ANALYSIS OF MULTINOMIAL DATA WITH OVERDISPERSION: DIAGNOSTICS AND APPLICATION

Maria Letícia SALVADOR [1]
Eduardo Eliás RIBEIRO JR [2]
César Augusto TACONELI [3]
Idemauro Antonio Rodrigues de LARA[1]

▪ ABSTRACT: In agronomic experiments, the presence of polytomous variables is common, and the generalized logit model can be used to analyze these data. One of the characteristics of the generalized logit model is the assumption that the variance is a known function of the mean, and the observed variance is expected to be close to that assumed by the model. However, it is not uncommon for extra-multinomial variation to occur, due to the systematic observation of data that are more heterogeneous than the variance specified by the model, a phenomenon known as overdispersion. In this context, the present work discusses a diagnostic of overdispersion in multinomial data, with the proposal of a descriptive measure for this problem, as well as presenting a methodological alternative through the Dirichlet-multinomial model. The descriptive measure is evaluated through simulation, based on two particular scenarios. As a motivational study, we report an experiment applied to fruit growing, whose objective was to compare the flowering of adult plants of an orange tree, grafted on "Rangpur" lime or "Swingle" citrumelo, with as response variable the classification of branches into three categories: lateral flower, no flower or aborted flower, terminal flower. Through the proposed descriptive measure, evidence of overdispersion was verified, indicating that the generalized logit model may not be the most appropriate. Thus, as a methodological alternative, the Dirichlet-multinomial model was used. Compared to the generalized logit model, the Dirichlet-multinomial proved to be more suitable to fit the data with overdispersion, by allowing the inclusion of an additional parameter to accommodate the excessive extra-multinomial dispersion.

[1] Universidade de São Paulo - USP, Escola Superior de Agricultura Luiz de Queiroz, Departamento de Ciências Exatas, CEP: 13418-900, Piracicaba, SP, Brasil. E-mail: *mariale_salvador@usp.br, idemauro@usp.br*

[2] Universidade de São Paulo - USP, Departamento de Estatística, CEP: 05508-090, São Paulo, SP, Brasil. E-mail: *jreduardo@ime.usp.br*

[3] Universidade Federal do Paraná - UFPR, Departamento de Estatística, CEP: 81531-980, Curitiba, PR, Brazil. E-mail: *taconeli@ufpr.br*

▪ KEYWORDS: Model Selection, Dirichlet-Multinomial, Maximum Likelihood, Dispersion Index.

1 Introduction

Categorized data result from the observation of characteristics of individuals that relate to quality or other attributes expressed in mutually exclusive categories. They can be classified according to the number of categories, as dichotomous or polytomous. In turn, polytomous variables can be classified according to their nature, being ordinal or nominal. According to Agresti (2019), when the response variable is polytomous, the probabilistic model most commonly associated is the multinomial distribution. Models involving this distribution are an extension of the generalized linear models (GLM) proposed by Nelder and Wedderburn (1972).

When the polytomous variable is nominal, the most common model is the generalized logit model. One of the characteristics of this model is the assumption that the variance is a known function of the mean, and the observed variance is expected to be close to that assumed by the model. But this does not always occur. In the field of agricultural sciences, for example, when working with data from entomology, animal behavior or species flowering, among others, it is not uncommon to find heterogeneity of the response variable, resulting in greater variance than specified by the proposed model. According to Hinde and Demétrio (1998), the experimental situations in which this discrepancy occurs, i.e., the observed variance is greater than the nominal one, are typical of the overdispersion phenomenon.

Further according to various researchers, the phenomenon of overdispersion can occur for several reasons, such as the variability of the experimental material, omission of covariates that can explain the lack of homogeneity, and excess of zeros in the data, among others. Furthermore, different causes can coexist, making it difficult to infer the precise cause of this phenomenon. Thus, diagnostic measures are important tools to study overdispersion, and it is essential to recognize that this phenomenon is present in the data so that alternatives can be applied to support model choice in order to ensure the validity of the estimation and resulting inferences. On the other hand, according to Olsson (2002), care should be taken not to confuse the phenomenon of overdispersion with unsatisfactory fit of the model, which can be caused, for example by the wrong choice of the link function or incorrect linear predictor.

The literature contains models able to deal with the overdispersion problem in the fitting process, in particular for proportion and count data, such as quasi-likelihood models (Wedderburn, 1974), which admit a more general form for the variance function, or alternatively by allowing an additional parameter and a two-stage model (Hinde and Demétrio, 1998) that assumes a distribution for the response variable and additionally another for the parameters. In this work, the two-stage model approach is used, through the Dirichlet-multinomial composite distribution (Mosimann, 1962), which has been used for the analysis of polytomous datasets that contain extra-multinomial variation and is a multicategory extension of the beta-binomial model.

Therefore, this work characterizes the problem of overdispersion in the context of categorized nominal polytomous data, proposes a descriptive dispersion index as a diagnostic measure for this overdispersion phenomenon. The performance of this index is assessed by simulation and we use the Dirichlet-multinomial model as an alternative in cases of overdispersion of multinomial data. An application in the field of agricultural sciences it also presented to illustrate the methodological procedure.

2 Material and methods

2.1 Material

As a motivational study, we consider part of an experiment developed by Voigt (2013), conducted in a greenhouse in 2011. The experiment was carried out in a completely randomized design involving the orange tree “x11”, which has as main characteristic the fact of presenting a short juvenile period.

The objective of this experiment was to evaluate the flowering of adult trees of this variety formed by grafting on “Rangpur” lime or “Swingle” citrumelo rootstock, considered as the treatments. For application purposes, this work considers the data referring to the winter season given the fixed total of items the multinomial distribution is an appropriate starting point. The response variable is the count of branches in three mutually exclusive categories: lateral flower (category 1), no flower or aborted flower (category 2), terminal flower (category 3).

2.2 Methods

2.2.1 Generalized logit model

The generalized logit model has the property of comparing each response category against reference, often the last.

In order to establish notation, the response variable, Y_{ijk}, denotes the number of branches of the i-th plant, in the j-th branch classification and k-th treatment, $i = 1, \cdots, 16$, $j = 1, 2, 3$ (lateral flower, no flower or aborted flower, terminal flower) and $k = 1, 2$ (“Rangpur” lime or “Swingle” citrumelo rootstocks). It is assumed that Y_{ijk} has a multinomial distribution, where its components represent the occurrences of response categories.

For an experimental situation, consider a random sample of this distribution and let $\boldsymbol{x}$ (rootstocks effect) be the vector of explanatory variables, where $\boldsymbol{\beta} = (\beta_{0j}, \beta_{jk})^T$ is the vector of unknown parameters.

In this context, fixing the J-th category as a reference, in this work represented by the classification of terminal flowers $J = 3$, we have that the model is that defined by:

$$\eta_{jk} = \ln\left(\frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}\right) = \beta_{0j} + \beta_{jk}\boldsymbol{x}, \tag{1}$$

where β_{jk} is the parameter associated with the k-th rootstock effect in the j-th logit, with $j = 1, 2, 3$ and $k = 1, 2$ and $\pi_j(\boldsymbol{x}) = P(\mathrm{Y} = j|\boldsymbol{x})$ represents the probability for each category, with $\sum_{j=1}^{3} \pi_j(\boldsymbol{x}) = 1$.

In the generalized logit model, the intercept (β_{0j}) and the regression parameter vector (β_{jk}) are different for each logit, which implies that the effects of the explanatory variables vary according to the category (Agresti, 2002). The model (1) considers $J-1$ logits for all possible pairs of categories, in our case, there are 2 logits, that are refered to the category 3 (terminal flower). Regarding the estimation of the model parameters (1), we use the maximum likelihood method via an iterative process. Here, we have used package "nnet" (Venables e Ripley, 2002).

Assuming model (1), the probabilities predicted by the generalized logit model are given by:

$$\hat{\pi}_j(\boldsymbol{x}) = \frac{\exp\left(\hat{\beta}_{0j} + \hat{\boldsymbol{\beta}}_j\boldsymbol{x}\right)}{1 + \sum_{j=1}^{2} \exp\left(\hat{\beta}_{0j} + \hat{\boldsymbol{\beta}}_j\boldsymbol{x}\right)},$$

where, $j = 1, 2$ (lateral flower, no flower or aborted flower).

However, to use the generalized logit model, as already reported, it is necessary to check whether the observed variance is close to that expected for the model. Otherwise, if the observed variance exceeds that, it may be an indication of overdispersion, hence we consider diagnostics through measures such as residual deviance, the estimation of an additional overdispersion parameter, among others. Additionally, one can use the half-normal plot with simulated envelopes (Moral, Hinde and Demétrio, 2017) to check the model's goodness of fit.

Once the phenomenon of overdispersion is verified, it is necessary to use models that take into account the extra-multinomial variation, and the model with the Dirichlet-multinomial distribution (Mosimann, 1962) for the polytomous response variable is a good option (Freitas, 2001) .

2.2.2 Dirichlet-multinomial model

In the Dirichlet-multinomial model (Morel & Nagaraj, 1992), the Dirichlet-multinomial distribution described by Mosimann (1962) is assumed for the response variable vector. In this context, let $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_J)$ a vector for polytomous response with the probability vetor $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_J)$, the function is

$$
\begin{aligned}
f(y|\alpha) &= \int_{\Omega} f(y|\pi) f(\pi|\alpha) d\pi \\
&= \int_{\Omega} \frac{n!}{n_1! n_2! \cdots n_J!} \prod_{j=1}^{J} (\pi_j)^{n_j} \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \prod_{j=1}^{J} (\pi_j)^{\alpha_j - 1} d\pi \\
&= \frac{n!}{n_1! n_2! \cdots n_J!} \frac{\Gamma(\sum_{j=1}^{J} \alpha_{ij})}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \int_{\Omega} \prod_{j=1}^{J} (\pi_{ij})^{n_j + \alpha_j - 1} d\pi \\
&= \frac{n!}{n_1! n_2! \cdots n_J!} \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\Gamma(n + \sum_{j=1}^{J} \alpha_j)} \prod_{j=1}^{J} \frac{\Gamma(n + \alpha_j)}{\Gamma(\alpha_j)},
\end{aligned} \qquad (2)
$$

where, $\Omega = \{\pi; \pi_j \in (0,1), j = 1, \cdots, J; \sum_{j=1}^{J} \pi_j = 1\}$, the parameters $\boldsymbol{\alpha} = (\alpha_1, \cdots \alpha_J)$ are strictly positive and $\Gamma(.)$ is the gamma function. The Dirichlet-multinomial distribution is composed by taking the random vectors $\boldsymbol{Y}$ and $\boldsymbol{\pi}$, where a first stage $\boldsymbol{Y} \mid \boldsymbol{\pi}$ follows the multinomial distribution and in a second stage the parameter vector $\boldsymbol{\pi}$ follows the Dirichlet distribution.

As in Section 2.2.1, let $\boldsymbol{x}$ (rootstocks effect) also be the vector of explanatory variables, where $\boldsymbol{\beta} = (\beta_{0j}, \beta_{jk})^T$ is the unknown parameter vector of interest, where $j = 1, 2, 3$ and $k = 1, 2$. To incorporate the effect of the explanatory variables, the log-linear link function is used:

$$\boldsymbol{\eta} = \ln(\alpha_j) = \ln(\beta_{0j} + \boldsymbol{\beta}_j^T \boldsymbol{x}),$$

where, $\alpha_j = \exp(\beta_{0j} + \boldsymbol{\beta}_j^T \boldsymbol{x})$. Thus, the Dirichlet-multinomial model is given by:

$$\pi_j(\boldsymbol{x}) = \frac{\alpha_j(\boldsymbol{x})}{\sum_{j=1}^{J} \alpha_j(\boldsymbol{x})}, \; j = 1, 2, 3. \qquad (3)$$

where $\pi_j(\boldsymbol{x}) = P(\mathrm{Y} = j|\boldsymbol{x})$ represents the probability for each category, with $\sum_{j=1}^{3} \pi_j(\boldsymbol{x}) = 1$

Note that in comparison with the generalized logit model, the Dirichlet-multinomial model does not compare each category with a reference category.

Regarding the estimation of the parameters, model (3) can be done by the maximum likelihood method (Chen & Li, 2013).

According to Paul *et al.* (1989), one can reparameterize the probability function of the Dirichlet-multinomial distribution considering $\gamma = \frac{1}{\sum_{j=1}^{J} \alpha_j}$. This additional parameter plays an important role in this distribution, not only to characterize it but also to control the variability. In this context, if $\gamma = 0$, the Dirichlet-multinomial model reduces to a multinomial model. A positive value of the parameter γ is characteristic of the Dirichlet- multinomial model, being usual when there is overdispersion of data. Thus, a likelihood ratio test to discriminate between the two model structures has the following hypotheses:

$$\begin{cases} H_0 : \gamma = 0 \ \text{(multinomial model).} \\ H_a : \gamma > 0 \ \text{(Dirichlet-multinomial model).} \end{cases}$$

whose test statistic likelihood ratio test (LRT) statistic is given by:

$$\text{LRT} = -2\ln(\ell_1 - \ell_0),$$

where ℓ_0 is the logarithm of the likelihood function under the null hypothesis, and ℓ_1 represents the logarithm of the likelihood function under the alternative hypothesis. Under the true null hypothesis, LRT has a chi-square distribution χ^2_g, where g is equal to the difference between the number of parameters (np) of the multinomial model and the Dirichlet-multinomial model. Freitas (2001) pointed out that if the hypothesis test proposed by Paul *et al.* (1989) is significant, there are indications of the presence of overdispersion. This extra- multinomial variation will be accommodated in the Dirichlet-multinomial model because of the additional parameter it contains. Here, we have used the package "MGLM" (Zhang and Zhou, 2016).

Additionally, in this work, the comparison between the two structures, in terms of goodness of fit, can be done by the half-normal plot (Moral, Hinde & Demétrio, 2018).

2.2.3 Dispersion index

In this section, a dispersion index proposal is presented, as a descriptive measure of overdispersion diagnostic, for nominal polytomous data, constructed as follows:

1. Consider a response variable Y_j, where j represents the j-th category. A dispersion index is calculated for each category:

   $$\text{ID}_j = \frac{\text{Var}_j(\text{Observed})}{\text{Var}_j(\text{Expected})}$$

   where Var_j(Observed) is given by the j-th observed variance of the data and Var_j(Expected) is the j-th variance expected of the assumed model, in which in the present work $j = 1, 2, 3$.

2. Finally, the mean of the dispersion indices obtained in the previous item for each of the categories is calculated, so the multinomial dispersion index (MDI) is given by:

   $$\text{MDI} = \frac{\sum_{j=1}^{J} \text{ID}_j}{\text{J}} \qquad (4)$$

   where J is the total number of categories.

Thus, it is expected that if the descriptive measure of the overdispersion diagnostic is around one, there is evidence that the data are equidispersed, that is, the observed variance is close to that fitted by the proposed model. On the other hand, if MDI is greater than one, there is evidence of the presence of overdispersed data.

In order to investigate the behavior of the dispersion index proposed in this work, a simulation study was carried out next (Section 2.2.4).

2.2.4 Simulation study

Under the equidispersion and overdispersion hypothesis, data were simulated from two particular scenarios.

For the simulation, we considered three response categories ($j = 1, 2, 3$) with two explanatory variables ($k = 1, 2$), for $N = 60$ groups containing 20 individuals each ($n = 20$). The simulations were carried out under the hypothesis of two scenarios: equidispersed and overdispersed data. In the first scenario, we applied the multinomial distribution and in the second the Dirichlet-multinomial distribution, since this distribution has an additional parameter to accommodate the extra-multinomial scattering (γ).

For the multinomial distribution, the probability vectors used in the simulation process associated with each explanatory variable were $\pi_{1j} = (0.66; 0.28; 0.06)$ and $\pi_{2j} = (0.59; 0.37; 0.03)$. For the Dirichlet-multinomial distribution, additionally the parameters $\alpha_{1j} = (0.1; 0.4; 0.1)$ and $\alpha_{2j} = (0.2; 0.5; 0.3)$ were used, whose parameters values were chosen randomly.

The simulation was performed using the R software (R Core Team, 2020). Based on these fixed parameters, 1,000 datasets were simulated for each scenario considered. In this context, for the first scenario, the probability vectors of the multinomial distribution were fixed, in which the dataset was simulated using the rmultinom ( ) function. In the second scenario, the dataset was simulated using the rdirm ( ) function, fixing the parameters of the Dirichlet-multinomial distribution.

For each simulated dataset, the multinomial and Dirichlet-multinomial models were fitted. It was possible to calculate for each simulated dataset the variance observed, the variance expected by the multinomial model, and the finally multinomial dispersion index, as expressed by equation (4). The indexes were then analyzed descriptively, allowing empirically obtained acceptable ranges for the assumed existence of overdispersion.

3 Results and discussion

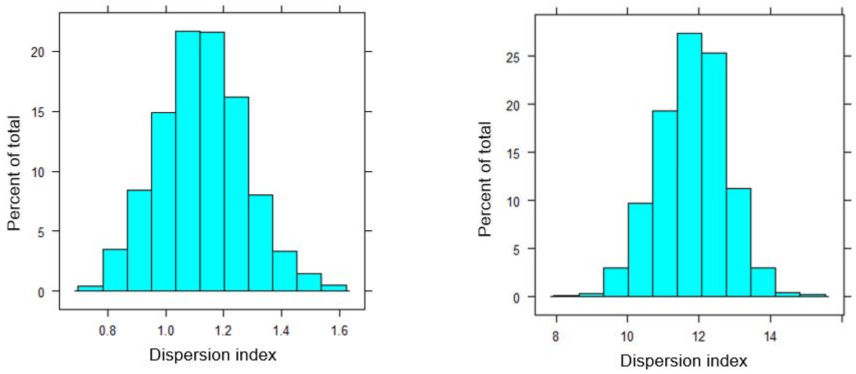
3.1 Simulation study

Based on the simulation study described in Section 2.2.4, some descriptive statistics are presented below.

Under the hypothesis of equidispersed data, the multinomial dispersion index (MDI) varied approximately from 0.7 to 1.5 (Table 1). On the other hand,

Table 1 - Descriptive measures referring to the dispersion index based on simulated data, according to each scenario considered in the estimation process

Descriptive measures	First scenario (equidispersion)	Second scenario (overdispersion)
Maximum	1.59	15.25
Minimum	0.73	8.22
Amplitude	0.86	7.03
Mean	1.12	11.82
Standard deviation	0.15	0.95



(a) First scenario (equidispersion).

(b) Second scenario (overdispersion).

Figure 1 - Histograms of the dispersion index according to equidispersion (a) and overdispersion (b) scenarios considered in the simulation study.

under the variance perturbation, the multinomial dispersion index (MDI) varied approximately from 8 to 15, as shown in the Table 1. Based on the histograms presented in the Figure 1 (a) and (b), the distribution of the multinomial dispersion index apparently has normal distribution.

Comparison of the simulations carried out by the two processes showed that the mean dispersion index obtained in the second scenario was approximately 10.5 times greater than that obtained in the first. In turn, the standard deviation was approximately six times higher in the second scenario than the first one, indicating extra variation between the dispersion index when considering the multinomial model. Therefore, empirically for our scenarios, it can be assumed that the acceptable values of the multinomial dispersion index range from 0.7 to 1.5, indicating there is no extra-multinomial variation present in the data.

3.2 Motivational study

Descriptive analysis was performed to explore the data, through mean and variance, that are presented in the Table 2.

Table 2 - Descriptive summary of the number of branches in relation to classification, according to the experiment performed by Voigt (2013) during the winter season

	Rootstocks			
	Rangpur lime		Swingle citrumelo	
Branch classification	Mean	Variance	Mean	Variance
Lateral flower	89.00	939.25	70.43	496.25
No flower or aborted	3.67	10.00	4.43	7.62
Terminal flower	12.33	71.50	27.71	430.57

There were on average a greater number of branches classified as lateral flower, followed by terminal flower, and no flower or aborted flower. Furthermore, there was a strong mean-variance relationship, that is, the variance increased with the mean.

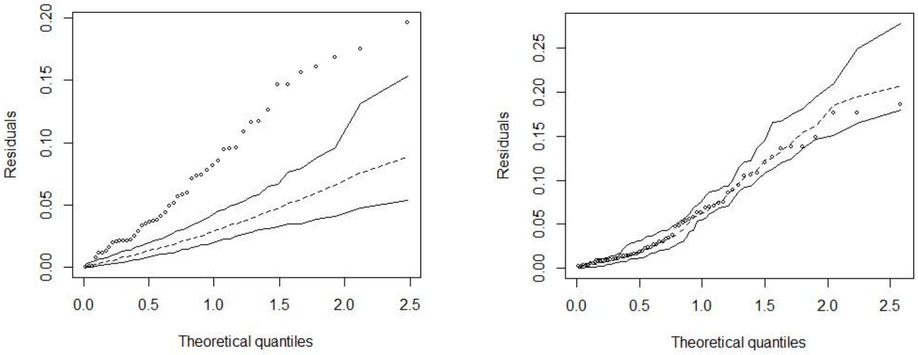
Then, the generalized logit model with canonical link function was fitted to data and with this to observe the possible existence of overdispersion, by comparing the variances observed and those assumed by the model, presented in the Table 3

There was a lack of homogeneity in the treatments, since both the observed variances were greater than those obtained with fitted model. Also, the value of the residual deviance exceeded the number of degrees of freedom (deviance = 159.11; df= 28). Furthermore, the proposed multinomial dispersion index was MDI = 7.7. Based on these three diagnostics measures, there is evidence of overdispersed data. In this context, an alternative model would be the Dirichlet-multinomial.

The diagnostic by the half-normal plots for both models are presented in the Figures 2 (a) and (b). In the generalized logit model, most of the points

Table 3 - Observed variances and expected from fitted generalized logit model, according to the data from the experiment performed by Voigt (2013) in the winter season

	Rootstocks			
	Rangpur lime		Swingle citrumelo	
Branch classification	Observed	Expected	Observed	Expected
Lateral flower	939.25	103.46	496.95	106.08
No flower or aborted	10.00	1.11	7.62	1.28
Terminal flower	71.50	11.51	430.57	38.25



(a) Generalized logit model.

(b) Dirichlet-multinomial model.

Figure 2 - Diagnostic graph (half-normal plot) to assess the fit goodness of the generalized logit model and Dirichlet-multinomial model, fitted to the data referring to the experiment performed by Voigt (2013) in the winter.

Table 4 - Comparison between the generalized logit models, with multinomial and Dirichlet-multinomial distributions for response and through the values of the AIC, log-likelihoods (ℓ), the parameter number (np) and the likelihood ratio test (LRT)

Models	np	ℓ	AIC	LRT	p-value
Multinomial	4	-136.62	281.24		
Dirichlet-multinomial	6	-68.34	208.68	76.56	<0.001

are outside the simulation envelope, indicating that this model has unsatisfactory fit. Thus, as previously mentioned, to accommodate the overdispersion, the Dirichlet-multinomial model was considered, which presented a satisfactory fit, as the points were accommodated within the simulation envelope. Furthermore, both the likelihood ratio test (LRT) and the Akaike information criterion (AIC), presented in the Table 4, also indicate the selection of the Dirichlet-multinomial model.

Table 5 - Likelihood ratio test for the treatment effect using generalized logit and Dirichlet- multinomial models, where np= number of parameters, ℓ = log-likelihood and LRT = likelihood ratio test statistic

	Models							
	Generalized Logit				Dirichlet-multionomial			
Linear Predictor	np	ℓ	LRT	p-value	np	ℓ	LRT	p-value
$\eta_{jk} = \beta_{0j}$	2	-169.55			3	-102.15		
$\eta_{jk} = \beta_{0j} + \beta_{jk}$	4	-136.62	65.87	<0.01	6	-98.33	7.63	0.0543

In order to verify the possible existence of a treatment effect in the experiment, the likelihood ratio test was used (Table 5), showing at 5% significance that the Dirichlet-multinomial model did not present a rootstock effect, but can consider them marginally. On the other hand, when applying the generalized logit model with multinomial distribution, there were significant rootstock effects. Therefore, this corroborates the importance of overdispersion diagnostics in the data to choose the appropriate model to ensure more reliable estimation and avoid erroneous conclusions.

Table 6 - Estimated parameters and standard errors in relation to the Dirichlet-multinomial model with rootstock effect in the winter season

Parameter	Estimate	Standard error	p-value
β_{01}(lateral flower)	2.977	0.457	<0.01
β_{02}(no flower or aborted)	0.004	0.442	0.993
β_{03}(terminal flower)	1.038	0.445	0.020
β_{11}(lateral flower “Swingle”)	-0.206	0.677	0.761
β_{21}(no flower or aborted “Swingle”)	0.274	0.646	0.671
β_{31}(terminal flower “Swingle”)	0.666	0.677	0.326

The estimates and standard errors obtained by fitting the Dirichlet-multinomial model are presented in the Table 6, indicating the parameters β_{jk} are not significant, due to the marginal treatment effect.

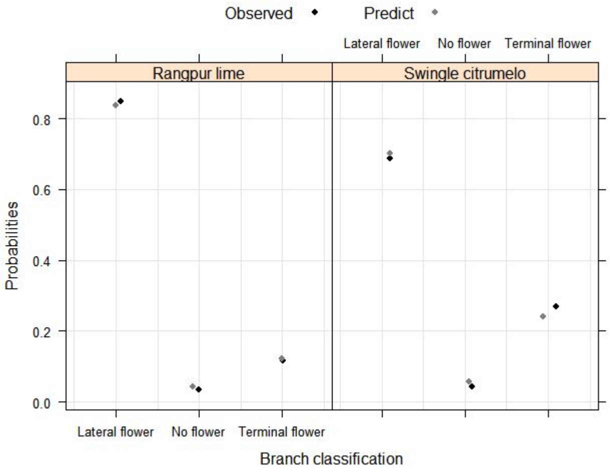


Figure 3 - Observed and predict probabilities by the Dirichlet-multinomial model regarding rootstocks and branch classification.

Finally, the predicted and observed probabilities according to the Dirichlet-multinomial model are illustrated in the Figure 3. The most likely category of branches during the winter season was lateral flowers, estimated at 0.837 for plants with “Rangpur” lime rootstock and 0.700 for “Swingle” citrumelo rootstock. The second most likely category was terminal flowers, and its probability was higher in plants with “Swingle” citrumelo rootstock, but this difference was not significant.

Conclusions

This work presents an introductory study to analyze polytomous data, with overdispersion in a cross-sectional study. The proposed alternative is the use of the Dirichlet-multinomial model, a two-stage model that contains an additional parameter in comparison with the multinomial one, allowing accommodation of extra variability. We also proposed the dispersion index as a diagnostic measure of overdispersion in nominal polytomous data, and evaluated its performance through the initial simulation study.

Both overdispersion diagnostics and the choice of the appropriate model are important to avoid wrong conclusions. As observed in the analysis of data in the motivational study, if the presence of overdispersion is not considered, it can be concluded, by using the generalized logit model, that the treatment effect is significant. However, with the Dirichlet-multinomial model, this fact does not occur.

Although the Dirichlet-multinomial model presented a satisfactory fit and the dispersion index initially presented satisfactory performance we reinforce, this work presents only an initial study for modeling of multinomial categorical data with overdispersion, which is a topic for future research. As perspectives for future work, we intend to carry out a more detailed study of the proposed dispersion index, by evaluating its performance with confidence intervals and applying other distributions such as the negative multinomial for example.

Acknowledgments

Work derived from the master's dissertation "The problem of overdispersion in categorized nominal pilot data in agrarian studies" with support from the Brazilian Foundation, "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" (CAPES) process number 23038.013648/2018-51.

SALVADOR, M. L.; RIBEIRO JR, E. E.; TACONELI, C.A.; LARA, I.A.R., Uma alternativa para diagnóstico e análise de dados multinomiais com superdispersão: aplicação à Fruticultura. *Braz. J. Biom.,* Lavras, v.40, n.3, p.342-355, 2022.

- *RESUMO: Em experimentos agronômicos é comum a presença de variáveis politômicas e o modelo dos logitos generalizados é uma classe de modelos de regressão que podem ser empregadas para a análise desses dados. Uma das características do modelo dos logitos generalizados é a pressuposição de que a variância é uma função conhecida da média e, espera-se que a variância observada esteja próxima da pressuposta pelo modelo. No entanto, não é raro ocorrer uma variação extra-multinomial, devido à observação sistemática de dados mais heterogêneos do que a variância especificada pelo modelo, fenômeno este conhecido como superdispersão. Neste contexto, o presente trabalho tem como objetivo discutir a importância de diagnosticar o problema da superdispersão em dados multinomiais, com proposta de uma medida descritiva para diagnóstico, assim como apresentar uma alternativa metodológica por meio do modelo Dirichlet-multinomial. A medida descritiva foi avaliada por meio de simulação, restringindo-se a dois cenários particulares. Como um estudo de motivação tem-se um experimento aplicado à fluticultura, cujo o objetivo foi avaliar o florescimento de plantas adultas de uma laranjeira, enxertadas sobre o limão "Cravo" e o citrumelo "Swingle", tendo com variável resposta à classificação de ramos das plantas em três categorias: terminal, lateral, sem flor ou abortada. Por meio da medida descritiva proposta verificou-se evidências de superdispersão, indicando que o modelo dos logitos generalizados pode não ser o mais apropriado. Dessa forma, como alternativa metodológica utilizou-se o modelo Dirichlet-multinomial. Comparativamente ao modelo dos logitos generalizados o Dirichlet-multinomial mostrou-se mais adequado para o ajuste de dados com superdispersão por permitir a inclusão de um parâmetro adicional para acomodar a dispersão excessiva extra-multinomail.*

- *PALAVRAS-CHAVE: Seleção de modelos, Dirichlet-multinomial, Máxima-verossimilhança, Índice de dispersão.*

References

AGRESTI, *An introduction to categorical data analysis*, John Wiley & Sons, 2019.

CHEN, J.; LI, H. *Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. Annals of Applied Statistics*, v.7, n.1, p.418–442, 2013. ISSN 19326157.

FREITAS, S. M. *Modelos para proporções com superdispersão provenienetes de ensaios toxicológicos no tempo.* 124p. Tese (Doutorado) — Universidade de São Paulo / Escola Superior Agricola "Luiz de Queiroz", 2001.

MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. B. *Half-Normal Plots and Overdispersed Models in R : The hnp Package. Journal of Statistical Software*, v.81, n.10, 23p., 2017.

MORAL, R. d. A.; HINDE, J.; DEMÉTRIO, C. G. B. *Half-Normal Plots with Simulation Envelopes.* [S.l.]: 2018-05-21, 2018. 28p.

MOREL, J. G.; NAGARAJ, N. K. A *Finite Mixture Distribution for Modelling Multinomial Extra Variation.* Biometrika Trust, Oxford University Press, v.80, n.2, p.363–371, 1992.

MOSIMANN, J. E. *On the Compound Multinomial Distribution , the Multivariate β - Distribution , and Correlations Among Proportions.* Biometrika Trust, Oxford University Press, v.49, n. 1, p.65–82, 1962.

NELDER, J. A.; WEDDERBURN, R. W. M. G*eneralized Linear Models.* v.135, n.3, p.370–384, 1972.

OLSSON, U. *Generalized Linear Models An Applied Approach.* [S.l.]: Lund: Studentlitteratur, 2002. 232p. ISBN 9789144031415.

PAUL, S. R.; LIANG, K. Y.; SELF, S. G. *On Testing Departure from the Binomial and Multinomial Assumptions.* v.45, n.1, p.231–236, 1989.

VENABLES, W. N.; RIPLEY, B. D. *Modern Applied Statistics with S.* Fourth. New York: Springer,2002. ISBN 0-387-95457-0.

VOIGT, V. *Caracterização fenotípica e avaliação da expressão de genes envolvidos na indução e no florescimento da laranjeirax11.* Tese (Doutorado) — Universidade de São Paulo, 2013.

ZHANG, Y.; ZHOU, H. *Mglm: Multivariate response generalized linear models.* R package version 0.2.0, v.7, 2016.

Received on 03.09.2021.

Approved after revised on 22.11.2021.