

Superdispersão em Dados Categorizados Multinomiais: uma Aplicação em Ciências Agrárias

Aluna: Maria Letícia Salvador

Orientador: Prof. Dr. Idemauro Antonio Rodrigues de Lara



Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"
Pós-Graduação em Est. e Exp. Agrônômica

21 de novembro de 2018

Sumário

① Motivação

② Revisão de Literatura

③ Materiais

Florescimento da Laranjeira x11

④ Resultados Parciais

Resultados Parciais: Florescimento da Laranjeira x11

⑤ Conclusão

⑥ Referências

Motivação

- Dados categorizados são frequentes na prática em diversas áreas, em especial nas Ciências Agrárias.
- Na análise que envolve dados categorizados, espera-se que a variância observada esteja próxima da variância pressuposta pelo modelo assumido.
- Existem casos em que os dados são mais heterogêneos do que a variância especificada pelo modelo proposto.

Objetivo:

- Caracterizar o problema da superdispersão;
- Apresentar modelos para solucionar o problema;

Dados Categorizados

Dados categorizados decorrem de observações de características dos indivíduos que dizem respeito a uma qualidade ou atributo, expresso em categorias mutuamente exclusivas.

Segundo Maccullagh e Nelder(1989), estas variáveis podem ser classificadas de acordo com as quantidades de categorias.

Variável Categorizada { Dicotômica
 { Politômica { Nominal
 { Ordinal

Modelo Logito de Categoria de Referência

Considere uma variável resposta Y politômica nominal com J categorias, sendo $\pi_j(\mathbf{x}) = P(Y = j|\mathbf{x})$, com $\sum_{j=1}^J \pi_j(\mathbf{x}) = 1$.

O modelo é definido por [Agresti(1996)]:

$$\text{logito}(\pi_j) = \ln \left(\frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})} \right) = \alpha_j + \beta_j \mathbf{x}$$

em que $j = 1, \dots, J - 1$.

Superdispersão

- De acordo com Mcculagh e Nelder(1989), a superdispersão ocorre quando a variância da variável resposta excede a variação nominal;
- Segundo Olsson (2002) a superdispersão se da pelo fato do ajuste do modelo ser insatisfatório;
- Deve-se tomar cuidado o fenômeno da superdispersão com o ajuste insatisfatório do modelo;

Superdispersão

O fenômeno da superdispersão pode ser identificado por meio:

- Do valor da *deviance* residual e do número de graus de liberdade do resíduo;
- Verificando se a variância observada excede a variação obtida por meio do ajuste do modelo;

Dirichlet-multinomial

- A distribuição Dirichlet-multinomial foi introduzida por Mosimann (1962).
- Esta distribuição tem sido utilizada para modelar dados categorizados que apresentam superdispersão.
- Uma alternativa para modelar a superdispersão é assumir o modelo de dois estágios. Ou seja, a variável resposta segue uma distribuição composta.

Dirichlet-multinomial

- Considere que $\mathbf{Y}|\boldsymbol{\pi} \sim \text{Multinomial}(n, \boldsymbol{\pi})$;
- Considere também que $\boldsymbol{\pi}$ segue a distribuição Dirichlet sob o espaço amostral Ω , em que $\Omega = \{\boldsymbol{\pi}; \pi_j \in (0, 1), j = 1, \dots, J; \sum_{j=1}^J \pi_j = 1\}$.
- A distribuição Dirichlet-multinomial, é definida por:

$$f(n|\boldsymbol{\alpha}) = \frac{n!}{y_1!y_2!\cdots y_J!} \prod_{j=1}^J (\pi_j)^{y_j} \frac{\Gamma(\sum_{j=1}^J \alpha_j)}{\Gamma(n + \sum_{j=1}^J \alpha_j)} \prod_{j=1}^J \frac{\Gamma(y_j + \alpha_j)}{\Gamma(y_j + \alpha_j)}$$

em que, os parâmetros $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_J)$ são estritamente positivos, $\mathbf{Y} = (Y_1, \dots, Y_J)$ e $\boldsymbol{\pi} = (\pi_1, \dots, \pi_J)$.

Teste de Hipótese

Segundo Paul et al. (1989), uma maneira de reparametrizar o modelo é considerando que $\theta_i = \frac{1}{\sum_{j=1}^J \alpha_j}$.

$$\begin{cases} H_0 : \theta_i = 0 \text{ (multinomial)} \\ H_a : \theta_i > 0 \text{ (Dirichlet-Multinomial)} \end{cases}$$

Dirichlet-multinomial

Multinomial

- $E(X_j) = n\pi_j$
- $Var(X_j) = n\pi_j(1 - \pi_j)$.

Dirichlet-multinomial

- $E(\mathbf{Y}) = n \cdot \frac{\alpha_j}{\sum_{j=1}^J \alpha_j} = n\mu_j$
- $Var(\mathbf{Y}) = n\mu_j(1 - \mu_j)[1 + (n - 1)\rho_j]$

Florescimento da Laranjeira x11



- O experimento foi desenvolvido por Voigt (2013), realizado durante o ano de 2011;
- Os dados são referentes a Estação Inverno;
- O objeto de estudo é a laranjeira da variedade “x11”;
- 9 plantas foram enxertadas sobre o limão “Cravo”;
- 7 plantas foram enxertadas sobre o citrumelo “*Swingle*”;
- A variável resposta deste experimento é o tipo de ramo;

Florescimento da Laranjeira x11

Classificação dos ramos:

- Número de ramos com flor terminal;
- Número de ramos com flor lateral;
- Número de ramos sem flor
- Número de ramos com flor abortada;

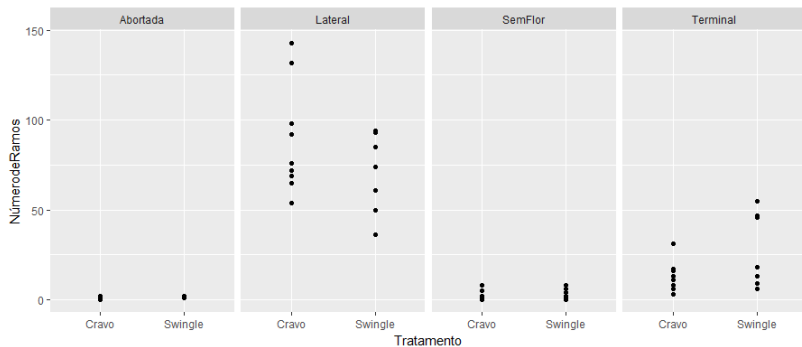
Objetivo do Experimento

Avaliar o florescimento de plantas adultas da variedade “x11”.



Resultados Parciais: Florescimento da Laranjeira x11

Figura: Gráfico de pontos das variedades de porta-enxertos limão “Cravo” e citrumelo “Swingle” em relação a classificação dos ramos no Inverno.



Resultados Parciais: Florescimento da Laranjeira x11

Considere uma variável resposta Y_{ijk} que segue a distribuição Multinomial.

- **Modelo 1:** $\eta_j = \ln \left(\frac{\pi_j}{\pi_J} \right) = \beta_{0j}$ em que, $j = 1, 2, 3$.
- **Modelo 2:** $\eta_{jk} = \ln \left(\frac{\pi_{jk}}{\pi_{JK}} \right) = \beta_{0j} + \beta_{j\text{porta-enxerto}_k}$ em que, $j = 1, 2, 3$.

Tabela: Seleção de modelos levando em consideração o valor do AIC.

Modelos	η	AIC	Deviances	G.L.
1	$\eta_j = \beta_{0j}$	375,34	2170,548	45
2	$\eta_j = \beta_{0j} + \beta_{j\text{porta-enxerto}_k}$	315,28	2104,486	42

Resultados Parciais: Florescimento da Laranjeira x11

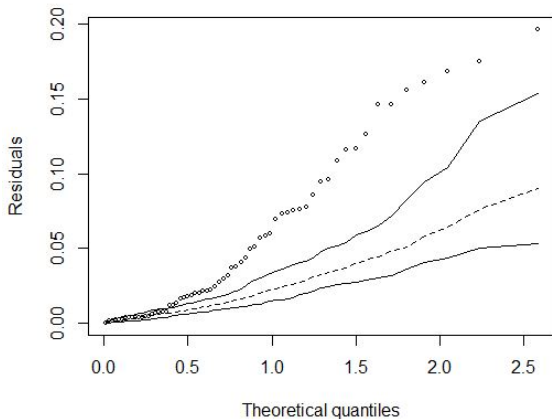
Tabela: Resumo descritivo dos dados em relação a cada classificação de ramos e dos porta-enxertos limão “Cravo” e citrumelo “Swingle” na estação inverno.

Limão “Cravo”		
Classificação	Var. Obs.	Var. Ajust.
Terminal	71,5	0,933
Lateral	939,25	1,162
Sem Flor	6,61	0,2408
Com Flor Abortada	0,44	0,066

Citrumelo “Swingle”		
Classificação	Var. Obs.	Var. Ajust.
Terminal	430,57	1,380
Lateral	496,95	1,506
Sem Flor	8,24	0,217
Com Flor Abortada	0,14	0,077

Resultados Parciais: Florescimento da Laranjeira x11

Figura: *Half-normal plot* Modelo 2 do experimento florescimento da laranjeira x11.



Resultados Parciais: Florescimento da Laranjeira x11

Teste de Hipótese





$$\begin{cases} H_0 : \theta_i = 0(\text{multinomial}) \\ H_a : \theta_i > 0(\text{Dirichlet-Multinomial}) \end{cases}$$

Modelos	Nº de Par.	AIC	BIC	P-valor
Multinomial	6	315,28	319,9	-
Dirichlet-Multinomial	8	257,5	263,68	< 0,01




Conclusão

- Tanto a distribuição multinomial quando a Dirichlet-multinomial modelam dados politômicos, porém eles apresentam estruturas de média e variância muito diferentes;
- O modelo Multinomial tem uma estrutura de média e variância mais restrita;
- O Dirichlet-multinomial tem uma estrutura de média e variância mais flexível.

Referências bibliográficas

-  AGRESTI, A., 1996 *An introduction to categorical data analysis*, volume 135. Wiley New York.
-  AGRESTI, A., 2002 *Categorical data analysis*, volume 482. John Wiley & Sons.
-  McULLAGH, P. and NELDER, 1989a Binary data. In *Generalized linear models*, pp., Springer.
-  MOSIMANN, J.E. On the compound multinomial distribution, the multivariate β -distribution, and correlations among proportions. **Biometrika**, 49, p.65, 1962.

Referências bibliográficas

-  NELDER, J.A.; WEDDERBURN, R. W. M. *Generalized linear models. Journal of the Royal Statistical Society A*, Hoboken, v. 135, n. 3, p. 370-384, 1972.
-  OLSSON, U., 2002. *Generalized linear models. An applied approach. Studentlitteratur, Lund* **18**.
-  VOIGT, V., 2013 *Caracterização fenotípica e avaliação da expressão de genes envolvidos na introdução e no florescimento da laranjeira à 11*. Ph.D. thesis, Universidade de São Paulo.