Mycorrhiza (1996) 6:197–200

ORIGINAL PAPER

Sonia A. Alvarez-Santiago · Felipe García-Oliva
Lucía Varela

Analysis of vesicular-arbuscular mycorrhizal colonization data with a logistic regression model

Accepted: 3 November 1995

Abstract Vesicular-arbuscular mycorrhizal (VAM) infection is usually expressed as percentage of root length colonized. The frequency distributions of the data are often non-normal and may follow a negative binomial distribution. Data transformation, such as an arcsin of percentage colonization, may be used to help colonization data satisfy the normal distribution assumption, but is not always successful. In this paper, we compare ANOVA and logistic regression model (LRM) analysis of data on the effect of phosphorus fertilization and corn cultivar on VAM colonization. Transformed data did not fit a normal distribution, and we propose the LRM as a better model for statistical analysis of VAM colonization. The LRM is more accurate because (1) this model assumes a binomial distribution, (2) it incorporates the original sample size into the probability estimation, and (3) the model uses non-transformed data, which are easier to interpret.

Key words Binomial distribution · Statistical analysis · ANOVA · Maize · P fertilization

Introduction

The clumped distribution of mycorrhizal fungi within roots makes VAM colonization data highly variable among replicate pots of similarly treated plants (Daft and Nicolson 1972; Barea et al. 1980; St. John and Hunt 1983). As a consequence, the data may approach a binomial distribution (St. John and Hunt 1983; Raju et al. 1990). The statistical analysis most often used is analysis of variance (ANOVA), which assumes a normal distribution of the data and homogeneity of variances. However, data with a skewed distribution (like percentage of colonization) often do not satisfy both assumptions and, for this reason, application of this model is invalid (St. John and Koske 1988).

Data transformation, e.g., arcsin of percentage colonization, may be used to make colonization data satisfy the above assumptions (Sokal and Rohlf 1981). However, transformed data may still not meet the assumption of normality, and it is necessary, therefore, to apply a goodness-of-fit test. Also, calculation of a percentage, obscures the size of the sample. This can lead to a reduction in the power of the analysis, because the probability estimation to reject the null hypothesis is affected by the size of the data used (degrees of freedom). An alternative method for analyzing VAM colonization data is the logistic regression model (LRM), which assumes a binomial distribution (Dobson 1983). In the present paper, we examine the differences in data interpretation arising from the use of ANOVA and LRM in the analysis of VAM colonization data on the effect of fertilization and cultivar on the colonization of maize (*Zea mays* L.) plants.

Materials and methods

The logistic regression model

The logistic regression model (LRM) is based on a binary response variable, while the explanatory variables are categorical (Aitkin et al. 1989). We define the response variable as:

$y=1$ "success" (occurrence of VAM colonization)
$y=0$ "failure" (non-occurrence)

and let p be the probability of success for a randomly chosen individual at given values of the explanatory variables (explanatory variables were fertilization levels and maize varieties in the present study). Then y has a Bernoulli distribution:

S. A. Alvarez-Santiago · L. Varela
Escuela Nacional de Ciencias Biológicas, IPN, AP 63-389, CP 02800, Mexico DF, Mexico

F. García-Oliva (✉)
Centro de Ecología, UNAM, AP 70-275, CP 04510, Mexico DF, Mexico
Fax: +52-5-616-1976
e-mail: fgarcia@miranda.ecologia.unam.mx

$$P(y)=p^{y}(1-p)^{1-y} \qquad y=0,1 \tag{1}$$

where

$P(y=1)=p$
$P(y=0)=1-p$

The distribution has a mean p and a variance $p(1-p)$. The Bernoulli distribution is a special case of the binomial distribution where the number of trials, n, is equal to 1 (Aitkin et al. 1989). Since the binomial distribution depends on both n and p, the number of trials n_i (total number of counted segments for each plant in our case) on which the observed number of successes (r_i) is based (occurrence of VAM colonization) must be specified as a second argument. As a consequence, each r_i is weighted by its respective n_i, and p can be estimated. A second characteristic of the binomial distribution is that the variance is not constant and changes with the mean. The variance is low when p is very high or low, and greatest when $p=q=0.5$. To avoid exceeding the natural bounds for p $(0,1)$, the logit link function or transformation is used, defined by

$$\alpha=\text{logit}\,p=\ln[p/(1-p)] \tag{2}$$

so that

$$p=e^{\alpha}/(1+e^{\alpha}) \tag{3}$$

In this way the natural bounds of p are converted from $(0,1)$ to $(-\infty,\infty)$.

Experimental test of the model

The data were obtained from a greenhouse experiment with two factors: three maize varieties (Criollo, H-34 and VS-22) and three levels of phosphorus fertilization (0, 16.6 and 33.2 μg P g^{-1}). The soil, collected from the foot-slope of the Malitzin volcano in the State of Tlaxcala, Mexico, was sandy-clay loam with a slightly to moderately acid pH. Total N remained low during the plant cycle (less than 0.07%), and available P was 30 μg g^{-1} at the onset of rains and decreased to 10 μg g^{-1} in the dry season (Gavito and Varela 1993). The pots were arranged in a completely randomized design with five replications, and plants were grown in a greenhouse maintained at 18–25 °C. Before seeding with maize, P solutions (16.6 and 33.2 μg P g^{-1}) were individually prepared and 10 ml of each solution was added to each pot. Distilled water (10 ml) was applied to non-fertilized pots as a control.

After 70 days of growth, roots were harvested and stained with trypan blue (Phillips and Hayman 1970). Approximately 400 segments 1 cm in length were cut from each stained root system, and spread out evenly in a plastic Petri dish with a grid (0.5-in. squares); VAM colonization was evaluated with the gridline intersect method (Giovannetti and Mosse 1980). The total number of intersections and the frequency of infected roots (mycelium, vesicles, arbuscules and/or spores) were recorded for each individual plant.

Percentage of colonization was calculated on the basis of the total number of segments examined and transformed to arcsin square root for a two-way factorial ANOVA. The model used had this expression:

$$y_{ij}=\mu+V_i+F_j+VF_{ij}+e_{ij} \tag{4}$$

where y_{ij} was percentage of colonization, μ was the overall mean, V_i was the cultivar effect (three varieties), F_j was the fertilization effect (three levels), VF_{ij} was the interaction effect, and e_{ij} was a random error component. The model was fitted by least squares, assuming a normal distribution of e_{ij} and homogeneity of variances. This model was used for calculation of fitted values, and residuals were estimated by subtraction of the observed from the fitted values. Frequency distributions of residuals were analyzed for normality by the Kruskal-Wallis test. Homogeneity of variances was assessed with Bartlett's test (Sokal and Rohlf 1981), and residual analysis was performed by plotting residual versus fitted values (Montgomery 1984). Multiple comparisons among means was made with Tukey's test.

The frequency of VAM colonization among the root segments was fitted to a LRM with two factors (variety and fertilization) using a generalized linear model with the GLIM statistical package (Anonymous 1985). The model was similar to Eq. 4, but the estimation of the terms was different. The model was fitted by maximum likelihood for the Bernoulli distribution, which was expressed as:

$$L(\beta)=\prod_{j=1}^{N} p_j^{rj}(1-p_j)^{nj-rj} \tag{5}$$

where r_j is the number of VAM colonized segments in each plant, n_j is the total number of segments in this plant, p_j is the probability of colonization, and N is the total number of plants. The random error had a binomial distribution, and the link function was logit (Eq. 2). The scaled deviance estimated by the LRM plays the role of the residual sum of squares in the normal model, providing a significance test for the importance of model factors, which is distributed in large samples like chi-square (Aitkin et al. 1989).

Results and discussion

Figure 1 shows that the frequency distribution of percentage colonization for the total data is strongly skewed to lower values (50% of the data is under 20% VAM colonization). The major difficulty with percentage data is that the values are strictly bounded; they can be no greater than 100% or less than 0%. The arcsin transformation could be used to adjust to a normal distribution of error; however, the residuals from ANOVA with the transformed data did not fit this distribution ($P=0.04$, Fig. 2). The lack of fit to a normal distribution after data transformation is common when data distribution is binomial, as has been reported for VAM colonization data (St. John and Hunt 1983; Raju et al. 1990). The assumption of homogeneity of variances was not violated (Bartlett's test $P=0.12$) among variances.

Departure from normality usually causes both the true significance level and the power to differ slightly

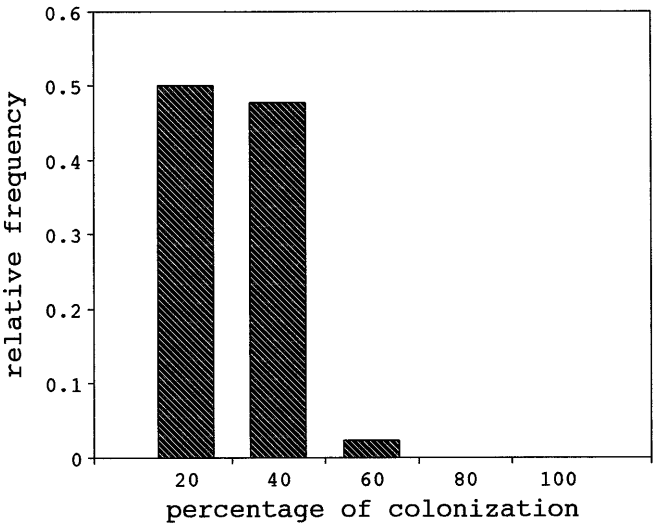

Fig. 1 Frequency distribution of percentage VAM colonization data

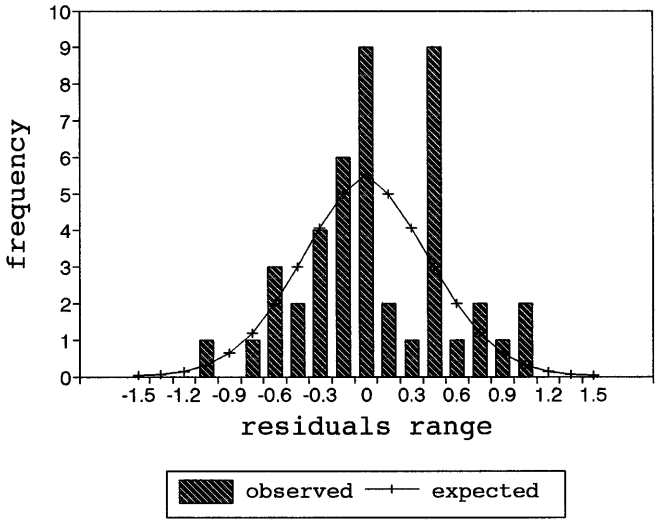

Fig. 2 Frequency distribution of residuals from ANOVA and expected values of a normal distribution

from the expected values (Montgomery 1984). Although ANOVA models are robust with respect to deviation from the normal distribution assumption, they may be affected by the type of effects (fixed-effects models are more robust than random-effects models; Montgomery 1984) and by the degrees of freedom (robustness decreases with lower degrees of freedom). In the present study, the ANOVA showed significant differences due to fertilization only, while LRM indicated that significant differences in colonization could be attributed to both factors and their interaction (Table 1). This suggests that ANOVA was less powerful than LRM, mainly in the interaction term, but this form of comparison is not strong enough.

Residual analysis is another tool to test the ANOVA model assumptions. Residuals are the non-explained part of the model and thus belong to the random error term (e_{ij}). For this reason, residuals must not be related to any other variable, including the response of y_{ij} (VAM colonization in our case). To check this, we plotted the residuals versus the fitted values (y_{ij}), which should not reveal any obvious pattern. The ANOVA residual plot showed two clusters of data explained by

Table 1 Significance levels (P) and determination coefficients (R^2) obtained for each model tested (DF degrees of freedom, *ANOVA* analysis of variance)

Source of variation	DF	ANOVA		Logistic regression	
		R^2	P	R^2	P
Fertilization	2	26.0	0.002	28.4	<0.001
Variety	2	7.1	0.148	8.1	<0.001
Interaction	4	5.2	0.747	2.5	<0.001
Total		38.4		39.0	

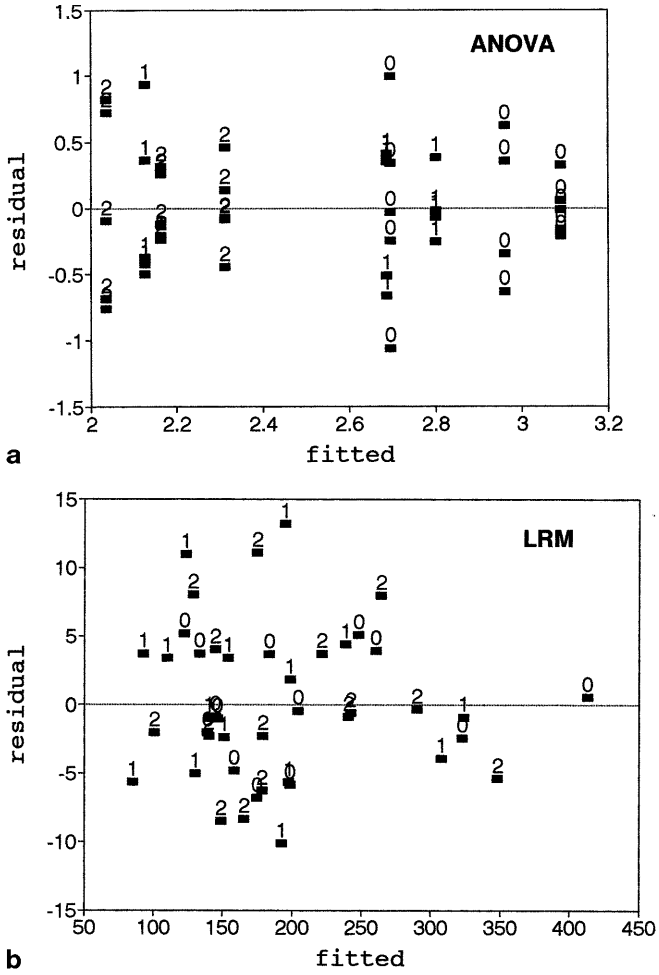

Fig. 3 Plot of residuals versus fitted values from the models tested: **a** ANOVA and **b** LRM. Numbers correspond to fertilization level: *0* non-fertilized, *1* 16.6 μg P g^{-1}, *2* 33.2 μg P g^{-1}

the fertilization factor (Fig. 3). The right cluster was dominated by the non-fertilized data (0 in the graph) and the left cluster was dominated by the 33.2 μg P g^{-1} fertilization level (2 in the graph). This pattern suggests that the e_{ij} were correlated to the fertilization factor and, as a result, the F-test was biased and the independence assumption was violated. In general, the ANOVA test is robust with respect to violations of normality but not to violations of independence (Lindman 1992). Therefore, because of the non-normal distribution and non-independence of data, the use of ANOVA was invalid.

In contrast, the LRM showed a highly significant effect of all factors included in the model (Table 1). This suggests that LRM was more powerful than ANOVA. The plot of residuals against fitted values from LRM did not show any unusual pattern (Fig. 3), which suggests that the e_{ij} values are not correlated with any other factor.

The decrease in sample size when the proportion is estimated is a constraint in using percentage data. This

Table 2 Mean comparison of percent VAM colonization among treatments between a Tukey's test performed on transformed data and the logistic regression model (LRM). Means followed by the same letter or number are not significantly different at $P=0.05$. The comparison in the LRM must be for each factor: letters compare varieties within each fertilization level and numbers compare fertilization level within each corn variety. The comparison among nine means is not valid

Variety	Fertilization (μg P g^{-1})			Mean
	0	16.6	33.2	
Tukey				
Criollo	29.16	22.66	16.54	22.78
H-34	23.50	14.72	14.42	17.54
VS-22	27.46	24.00	13.99	21.41
Mean	26.65 (a)	20.46 (b)	14.98 (c)	
LRM				
Criollo	29.16 (a, 1)	22.66 (a, 2)	16.54 (a, 3)	
H-34	23.50 (b, 1)	14.72 (b, 2)	14.42 (b, 2)	
VS-22	27.46 (a, 1)	24.00 (a, 2)	13.99 (b, 3)	

means that any percentage data assumes an equal estimation effort. For example, a 10% value can be obtained either from 100 or 400 root segments. LRM performs weighted regression using the individual sample sizes as weights and they are considered in probability estimation.

Table 2 shows mean comparisons among treatments using both models. As indicated by the Tukey's test, mycorrhizal colonization decreased proportionally with increased fertilization for all maize varieties. However, the individual varieties did not respond similarly to fertilization when analyzed by LRM, because the interaction was significant. VAM colonization in Criollo and VS-22 decreased gradually with fertilization, while it dropped with the first fertilization level (16.6 μg P g^{-1}) in H-34. These results illustrate the different conclusions that can be obtained by using different statistical tests.

Where data transformation does not violate the independence assumption, the robustness is strongly affected by the degrees of freedom. This is more so when the model is more complex (e.g., a nested model), and thus the number of replicates must be higher. However, this option is limited by such time-consuming methods as the gridline intersect method (Giovannetti and Mosse 1980). One alternative to LRM is the non-parametric Kruskal-Wallis ANOVA by ranks, but this has reduced power as a result of ranking the data (Sokal and Rohlf 1981). This constraint is not present in LRM.

We conclude that analysis of VAM colonization data by the LRM is more powerful than ANOVA because (1) binomial models consider the binary nature of data, (2) the LRM incorporates the sample size from which the percentage is estimated into the probability estimation, and (3) the LRM is fitted to non-transformed data, which are often easier to interpret.

Acknowledgements We wish to thank Arturo Estrada, Hugo Riemann, Victor Jaramillo and anonymous reviewers for helpful comments on the manuscript. This research was supported by DEPI, IPN, Mexico.

References

Aitkin M, Anderson D, Francis B, Hinde J (1989) Statistical modelling in GLIM. Clarendon, Oxford, UK

Anonymous (1985) GLIM version 3.7. The Royal Statistical Society, London

Barea JM, Escudero JL, Azcon-G de Aguilar C (1980) Effects of introduced and indigenous VA mycorrhizal fungi on nodulation, growth, and nutrition of *Medicago sativa* in phosphate-fixing soils as affected by P fertilizers. Plant Soil 54:283–296

Daft MJ, Nicolson TH (1972) Effect of Endogone mycorrhiza on plant growth. IV. Quantitative relationships between the growth of the host and the development of the endophyte in tomato and maize. New Phytol 71:287–295

Dobson AJ (1983) Introduction to statistical modelling. Chapman and Hall, London

Gavito ME, Varela L (1993) Seasonal dynamics of mycorrhizal associations in maize fields under low input agriculture. Agric Ecosyst Environ 45:275–282

Giovannetti M, Mosse B (1980) An evaluation of techniques for measuring vesicular-arbuscular mycorrhizal infection in roots. New Phytol 84:489–500

Lindman HC (1992) Analysis of variance in experimental design. Springer, New York Berlin Heidelberg

Montgomery DC (1984) Design and analysis of experiments. Wiley, New York

Phillips JM, Hayman DS (1970) Improved procedures for clearing roots and staining parasitic and vesicular-arbuscular mycorrhizal fungi for rapid assessment of infection. Trans Br Mycol Soc 55:158–161

Raju PS, Clark RB, Ellis JR, Duncan RR, Maranville JW (1990) Benefit and cost analysis and phosphorus efficiency of VA mycorrhizal fungi colonizations with sorghum (*Sorghum bicolor*) genotypes grown at varied phosphorus levels. Plant Soil 124:199–204

Sokal RR, Rohlf FJ (1981) Biometry. Freeman, New York

St. John TV, Hunt HW (1983) Statistical treatment of VAM infection data. Plant Soil 73:307–313

St. John TV, Koske RE (1988) Statistical treatment of endogonaceous spore counts. Trans Br Mycol Soc 91:117–121