Modelos de regressão para dados de contagem inflacionados de zeros: Uma aplicação sobre o percevejo-do-colmo em arroz irrigado por inundação no sul do Brasil

Enio Júnior Seidel[1], Vera Lucia Damasceno Tomazella[2], Juliano de Bastos Pazini[3], Afrânio Márcio Corrêa Vieira[4]

1. Introdução

Os modelos de regressão para dados de contagem são muito utilizados nas mais variadas áreas de estudo para a modelagem de fenômenos. A distribuição de Poisson é a mais conhecida, e a mais utilizada para modelar dados de contagem, no entanto sempre que existe sobredispersão, torna-se necessário recorrer a outras distribuições. Outro problema comum nos dados de contagem é o excesso de zeros na variável resposta. Os modelos de regressão de zeros inflacionados são amplamente usados para modelar esse tipo de dados. Estes modelos modelam as contagens como uma mistura de duas distribuições com dois processos subjacentes, um que trata do excesso de zeros modelado por uma massa pontual, e um outro que trata das contagens sendo modelado por uma distribuição de Poisson, Binomial Negativa entre outras.

Neste trabalho são estudados os modelos de regressão para dados de contagem e a sua aplicação a dados de contagem de percevejo-do-colmo *Tibraca limbativentris* Stal, 1860 (Hemiptera: *Pentatomidae*). Este é um dos insetos mais prejudiciais à cultura do arroz irrigado por inundação no Brasil, sendo que sua ocorrência é maior na região do Planalto da Campanha do Rio Grande do Sul, onde predominam lavouras implantadas em áreas inclinadas. Esse tipo de lavoura, por exigir maior proximidade entre taipas, sobre as quais o arroz também é semeado, favorece o estabelecimento do inseto, podendo ser encontrado nas fases vegetativa e reprodutiva dos arrozais (REUNIÃO, 2018).

Recentemente, Pazini et al. (2015) e Pazini et al. (2017) estudaram o comportamento de distribuição espaço-temporal do inseto, no sul do Brasil, e identificaram que ocorrem agrupamentos espaciais de maior densidade em zonas preferenciais da lavoura. Contudo, os autores não consideraram covariáveis na modelagem.

Deste modo, o objetivo deste trabalho é obter um modelo de regressão para a contagem do percevejo-do-colmo em função do tempo (em dias) de implantação da lavoura.

2. Material e Métodos

O estudo foi realizado na safra de 2010/11, em lavoura comercial de arroz instalada num Luvissolo, com declividade média de 4,8%, na Fazenda Pitangueira, situada a 29°09’56.52”S e 56°29’20.06”W, em Itaqui, RS. O clima predominante da região é “Cfa”, subtropical, temperado quente, com chuvas bem distribuídas e estações bem definidas, segundo classificação de Köppen-Geiger.

1 UFSM e UFSCar. email: *enioseidel@gmail.com*.
2 UFSCar. email: *vera@ufscar.br*.
3 UFPel. email: *julianopazzini@hotmail.com*.
4 UFSCar. email: *afranio@ufscar.br*.

A cultura foi implantada via cultivo mínimo, semeando a cultivar IRGA 417, na primeira semana de outubro, na densidade de 60 sementes/m linear, num espaçamento de 0,17m entre linhas. A adubação foi de 286kg ha-1 de 4-17-27 N-P-K na semeadura, 150kg ha-1 de 45-0-0 N-P-K aos 15 dias pós-emergência das plântulas, antes da inundação do arrozal, e 75kg ha-1 de 30-0-20 N-P-K na diferenciação do primórdio floral. O manejo fitossanitário foi feito conforme as recomendações técnicas para a cultura do arroz irrigado (REUNIÃO, 2010), porém, sem aplicações de inseticidas. No pós-colheita do arroz, em meados de fevereiro, a lavoura foi submetida a duas passadas de grade aradora e uma de grade niveladora para destruição dos restos culturais (resteva).

Foram realizadas cinco amostragens durante a safra para monitorar a quantidade de insetos na área. Em todos os levantamentos de T. limbativentris, sobre cada ponto amostral, foi lançada uma estrutura metálica com 0,5m x 0,5m, sendo as plantas inseridas na área de 0,25m², examinadas visualmente para a contagem do percevejo-do-colmo.

Os dados foram submetidos à análise estatística por meio de gráficos, Tabelas e ajuste de modelos de regressão para contagens. A modelagem GAMLSS (*Generalized additive models for location, scale and shape*) foi utilizada por permitir uma quantidade maior de distribuições de probabilidade e ser bem flexível na modelagem dos parâmetros de interesse. Os modelos GAMLSS testados foram os que consideram as distribuições Poisson, Binomial negativa, Poisson inflacionada de zeros (ZIP), Poisson inflacionada de zeros reparametrizada (ZIP2) e Binomial negativa inflacionada de zeros (ZIBN). O modelo GAMLSS pode ser definido como:

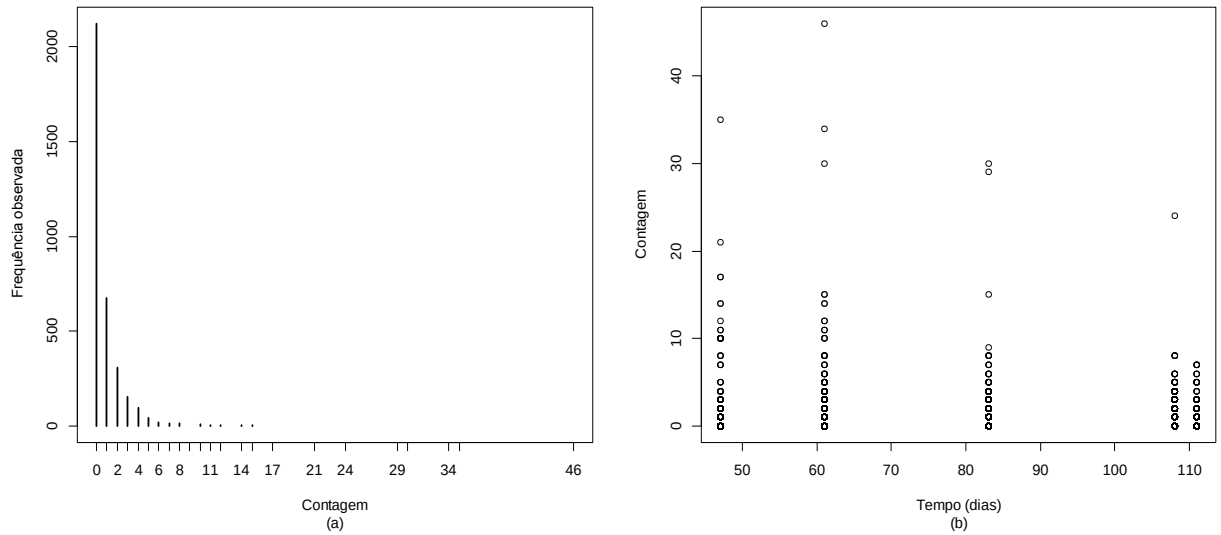
$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} Z_{jk}\gamma_{jk}$$

em que g(.) é uma função de ligação para o k-ésimo parâmetro θ, k=1,2, ..., K, η é o preditor linear, β é o vetor de parâmetros de efeitos fixos do modelo linear e X é uma matriz experimental, γ é uma variável aleatória e Z é uma matriz experimental. Mais detalhes sobre os modelos GAMLSS podem ser obtidos em Rigby e Stasinopoulos (2005).

Para a escolha do modelo mais adequado utiliza-se o critério de informação de Akaike (AIC). A análise estatística foi realizada no pacote GAMLSS (RIGBY; STASINOPOULOS, 2005) do programa R (R CORE TEAM, 2019).

3. Resultados e Discussão

A Figura 1 mostra o comportamento da contagem dos insetos e sua relacão com o tempo de implantação da lavoura (em dias).



Fonte: autores

Figura 1 – Frequência observada da contagem do inseto (a) e relacionamento da contagem com o tempo de implantação da lavoura (em dias) (b).

Percebesse que ocorrem muitos zeros nos dados (Figura 1a), ou seja, existe uma inflação de zeros nas contagens realizadas. Além disso, observasse que a contagem do percevejo-do-colmo decresce conforme evolui o tempo de implantação da lavoura na safra (Figura 1b). Este número elevado de zeros observados leva a uma sobredispersão dos dados com variância maior que a média, conforme já observado em Pazini et al. (2017).

Como observa-se na Figura 1, é necessário considerar um modelo de regressão que modele o excesso de zeros. Contudo, para fins de comparação, serão considerados modelos de contagem sem inflação de zeros e com inflação de zeros. Além disso, serão considerados os modelos somente com efeitos fixos e, após, com inclusão de efeitos aleatórios, já que as contagens foram feitas nas mesmas unidades experimentais em cinco datas distintas no decorrer da safra. A Tabela 1 mostra as estimativas para os modelos de regressão com efeitos fixos e o critério AIC.

Tabela 1 – Estimativas dos parâmetros dos modelos de regressão Poisson, Binomial negativo (BN), Poisson inflacionado de zeros (ZIP), Poisson inflacionado de zeros 2 (ZIP2) e Binomial negativo inflacionado de zeros (ZIBN), via modelagem GAMLSS, com efeitos fixos, e critério AIC.

Modelo	Log(Média)		Log(DP)[#]	Logit(Prop)[##]		AIC
	Intercepto	Tempo	Intercepto	Interpecto	Tempo	
Poisson	-0,218*	0,003*				12158
BN	-0,233*	0,002[ns]	0,923*			9016
ZIP	1,341*	-0,008*		1,856*	-0,021*	10173
ZIP2	-0,092[ns]	0,001[ns]		1,866*	-0,021*	10157
ZIBN	1,191*	-0,012*	0,458*	4,636*	-0,086*	8850

[#]DP=Desvio padrão. [##]Prop=Proporção de zeros.

Fonte: autores

Verificasse na Tabela 1 que o modelo de regressão Binomial negativo inflacionado de zeros apresentou o menor valor no critério AIC, indicando ser o modelo mais adequado.

Já a Tabela 2 mostra as estimativas dos parâmetros para os modelos considerando a inclusão de efeitos aleatórios.

Tabela 2 – Estimativas dos parâmetros dos modelos de regressão Poisson, Binomial negativo (BN), Poisson inflacionado de zeros (ZIP), Poisson inflacionado de zeros 2 (ZIP2) e Binomial negativo inflacionado de zeros (ZIBN), via modelagem GAMLSS, considerando efeitos aleatórios, e critério AIC.

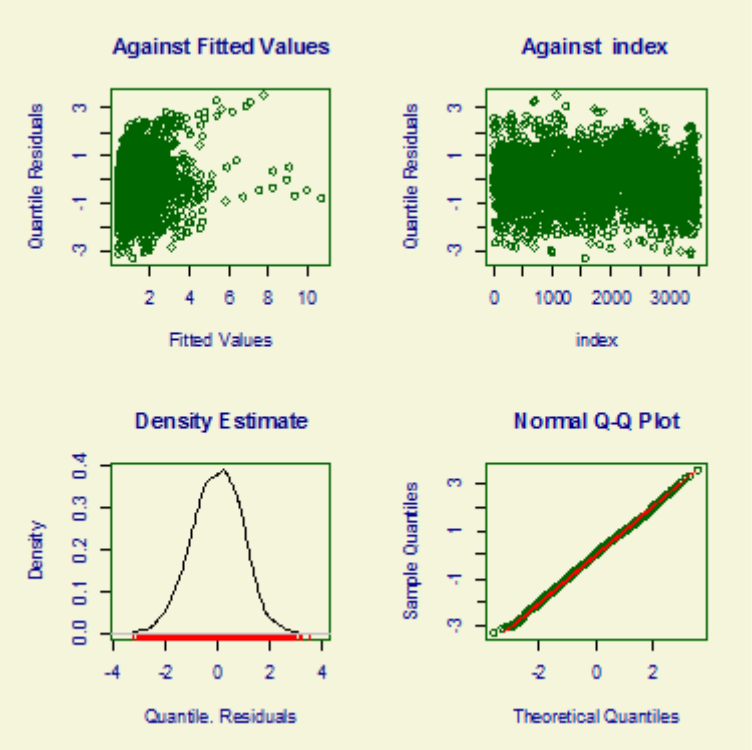
Modelo	Log(Média)			Log(DP)[#]	Logit(Prop)[##]			AIC
	Intercepto	Tempo	EA[###]	Intercepto	Interpecto	Tempo	EA	
Poisson	-0,293*	0,002*	0,567					10764
BN	-0,624*	0,005*	0,492	0,536*				8914
ZIP	1,212*	-0,011*	0,902		2,193*	-0,030*	0,016	8777
ZIP2	-0,354*	0,002[ns]	0,828		1,923*	-0,026*	0,0002	8825
ZIBN	1,214*	-0,013*	0,715	-0,565*	4,216*	-0,071*	0,468	8626

[#]DP=Desvio padrão. [##]Prop=Proporção de zeros. [###]EA=Efeito aleatório.
Fonte: autores

Verificasse na Tabela 2 que o modelo de regressão Binomial negativo inflacionado de zeros apresentou o menor valor no critério AIC, indicando ser o modelo mais adequado. Como os modelos BN e ZIBN possuem um parâmetro de dispersão, este foi modelado apenas por um intercepto de efeito fixo, de modo a considerer o efeito do tempo e o efeito aleatório somente nos parâmetros de média e de proporção de zeros.

Comparativamente, considerando as Tabelas 1 e 2, tem-se que o modelo de regressão ZIBN, com efeitos aleatórios, apresentou o menor valor no critério AIC, indicando ser o modelo mais adequado para modelar a contagem do percevejo-do-colmo em função do tempo de implantação da lavoura (em dias). Almeida et al. (2016), ao estudar a contagem de plantas doentes em pomares ao longo do tempo, também verificaram que o modelo Binomial negativo inflacionado de zeros foi o mais adequado.

Para complementar a avaliação do modelo ZIBN com efeito aleatório, considera-se a Figura 2, na qual são apresentados gráficos de avaliação dos resíduos do modelo. Percebe-se que os resíduos tem distribuição normal, indicando um bom ajuste do modelo.



Fonte: autores

Figura 2 – Gráficos de avaliação dos resíduos do modelo ZIBN com efeito aleatório.

4. Conclusão

O modelo de regressão Binomial negativo inflacionado de zeros, com efeitos aleatórios, foi o que melhor descreveu a contagem do percevejo-do-colmo como função do tempo de implantação da lavoura.

O tempo (em dias) de implantação da lavoura teve efeito significativo tanto na média quanto na proporção de zeros observados nas contagens do percevejo-do-colmo.

Referencias Bibliográficas

ALMEIDA, E. P.; JANEIRO, V.; GUEDES, T. A.; MULATI, F.; CARNEIRO, J. W. P.; NUNES, W. M. C. Modeling citrus huanglongbing data using a zero-inflated negative binomial distribution. Acta Scientiarum. Agronomy, v. 38, n. 3, p.299-306, 2016.

PAZINI, J. B.; BOTTA, R. A.; SEIDEL, E. J.; SILVA, F. F.; MARTINS, J. F. S.; BARRIGOSSI, J. A. F.; RÜBENICH, R. Geoestatística aplicada ao estudo da distribuição espacial de Tibraca limbativentris em arrozal irrigado por inundação. Ciencia Rural, v. 45, n. 6, p.1006-1012, 2015.

PAZINI, J. B.; SEIDEL, E. J.; SILVA, F. F.; BARRIGOSSI, J. A. F.; MARTINS, J. F. S.; BOTTA, R. A. Validação do arranjo espacial do percevejo-do-colmo em arroz irrigado por inundação. Ciência e Natura, v. 39, n. 2, p.221-232, 2017.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2019. ISBN 3-900051-07-0, URL http://www.R-project.org/.

REUNIÃO TÉCNICA DA CULTURA DO ARROZ IRRIGADO, 28. 2010. Arroz irrigado: recomendações técnicas da pesquisa para o Sul do Brasil. Porto Alegre: SOSBAI, 2010. 188p.

REUNIÃO TÉCNICA DA CULTURA DO ARROZ IRRIGADO, 32. 2018. Arroz irrigado: recomendações técnicas da pesquisa para o Sul do Brasil. Cachoeirinha: SOSBAI, 2018. 205p.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: SeriesC (AppliedStatistics), v.54, n.3, p.507–554, 2005.