Modelo linear generalizado para dados de contagem: uma aplicação em um experimento agronômico

Macedo, Caio M.[1] (IC); Medeiros, Simone Daniela Sartorio de[2](O); Oliveira, Tiago Almeida[3](CO); Micaroni, Isabela Rótoli[4]

[1]Graduando do curso de Engenharia Agronômica do CCA/UFSCar – Bolsista PIBIT ; [2]Docente do Departamento de informática e Estatística da UFSC; [3]Docente do departamento de estatística da UEPA; [4]Bacharel em Engenharia Agronômica do CCA/UFSCar

Introdução

O motivo da escolha da metodologia de MLG foi pela vantagem de não infringir as pressuposições do modelo, já que, se trata de uma variável discreta. Por isso, que de forma geral, as pressuposições do modelo clássico não são satisfeitas, além de supor, de forma rígida, que tal variável segue uma distribuição normal. Desta maneira, o intuito é de apresentar uma abordagem diferente das técnicas clássicas, sendo mais rica em informações e mais indicada de se tratar um banco de dados desta natureza, o que pode levar a resultados diferentes da abordagem clássica.

Serão utilizados dados de um experimento agronômico delineado inteiramente ao acaso, em que se observou a variável resposta “número de sementes por fruto” em quatro variedades de tangerina, onde foi feito 20 repetições. Assim realizou-se a análise da variável de contagem utilizando as metodologias: clássica e de MLG, para estudar e comparar os resultados obtidos. O objetivo do presente trabalho foi Estudar a metodologia de Modelos Lineares Generalizados (MLG), em particular, fazer uso das distribuições Poisson, ou de alguma de suas extensões, e da Binomial Negativa em comparação ao Modelo Clássico para explicar fenômenos que envolvem uma variável quantitativa discreta (contagem) como variável resposta. Além dos objetivos descritos anteriormente, propõe-se realizar mais quatro simulações (cenários) utilizando a amostragem aleatória estratificada, ou seja, diminuiremos o número de repetições desse experimento para 4, 5, 10 e 15 repetições;

Metodologia

Para tal abordagem, depois de fixadas as hipóteses e o nível de significância, foram realizadas as análises descritivas de todos os cenários, ou seja, média, variância e desvio padrão. O Modelo Clássico, necessitou verificar as pressuposições, dentre as características do experimento e as várias possibilidades de gráficos e testes. O MLG, por sua vez, possui pressuposições que se ajustam melhores aos dados e também foram verificadas a partir das características dos dados de contagem e o gráfico seminormais (*half normal plot*) de probabilidade com envelope simulado. Por final realizou-se o teste de comparação múltipla, no caso o Tukey, para verificar se existia pelo menos uma média do número de sementes inteiras que seja diferente das demais nos diferentes números amostrais.

Resultados e Discussão

No caso de cada amostra do modelo clássico, foram realizadas a análise descritiva dos dados, como média, variância e desvio padrão para os diferentes n amostrais (Tabela 1).

Tabela 1. Média, variância e desvio padrão de cada tratamento (Muscia, Híbrido, Pokan e Span) e coeficiente de variação (C.V.) em cada experimento (4, 5, 10, 15 e 20 repetições por tratamento).

N. repetição por tratamento	Tratamento	Média	Variância	Desvio padrão	C.V. (%)
4	Muscia	18,2500	18,6667	4,3204	26,22
	Hibrido	13,0000	3,5833	1,8929	
	Ponkan	15,7500	8,9167	2,9861	
	Span	11,5000	27,6667	5,2599	
5	Muscia	14,4000	19,8000	4,4497	29,36
	Hibrido	18,0000	12,5000	3,5355	
	Ponkan	14,4000	33,8000	5,8137	
	Span	15,6000	17,8000	4,2190	
10	Muscia	17,8000	25,7333	5,0728	32,29
	Hibrido	15,7000	30,2333	5,4984	
	Ponkan	14,9000	21,6557	4,6536	
	Span	12,4000	18,7111	4,3256	
15	Muscia	16,7333	21,9238	4,6822	33,64
	Hibrido	16,0000	22,1429	4,7056	
	Ponkan	14,5333	31,2668	5,5917	
	Span	11,0000	20,7143	4,5513	
20	Muscia	17,2000	23,3263	4,8297	32,34
	Hibrido	16,0000	20,2105	4,4956	
	Ponkan	14,7500	27,2500	5,2202	
	Span	12,3000	24,1158	4,9108	

Os erros só possuíram distribuição normal se o valor-*p* for maior que o nível de significância do estudo ($\alpha = 0{,}05$), o qual só não ocorreu para n=20 para o Teste de Shapiro-Wilk, porém para os testes de Lilliefors, Cramér-von Mises e Shapiro-Francia o valor-p foi maior, adotou-se como tendo normalidade.

A homogeneidade da variância e a presença de *outliers* foi verificado pelo Teste de Bartlett, Hartley e Levene, ou seja, o valor-p é maior que o nível de significância do estudo (**α** = 0,05) em todos os casos, mostrando que possivelmente as variâncias são homogêneas e não possuem *outliers* (Tabela 3).

No MLG, realizou-se os procedimentos semelhantes com os mesmos dados e mesmo número de repetições contento os mesmos valores. A partir disso definiu-se o tipo de distribuição e a sua função de ligação para cada n, levando em consideração o quão próximo o resíduo da deviance está dos graus de liberdade e também utilizando o gráfico de envelopes, tendo como permitido até 15% dos pontos fora do envelope.

Tabela 4. Valor do resíduo da deviance, dos graus de liberdade, o modelo que melhor se ajustou na repetição por tratamento, sua função de ligação e o parâmetro de dispersão da Binomial Negativa (θ).

Repetições por tratamento	Distribuição de probabilidade	Função de ligação	θ	Resíduo da deviance	Graus de liberdade
4	Poisson	log	-	14,4170	12
5	Binomial Negativa	log	193,1513	20,91	16
10	Binomial Negativa	log	34,4081	41,37	36
15	Binomial Negativa	log	24,2523	63,31	56
20	Binomial Negativa	log	27,2773	84,75	76

Como a ANOVA de ambos modelos foram aceitas, utiliza-se, para descobrir qual média difere-se das demais aplica-se o teste de comparação múltipla, foi escolhido o de Tukey, (Tabela 4).

Tabela 5. Valor-p da ANOVA e ANODEV para cada cenário, média, erro padrão por tratamento Muscia, Híbrido, Pokan e Span), a estimativa da média de cada tratamento pelo MLG e o resultado do Teste de Tukey a um nível de significância de 5%.

		Modelo clássico				MLG			
n_i	Variedade	Valor-p	Média	e.p.	Tukey	Valor-p	Média	e.p.	Tukey
4	Muscia	0,116	18,2500	1,0800	A	0,060	18,2506	0,1815	A
	Hibrido		13,0000	0,4732	A		13,0006	0,1170	A
	Ponkan		15,7500	0,7465	A		15,7509	0,1720	A
	Span		11,5000	1,3150	A		11,5006	0,1882	A
5	Muscia	0,573	14,4000	0,8899	A	0,480	14,4010	0,1645	A
	Hibrido		18,0000	0,7071	A		18,0005	0,1102	A
	Ponkan		14,4000	1,1628	A		14,4010	0,1645	A
	Span		15,6000	0,8438	A		15,6004	0,1613	A
10	Muscia	0,122	17,8000	0,5073	A	0,070	17,8003	0,1334	A
	Hibrido		15,7000	0,5498	A		15,7006	0,096	A
	Ponkan		14,9000	0,4653	A		14,9009	0,1374	A
	Span		12,4000	0,4325	A		12,4000	0,1422	A
15	Muscia	0,010	16,7333	0,3122	A	0,004	16,7333	0,1169	A
	Hibrido		16,0000	0,3137	A		16,0002	0,08323	A
	Ponkan		15,5333	0,3728	AB		15,5334	0,1194	AB
	Span		11,0000	0,3034	B		11,0005	0,1255	B
20	Muscia	0,010	17,2000	0,2415	A	0,009	17,1998	0,0984	A
	Hibrido		16,0000	0,2248	AB		16,0001	0,0704	A
	Ponkan		14,7500	0,2610	AB		14,7493	0,1009	AB
	Span		12,3000	0,2455	B		12,3000	0,1042	B

Conclusão

A mudança do número de repetições interfere diretamente na análise descritiva e na Análise da variância. Nota-se que, para ambos modelos, quando o n é menor ou igual a 10 não existe diferença estatística da média do número de sementes inteiras. Nas amostras de tamanho maiores (n = 15 e 20), em ambos, destacam-se o Híbrido e a Muscia como tendo uma média maior e por sua vez a Span é a que possui menor média do número de sementes inteiras, enquanto que a Ponkan se mostra estatisticamente igual aos dois grupos. Conclui-se que o MLG, para esse estilo de dados, se ajusta de maneira melhor, tendo em vista o atendimento das pressuposições e a flexibilidade do modelo, evitando erros estatísticos e ganhando maior exatidão nos resultados.

Referências Bibliográficas

DEMÉTRIO, C.G.B. **Modelos Lineares Generalizados em Experimentação** Agronômica. ESALQ/USP – Piracicaba, Mai. 2002.

MICARONI, I.R. **Estudo físico-químico e sensorial de tangerinas utilizando técnicas univariadas e multivariadas**. Monografia apresentada ao Curso de Engenharia Agronômica – CCA – UFSCar para a obtenção do título de Engenheira Agrônoma, 2018.