assumptions did not show these differences either with or without the data transformation. Thus, due to the occurrence of a significant interaction effect, it was concluded that the coconut medium was obviously higher than the Knudson medium only when the seeds were stored at 10°C. These results emphasized the importance of adopting the generalized linear modeling approach, improving and optimizing ex situ biodiversity conservation methods of this Brazilian native orchid.

RESUMO

A técnica de propagação in vitro é considerada efetiva para fins comerciais e de conservação de orquídeas. A metodologia de modelos lineares generalizados (MLG) foi usada para analisar um experimento de germinação de Cattleya bicolor. O propósito do experimento foi avaliar os efeitos da temperatura de armazenamento e dos meios de cultivo sobre a germinação, cuja resposta foi considerada binária. Análise convencional com ou sem transformação de dados foram também realizados. Todas as abordagens estatísticas indicaram a importância da temperatura sobre a viabilidade das sementes. Entretanto, os efeitos de meios de cultivo e interação foram significativos apenas para MLG. As sementes armazenadas a 10°C incrementaram sua viabilidade, onde o meio a base de coco atingiu o melhor desempenho. Os resultados enfatizam a importância de adotar MLG, para melhorar a confiabilidade em situações onde a variável resposta segue uma distribuição distinta à normal.

REFERENCES

Arditti, J. and Ernst, R. (1993), Micropropagation of orchids. John Wiley and Sons Press, New York, pp. 682.

Arditti, J. and Ghani, A. K. A. (2000), Numerical and physical properties of orchid seeds and their biological implications. New Phytol., 145, 367-421.
Bhadra, S. K. and Hossain, M. M. (2003), *In vitro* germination and micropropagation of *Geodorum densiflorum* (Lam.) Schltr., an endangered orchid species. *Plant Tissue Cult.*, 13, 165-171.

Buyun, L.; Lavrentyeva, A.; Kovalska, L. and Ivannikov, R. (2004), *In vitro* germination of seeds of some rare tropical orchids. *Acta Universitatis Lodziensis, Biology*, 676, 159–162.

Carrão-Panizzi, M. C.; Goés-Favoni, S. P. and Kikuchi, Damon, A.; Pérez-Soriano, M. and Rivera, M. L. (2002), Extraction time for soybean isoflavone determination. *Braz. arch. biol. Technol.*, 45(4), 515-518.

Carrão-Panizzi, M. C.; Goés-Favoni, S. P. and Kikuchi, A. (2004), Hydrothermal treatments in the development of isoflavone aglycones in soybean (Glycine max (L.) Merrill) grains. *Braz. arch. biol. Technol.*, 47(2), 225-232.

Chang, W. C.; Chen, M. H. and Lee, T. M. (1999), 2,3,5-riphenyltetrazolium reduction in the viability assay of *Ulva fasciata* (Chlorophyta) in response to salinity stress. *Bot. Bull. Acad. Sin.*, 40, 207-212.

Clauss, M. J. and Venable, D. L. (2000) Seed Germination in Desert Annuals: An empirical test of adaptive bet hedging. *The American Naturalist*, 155, 168-186.

Damon, A.; Pérez-Soriano, M. and Rivera, M. L. (2005), Substrates and fertilization for the rustic development of *Corallorrhiza trifida* and characterization of its mycorrhizal fungi. *New Phytol.*, 145, 523-537.

Myers, R. H.; Montgomery, D.C. and Vining, G.G. (2002), Generalized linear models, with applications in engineering and the sciences. John Wiley and Sons Press, New York, pp.342.

Moravcová, L.; Zákhravský, P. and Hroudová, Z. (2002), Germination response to temperature and flooding of four Central European species of Bolboschoenus. *Preslia*, 74, 333–343.

Nelder, J. A. and Wedderburn, R. W. M. (1972), Generalized linear model. *Journal of the Royal Statistical Society A*, 35, 370-384.

Oliveira, V. C. and Sajo, M. G. (1999), Root anatomy of nine Orchidaceae species. *Braz. arch. biol. Technol.*, 42(4), 405 - 413.

Prati, D. and Bossdorf, O. (2004), Allelopathic inhibition of germination by *Alliaria petiolata* (Brassicaceae). *American Journal of Botany*, 91, 285-288.

Raghavan, V. (2003), One hundred years of zygotic embryo culture investigations. In *in vitro Cell. Dev. Biol. Plant*, 39, 437-442.

Reddy, K. N. (2000), Factors affecting *Campsis radicans* seed germination and seedling emergence. *Weed Science*, 48(2), 212–216.

Rodríguez, G. R., Pratta, G. R., Zorzoli, R. and Picardi, L.A. 2006. Evaluación de caracteres de planta y fruto en líneas recombinantes autofecundadas de tomate obtenidas por cruzamiento entre *Lycopersicon esculentum* y *L. pimpinellifolium*. *Cien. Inv. Agr.*, 33, 133-141.

SAS-Institute (1996), Statistical analysis system: user's guide. SAS Institute, Cary, pp. 956.

Shiau, Y. J.; Sagare, A. P.; Chen, U. C.; Yang, S. R. and Tsay, H. S. (2002), Conservation of *Anoectochilus formosanus* Hayata by artificial cross-pollination and *in vitro* culture of seeds. *Bot. Bull. Acad. Sin.*, 43, 123-130.

Sokal, R. R. and Rohlf, F. J. (2003), *Biometry*: the principles and practice of statistics in biological research. Third edition. Freeman and Company Press, New York, pp. 850.

Spyrides-Cunha, M. H.; Demetrio, C. G. B. and Camargo, L.E.A. (2000), Proportional odds model applied to mapping of disease resistance genes in plants. *Genetics and Molecular Biology*, 23(1), 223-227.

Stenberg, M. L. and Kane, M. E. (1998), *In vitro* seed germination and greenhouse cultivation of *Encyclia boothiana* var. erythronioides, an endangered Florida orchid. *Lindleyana*, 13, 101–112.

Walck, J.L.; Hidayati, S. N. and Okamami, N. (2002), Seed germination ecophysiology of the Asian species *Osmorhiza aristata* (Apiaceae); comparison with its North American congeners and implications for
Mora, F. et al.  

evolution of types of dormancy. *American Journal of Botany*, 89, 829-835.

Webb, M. C.; Wilson, J. R. and Chong, J. (2004), An analysis of quasi-complete binary data with logistic models: applications to alcohol abuse Data. *Journal of Data Science*, 2, 273-285.

Willenborg, C. J.; Wildeman, J. C.; Miller, A. K.; Rossnagel, B. G. and Shirtliffe, S. J. (2005), Oat germination characteristics differ among genotypes, seed sizes, and osmotic potentials. *Crop Science*, 45, 2023–2029.

Received: April 10, 2006;  
Revised: November 19, 2007;  
Accepted: May 21, 2008.