Modelagem de dados subdispersos via distribuição Contagem Gama

Walmes Marques Zeviani, Wagner Hugo Bonat, Paulo Justiniano Ribeiro Jr*,
LEG/DEST - Paraná Federal University

Resumo

Variáveis aleatórias de contagem assumem valores inteiros não negativos que representam o número de vezes que um evento ocorre. Análise de contagens por modelos de regressão gaussianos não são eficientes pois não consideram assimetria e heterocedásticidade. O modelo de regressão Poisson é largamente aplicado à dados de contagem mas se a suposição de equidispersão não for verificada, as estimativas dos parâmetros apresentarão erros padrões inconsistentes. Uma abordagem para tratar a subdispersão é o modelo contagem gama, que a considera a custo da estimação de um parâmetro extra. Neste artigo comparamos através de um exemplo o desempenho do modelo Poisson com o contagem gama. A abordagem semi-paramétrica Quasi-Poisson também foi considerada para comparação dos resultados. Inferência para os parâmetros envolvidos nos modelos foi feita por máxima verossimilhança. Os resultados mostram que o modelo contagem gama é tão flexível quanto o Quasi-Poisson. O modelo Poisson mostrou-se inadequado para a situação de dados subdispersos sendo muito conservador. Recomenda-se o uso do modelo contagem gama para modelagem de contagens subdispersas.

keywords: poisson, likelihood inference, gamma-count, subdispersão, quasi-Poisson, algodão

*Corresponding author: paulojus@leg.ufpr.br, Dpto Estatística-UFPR, CP 19.081, Curitiba, PR Brazil, 81.531-990

1 Introduction

Modelos de regressão têm sido aplicados na análise de dados nas mais diversas áreas da ciência. O modelo de regressão linear (Gaussiano) é sem dúvida o modelo mais popular entre usuários de estatística aplicada. Dentre a grande variedade de aplicações é comum encontrar situações onde a variável aleatória de interesse (resposta) se apresenta na forma de contagens. De forma geral, contagens são variáveis aleatórias que assumem valores inteiros e não negativos, representam o número de vezes que um evento ocorre em um domínio fixo que pode ser contínuo, como um intervalo de tempo ou espaço, ou discreto, como a avaliação de um indivíduo ou setor censitário.

A construção de modelos de regressão para dados de contagens por modelos de regressão Gaussianos não é eficiente, tem erros padrões inconsistentes e podem produzir predições negativas para o número esperado de eventos (King, 1989). Isso ocorre, porque o modelo Gaussiano não considera o fato do dado ser discreto, heterocedástico, assimétrico e não negativo, características inerentes à dados de contagem. Esses problemas, são agravados quando o tamanho amostral é reduzido e as contagens são baixas.

Com a introdução dos modelos lineares generalizados por Nelder and Wedderburn (1972), a análise padrão para dados desta natureza passou a ser o modelo de regressão de Poisson. A distribuição de Poisson, considera naturalmente dados assimétricos e heterocedástico, além disso tem como seu domínio os naturais positivo o que a torna uma opção muito atraente para modelar dados de contagens. A distribuição de Poisson é indexada por um parâmetro λ que é a sua esperança e também sua variância, ou seja, a sua média é igual a sua variância. Esta particularidade do modelo Poisson, impõe algumas restrições quando constrõe-se modelos de regressão utilizando esta distribuição. O modelo impõe a suposição de equidispersão (média igual a variância) o que pode não ser adequado para diversas situações. Se o modelo Poisson for aplicado em situações em que a equidispersão não é verificada, estimativas dos parâmetros serão ineficientes e erros padrões inconsistentes (Winkelmann and Zimmermann, 1994), (Winkelmann, 1995).

O caso mais comum de fuga da equidispersão é a superdispersão, quando a variância é maior do que a média, que pode ocorrer pela ausência de covariáveis importantes, heterogeneidade de unidades amostrais, níveis de amostragem, excesso de zeros, entre outros (Grunwald et al., 2011). Neste caso, a abordagem padrão é adotar modelos com a presença de efeitos aleatórios, na estrutura dos Modelos Lineares Generezalizados Mistos (MLGM) que são capazes de descrever a variabilidade extra pela inclusão de variáveis latentes não observadas, responsáveis pelo excesso de variabilidade. Um caso interessante desta abordagem é quando tem-se um modelo Poisson com efeito aleatório Gama, que gera um modelo Binomial Negativo, uma das abordagens mais comuns e eficientes para modelar superdispersão, porém outras alternativas estão disponíveis conforme apresentado em El Shaarawi et al. (2011).

Uma situação menos comum é a presença de subdispersão, quando a variância é menor que a média. As explicações para este tipo de fenômeno são mais escassas e dependem muito da área da ciência em que se está atuando. A explicação geral encontra-se na origem do modelo Poisson, de que se os tempos entre eventos têm distribuição Exponencial, o número de eventos em um intervalo será Poisson. Porém, existem situações onde a ocorrência de um evento pode aumentar ou diminuir a probabilidade de outro evento acontecer tornando a suposição de tempos entre eventos exponencial inadequada, e consequentemente gerando contagens super ou subdispersas. Atribuir outra distribuição de probabilidade para os tempos entre eventos é uma abordagem comum, as mais usadas são a Gama (Winkelmann, 1995), (Toft et al., 2006), Lognormal (Gonzales-Barron and Butler, 2011) e Weibull (McShane et al., 2008).

Outras opções de abordagens são baseadas em ponderações na distribuição de Poisson (Ridout and Besbeas, 2004), a distribuição COM-Poisson (Lord et al., 2010), (Lord et al., 2008), distribuições com caudas pesadas (Zhu and Joe, 2009), entre diversas outras na literatura, para uma visão geral ver (Winkelmann, 1995).

Tratar da subdispersão em dados de contagens é mais desafiador do que tratar a superdispersão, uma vez que gerar variabilidade é mais simples porque não envolve mudança

de modelos, apenas a inclusão de efeitos aleatórios que são modelos bem estabelecidos na literatura. Tratar a subdispersão exige uma mudança mais profunda no modelo de análise. Winkelmann (1995) explorou a conexão entre modelos de contagem e modelos para duração (sobrevivência) para flexibilizar a suposição de equidispersão à custo da estimação de um parâmetro adicional. A proposta deste autor é flexibilizar o modelo Poisson, trocando a distribuição geradora dos tempos entre eventos de exponencial para Gama, uma vez que quando o parâmetro de forma da Gama $\alpha = 1$ tem-se a distribuição exponencial, tem-se que o modelo Poisson é um caso particular do modelo denominado Contagem Gama. Não fixando o parâmetro $\alpha = 1$ e sim estimando este parâmetro, tem-se uma distribuição de probabilidade bastante flexível capaz de modelar subdispersão, quando $\alpha > 1$ e também superdispersão quando $0 < \alpha < 1$. Para detalhes da construção do modelo contagem Gama ver (Winkelmann, 1995).

O objetivo deste artigo é comparar o desempenho do modelo contagem Gama com o modelo Poisson, através de uma aplicação a dados de cultura do algodão. A variável resposta neste experimento é o número de capulhos produzidos por planta, por condições da cultura é esperado que tais tipos de contagens apresentem subdispersão. O modelo de regressão contagem Gama, não é comum entre usuários de estatística aplicada e ainda não figura entre as rotinas de análise de softwares estatísticos. Sendo assim, foi implementada a estimação dos parâmetros deste modelo pelo método de Máxima Verossimilhança. Funções genéricas para a estimação dos parâmetros deste modelo são disponibilizadas no complemento *on line* do artigo.

Dado o pouco uso deste tipo de modelo na literatura aplicada, frente a grande quantidade de dados que podem apresentar subdispersão é também um objetivo deste artigo alertar a pesquisadores das mais diversas áreas da ciência a presença deste tipo de dados, os problemas que a análise convencional via o modelo Poisson pode apresentar e mostrar uma forma efetiva de analizar dados subdispersos. Aspectos referentes a inferência sobre os parâmetros do modelo contagem Gama, tais como, construção de intervalos de confiança, sejam assintóticos

ou perfilhados, condução de testes de hipóteses e critérios para comparação de modelos são discutidos através da analise do conjunto de dados. Optou-se também por analisar os dados por uma abordagem semi-paramétrica usando como modelo de benchmark a ser superado o modelo Quasi-Poisson.

O presente artigo encontra-se dividido da seguinte forma: esta primeira seção objetiva discutir sobre a construção de modelos para dados de contagens suas possíveis limitações e algumas abordagens para solucioná-las e apresenta os principais objetivos do artigo. A segunda seção apresenta o *framework* geral para a construção de modelos de regressão para dados de contagens, apresentando as duas opções de distribuições para a variável resposta que serão consideradas, Poisson e Contagem Gama. A seção 3 apresenta o conjunto de dados utilizado nas análises. A seção 4 apresenta os principais resultados da aplicação dos modelos ao conjunto de dados, destancando as principais diferenças, vantagens e desvantagens de cada abordagem. Por fim, a seção 5 apresenta as principais conclusões e recomendações, bem como, possíveis pontos para serem melhor elucidados por trabalhos futuros.

2 Background

A construção de modelos de regressão para dados de contagem seguindo a distribuição Poisson é facilmente feita usando a estrutura dos modelos lineares generalizados. Porém, quando trata-se a subdipersão/superdispersão por modelos alternativos como o caso do modelo contagem Gama, a construção é facilitada pensando em tempos entre eventos que é como se origina a distribuição de Poisson.

Considere que os tempos entre eventos são independentes, mas não exponenciais (que leva a distribuição Poisson), ao invés disso tem-se alguma outra distribuição com um função de risco não constante, neste caso considera-se a distribuição Gama. Considere a Figura 1.

Note que nos três casos a distribuição entre os eventos é Gama com a mesma média, porém com variâncias diferentes. No primeiro caso a variância é $5/5^2$ consequentemente a função

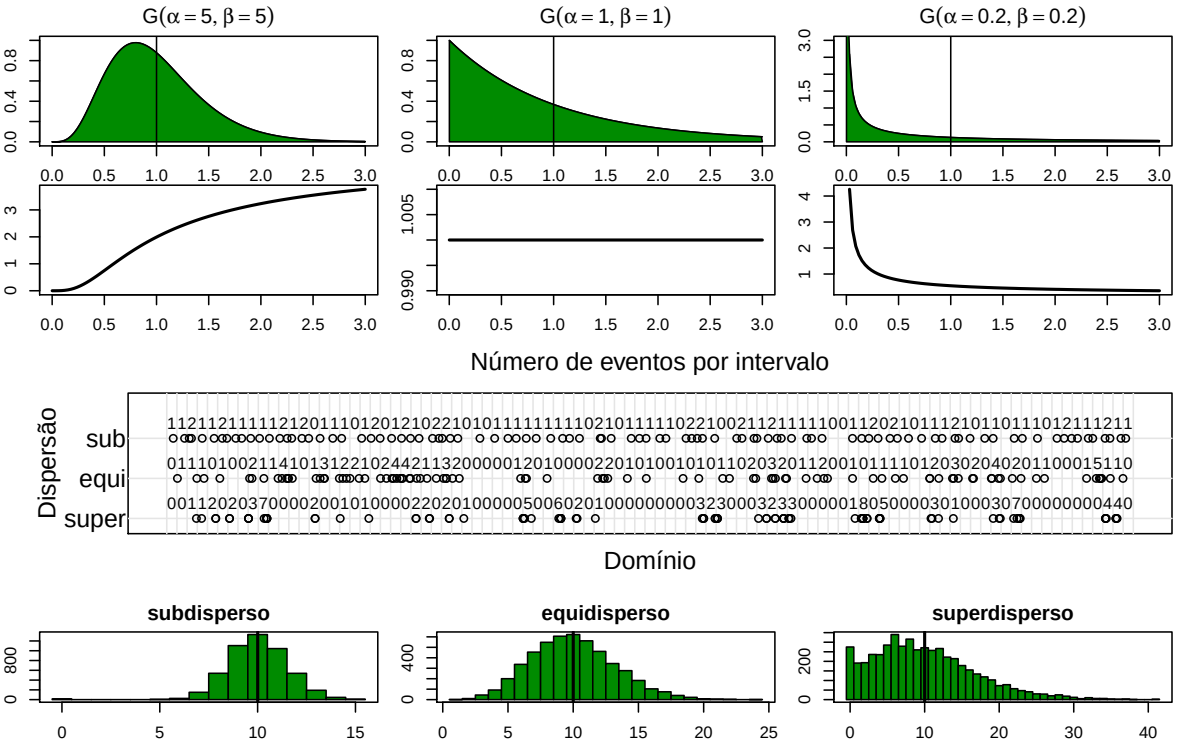

Figura 1: Representação do tipo de contagem de acordo com a distribuição de probabilidade para o tempo entre eventos.

de risco é crescente o que gera contagens subdispersas. Por outro lado, na situação onde a variância é grande $0.2/0.2^2$ a função de risco é decrescente gerando contagens superdispersas. Na situação canônica quando a variância é 1, que é quando $\alpha = \beta = 1$ tem-se que a Gama toma a forma da distribuição exponencial, a função de risco é constante gerando contagens equidispersas. Esta figura também mostra claramente a grande limitação da distribuição Poisson para modelar contagens, já que pode-se ter três situações de contagens e a Poisson só é adequada para uma delas.

Winkelmann (1995) discute a relação existente entre os tempos de ocorrências e as contagens chegando ao modelo contagem Gama. Para a estimação dos parâmetros envolvidos no modelo, é utilizado o método da máxima verossimilhança. A função de log-verossimilhança

é

$$l(\beta, \alpha; y) = \sum_{i=1}^{n} \log \left(G(\alpha y_i, \alpha \exp(x_i^\top \beta)) - G(\alpha y_i + \alpha, \alpha \exp(x_i^\top \beta)) \right) \quad (1)$$

em que β é o vetor de parâmetros do modelo de regressão que descreve o intervalo entre eventos, α é o parâmetro de dispersão, y é o vetor da amostra observada, x é a matriz de covariáveis associada às observações e $G()$ é a função densidade acumulada da distribuição Gama.

Por esta equação também fica claro que o modelo Poisson é um caso particular do contagem Gama, quando $\alpha = 1$ que é a diferença entre duas exponenciais.

Neste artigo a estimação foi feita maximizando numericamente a equação 1, usando o algoritmo BFGS (Byrd, 1995) implementado na função *optim*() do software estatístico R Development Core Team (2012). Para a construção dos intervalos de confiança foram considerados intervalos baseados em aproximação quadrática da verossimilhança (intervalos de Wald) e também intervalos baseados em perfil de verossimilhança para fins de comparação. Códigos genericos para os ajustes são disponibilizados nos complementos *on line*.

3 Conjunto de dados

O conjunto de dados utilizado na analise corresponde a um experimento em casa de vegetação com plantas de algodão *Gossypium hirsutum* submetidas à níveis de desfolha artificial $(0, 25, 50, 75, 100\%)$ de remoção foliar, em combinação com o estágio fenológico no qual a desfolha foi aplicada (vegetativo, botão floral, florescimento, maça e capulho) em um delineamento inteiramente casualizado com cinco repetições (da Silva et al., 2012). A unidade experimental foi um vaso com duas plantas onde foi avaliado o número de capulhos produzidos no final do ciclo da cultura. Uma análise exploratória deste conjunto de dados visa encontrar indicios de subdispersao, para isto um gráfico da média amostral em função da variância amostral, é uma boa ferramenta. A Figura 2 apresenta gráficos descritivos e um

diagrama de dispersão cruzando a média e variância amostral para o número de capulhos em cada combinação de nível de desfolha e estágio fenológico do algodão.

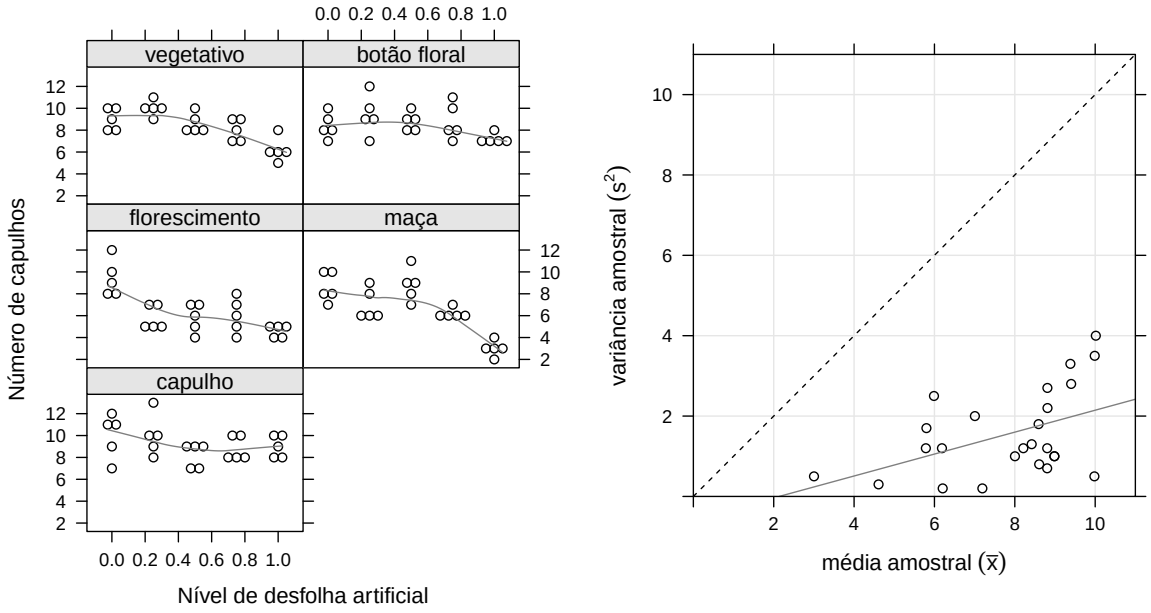

Figura 2: Variância amostral em função da média amostral para o número de capulhos em cada combinação de nível de desfolha e estágio fenológico do algodão. Valores são baseados em cinco repetições.

A Figura 2 aponta evidência de subdispersão, onde todos os pontos estão abaixo da linha de equidispersão. Como trata-se de dados experimentais existem diversos efeitos que precisam ser considerados para que uma decisão seja tomada. Diante da estrutura do experimento propomos uma sequência de cinco modelos crescente em complexidade que testam diversos aspectos interessantes sobre os fatores experimentais envolvidos no experimento. Em cada uma destas estruturas de modelos ajustamos os modelos com a distribuição de Poisson e Contagem Gama, para fins de comparação consideramos o valor da log verossimilhança maximizada, e o critério de *Akaike*, a estrutura dos modelos ajustados são as seguintes:

Model 1 - $g(\mu) = \beta_0$;

Model 2 - $g(\mu) = \beta_0 + \beta_1 def$ (efeito de primeira ordem da desfolha);

Model 3 - $g(\mu) = \beta_0 + \beta_1 def + \beta_2 def^2$ (efeito de segunda ordem da desfolha);

Model 4 - $g(\mu) = \beta_0 + \beta_{1i} def + \beta_2 def^2$ (efeito de desfolha para cada estágio de crescimento);

Model 5 - $g(\mu) = \beta_0 + \beta_{1i} def + \beta_{2i} def^2$ (efeito de segunda ordem de desfolha para cada estágio de crescimento).

Como é claro a estrutura dos modelos foi construída de forma aninhada para facilitar a condução de testes de hipóteses baseados em razão de verossimilhanças. Cada modelo foi proposto pensando em testar diferentes aspectos com relação a estrutura dos efeitos contidos no experimento. O modelo 1 contêm apenas o intercepto, é ajustado apenas como ponto de partida para verificar como modelos mais estruturados melhoram o ajuste. O modelo 2 apresenta apenas o efeito de desfolha de forma linear, o modelo 3 é o modelo 2 somado um efeito de desfolha de segunda ordem. O modelo 4, apresenta o efeito de desfolha linear mudando de acordo com o estágio de crescimento, e por fim o modelo 5 diz que não somente o efeito de primeira ordem muda com o estágio de crescimento, mais também o efeito de segunda ordem. Na próxima seção apresentamos os resultados dos ajustes dos modelos com a distribuição Poisson e Contagem Gama.

Para fins de comparações e confirmação dos resultados, optamos por ajustar também um modelo quasi-Poisson com cada uma das estruturas apresentadas. Como esta é uma classe semi-paramétrica de modelos, tende a ser mais geral que a abordagem estritamente paramétrica, caso dos modelos Poisson e Contagem Gama. O objetivo é mostrar que com um modelo paramétrico podemos atingir resultados tão bons quanto com o semi-paramétrico, porém com todas as vantagens de assumir um modelo paramétrico. (ESTE PARAGRAFO REPETE UM MONTE DE VEZ A MESMA COISA TEM QUE ESCREVER MELHOR). Para este ajuste foi utilizado a função $glm()$ do pacote R.

4 Resultados

Os resultados, em termos de medidas de ajuste e comparação entre sub-modelos, pelas três abordagens estão apresentados na tabela 1. O modelo contagem-Gama apresentou a maior log-verossimilhança comparado com o Poisson. Mesmo para o modelo nulo, sem uso de covariáveis, o contagem-Gama rejeitou a hipótese de equidispersão pelo teste da razão de verossimilhança. A estimativa de $\hat{\alpha} > 1$ indica que o número de capulhos produzidos apresenta duração dependência positiva, que implica em subdispersão. Assim, a probabilidade de surgimento de um novo capulho em uma planta aumenta com o passar do tempo pois a função de risco é crescente. Tal resultado apoia a hipótese de distribuição regular do número de capulhos por planta.

O modelo quasi-Poisson também apontou subdispersão mesmo no modelo nulo. No entando, o nível descritivo dos testes de hipótese para o parâmetro de dispersão (ϕ e α) foram maiores para o contagem-Gama (tabela 1). Ao contrário dos demais modelos, o modelo Poisson não apontou significância para o sub-modelo 5 o que pode ser atribuída a suposição de equidispersão.

O modelo contagem-Gama apresentou maior nível descritivo para comparação entre sub-modelos 3 vs 4 e 4 vs 5 ao passo que o quasi-Poisson o fez para 1 vs 2 e 2 vs 3. Estes resultados não permitem apontar que um dos modelos é mais conservador que o outro em termos de teste de hipótese para termos do modelo de regressão. Porém, fica evidente que o modelo Poisson, quando aplicado na presença de subdispersão, se torna conservador.

Os modelos contagem-Gama e quasi-Poisson apontaram efeito do estágio fenológico e nível de desfolha em relação ao número de capulhos produzidos. O efeito da desfolha, tanto de primeira quanto de segunda ordem, sob número de capulhos depende do estágio em que ela ocorre. Pela leitura da tabela 2 e figura 3 verifica-se não haver efeito de desfolha nos estágios botão floral e capulho. A razão entre a estimativa e o erro padrão nesses estágios foram, em valor absoluto, menor 1,96, valor de corte baseado na distribuição normal para nível nominal de significância de 5%. O modelo Poisson apenas apontou efeito de desfolha

Tabela 1: Medidas de ajuste e comparação entre sub-modelos.

Poisson	np	lv	AIC	dif np	2(dif lv)	P($>\chi^2$)		
1	1	-279,933	561,866					
2	2	-272,001	548,001	1	15,864	6,805E-05		
3	3	-271,354	548,709	1	1,293	2,556E-01		
4	7	-258,674	531,348	4	25,360	4,258E-05		
5	11	-255,803	533,606	4	5,742	2,193E-01		
Quasi-Poisson	np	deviance		dif np	dif dev	P($>F$)	$\hat{\phi}$	P($>\chi^2$)[a]
1	1	75,514					0,567	3,660E-04
2	2	59,650		1	34,214	4,235E-08	0,464	5,134E-07
3	3	58,357		1	2,810	9,630E-02	0,460	3,661E-07
4	7	32,997		4	22,768	7,676E-14	0,278	9,154E-16
5	11	27,255		4	5,956	2,241E-04	0,241	3,566E-18
Gamma-count	np	lv	AIC	dif np	2(dif lv)	P($>\chi^2$)	$\hat{\alpha}$	P($>\chi^2$)[a]
1	2	-272,396	548,792				1,764	1,034E-04
2	3	-257,350	520,701	1	30,092	4,121E-08	2,266	6,198E-08
3	4	-255,981	519,962	1	2,738	9,796E-02	2,317	2,940E-08
4	8	-220,145	456,291	4	71,671	1,007E-14	4,206	1,661E-18
5	12	-208,386	440,773	4	23,518	9,976E-05	5,112	2,071E-22

np - número de parâmetros; lv - log-verossimilhança; dif np - diferença em np; dif lv - diferença em lv; dif dev - diferença na deviance escalonada; [a] teste de hipótese para o parâmetro de dispersão ser igual à 1.

para o estágio de florescimento. Os demais modelos apontaram efeito de desfolha no estágio vegetativo, florescimento e maça.

Tabela 2: Estatimativas dos parâmetros e razão estimativa/erro padrão pelos três modelos.

	Poisson		quasi-Poisson		Gamma-count	
Parâmetro	Estimativa	Est/EP	Estimativa	Est/EP	Estimativa	Est/EP
β_0	2,1896	34,5724*	2,1896	70,4205*	2,2342	79,7128*
$\beta_{1vegetativo}$	0,4369	0,8473	0,4369	1,7260	0,4122	1,8080
$\beta_{2vegetativo}$	-0,8052	-1,3790	-0,8052	-2,8089*	-0,7628	-2,9544*
β_{1botao}	0,2897	0,5706	0,2897	1,1622	0,2744	1,2224
β_{2botao}	-0,4879	-0,8613	-0,4879	-1,7544	-0,4642	-1,8534
$\beta_{1floresc}$	-1,2425	-2,0581*	-1,2425	-4,1921*	-1,1821	-4,4348*
$\beta_{2floresc}$	0,6728	0,9892	0,6728	2,0149*	0,6453	2,1486*
β_{1maca}	0,3649	0,6449	0,3649	1,3135	0,3198	1,2797
β_{2maca}	-1,3103	-1,9477	-1,3103	-3,9672*	-1,1990	-4,0385*
$\beta_{1capulho}$	0,0089	0,0178	0,0089	0,0362	0,0070	0,0315
$\beta_{2capulho}$	-0,0200	-0,0361	-0,0200	-0,0736	-0,0185	-0,0756
α	-	-	-	-	5,1120	7,4228*

** indica $|\text{Est/EP}| > 1,96$. Função de ligação usada é a logaritmica.

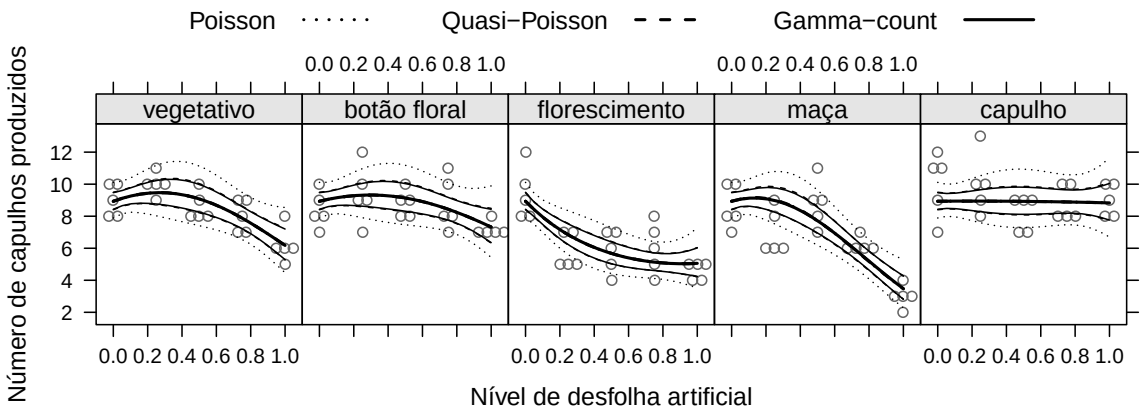

Figura 3: Diagrama de dispersão dos valores observados e da curva de valores preditos seguidos do intervalo de confiança (95%) em função do nível de desfolha artificial em cada estágio fenológico.

O estágio de florescimento apresentou estimativas com sinal contrário aos outros estágios. O termo linear negativo e significativo indica rápida queda no número de capulhos com o início da desfolha. O termo quadrático positivo indica concavidade para cima conforme vemos na Figura 3 para o estágio florescimento. Verifica-se portanto o maior impacto da desfolha no estágio de florescimento e uma certa tolerância à aproximadamente 40% de desfolha para o estágio vegetativo e de maça.

No modelo contagem-Gama, os parâmetros estimados se referem a distribuição do tempo entre eventos e não ao número de eventos como é para Poisson e quasi-Poisson. Apesar de não observado, a distribuição do tempo entre eventos pode ser estimada por meio do número observado de eventos aplicando o modelo contagem-Gama. Essa é uma propriedade interessante do modelo, uma vez que o número de eventos é uma variável aleatória mais fácil de registrar. Os parâmetros estimados entre o contagem-Gama e Poisson só terão mesma interpretação quando $\alpha = 1$.

A precisão das estimativas, de forma geral, foram maiores para modelo contagem-Gama, uma média de 3% em ordem de magnitude. Pórem, o significado das estimativas é diferente conforme já discutido no parágrafo anterior.

A precisão dos valores preditos foi a mesma entre o contagem-Gama e quasi-Poisson (Figura 3) o que pode ser observado pela sobreposição dos intervalos de predição. Tal propriedade não foi abordada nos demais estudos sobre modelos de duração dependência e portanto é uma contribuição desse trabalho. O modelo Poisson, por sua vez, apresentou intervalos amplos pela imposição da suposição de equidispersão.

Como pode ser visto com a apresentação deste resultados, o modelo contagem Gama e o Quasi-Poisson apresentam resultados muito parecidos em quase todos os aspectos de inferência (estimativas pontuais e intervalares, testes de hipóteses, comparações de modelos e bandas de predição). Isso mostra a grande flexibilidade do modelo contagem Gama, dado que o modelo Quasi-Poisson é um modelo semi-paramétrico é esperado que ele tenha um ajuste melhor para um particular conjunto de dados, uma vez que não há a declaração explicita de um modelo de probabilidade. Porém esta flexibilidade é acompanhada de diversos inconvenientes, por exemplo, a comparação entre submodelos e principalmente não é possível obter a distribuição de probabilidade das contagens, o que pode ser de grande interesse em diversas áreas da ciência.

Para mostrar esta vantagem do modelo contagem Gama foram calculadas as distribuições de probabilidade do número de capulhos no nível zero de desfolha para os modelos Poisson e contagem-Gama. No nível de desfolha zero o valor esperado é 8,93 capulhos por 2 plantas. Para o modelo contagem-Gama a distribuição de probabilidades mais concentrada ao redor desse valor médio. Embora o modelo quasi-Poisson tenha apresentado desempenho semelhante ao contagem-Gama nos aspectos mencionados, não é possível obter distribuição de probabilidades pois esse modelo é semi-paramétrico. O modelo contagem-Gama é totalmente paramétrico, e como já destacado, mostrou-se tão flexível na modelagem dos dados quanto o quasi-Poisson.

Outro aspecto relevante é em relação ao termo de dispersão que pela abordagem Contagem Gama é possível modelar esse parâmetro via modelo de regressão, em outras palavras, assim como podemos modelar o valor esperado, também podemos modelar a dispersão como função

de covariáveis. O que claramente não pode ser feito, pelo menos de forma simples no modelo Quasi-Poisson.

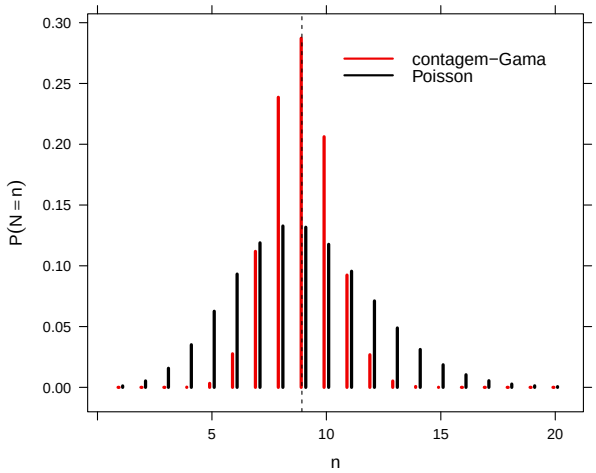

Figura 4: Probabilidades pelo modelo Poisson e contagem-Gama para o nível zero de desfolha. Para ambos o valor esperado é 8,93 capulhos.

Dado a relativa pouca aplicabilidade de modelos contagem Gama em dados agronômicos é interessante verificar alguns aspectos em relação a inferência por verossimilhança para os parâmetros do modelo. O primeiro aspecto interessante é verificar o formato do perfil da verossimilhança do parâmetro α, já que, este parâmetro é o responsável por mensurar a sub ou superdispersão dos dados.

O perfil de log-verossimilhança para o parâmetro α apresentou leve assimetria à direita (figura 5). O intervalo de confiança (95%) baseado na distribuição χ^2 foi (3,89; 6,59) ao passo que o intervalo assintótico foi (3,76; 6,46). Essa diferença nos limites é uma translação uniforme do intervalo em 0,13, mantendo a amplitude de ambos em 2,70, o que representa uma razão de 0,13/2,70=0,048. Essa diferença é pequena e, portanto, consideramos que a aproximação quadrática da verossimilhança foi satisfatória. Embora a precisão dos intervalos tenha sido a mesma, o baseado na log-verossimilhança é mais adequado no sentido de que a assimetria melhor representa a incerteza associada ao parâmetro α uma vez que este têm espaço paramétrico limitado à esquerda $(0; \infty)$.

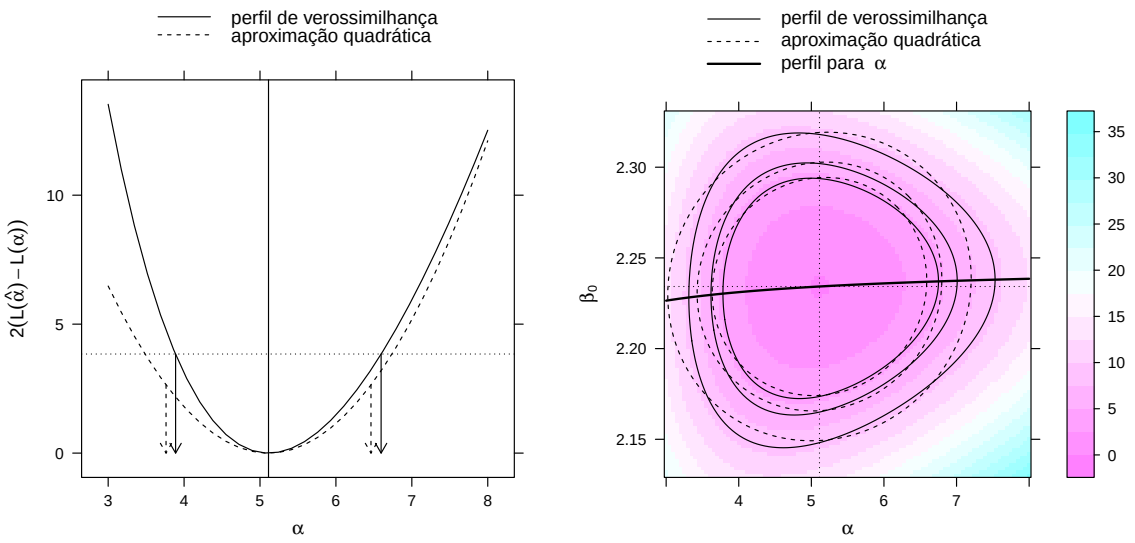

Figura 5: (esq) Perfil de verossimilhança exato e por aproximação quadrática para o parâmetro α. Setas indicam o intervalo de 95% de confiança. (dir) Regiões de confiança (90, 95, 99%) para os parâmetros β_0 e α por perfil de verossimilhança e aproximação quadrática.

Outro fato interessante é ver se este parâmetro tem alguma influência sobre os parâmetros da regressão (ortogonalidade). No gráfico da direita na Figura 5 temos a região de confiança conjunta para os parâmetros α e β_0 também obtida via perfil de verossimilhança e aproximação quadrática da verossimilhança. A região de confiança apresenta eixos paralelos aos eixos cartesianos o que aponta ortogonalidade entre os parâmetros na informação observada. Em decorrência disso, a trajetória para o perfil de log-verossimilhança do parâmetro α foi praticamente paralela ao seu eixo. Essa é uma propriedade interessante pois as inferências sobre um parâmetro passam a não ter forte influência dos valores dos outros.

Observamos simetria da região de confiança para o parâmetro β_0 e sobreposição entre os limites com relação às regiões de confiança assintóticas para o eixo vertical em $\hat{\alpha}$. Para esse parâmetro, o intervalo de confiança assintótico coincide com o de log-verossimilhança. Do ponto de vista computacional, o intervalo de confiança assintótico é mais fácil de obter pois envolve a inversão da matriz hessiana da função de log-verossimilhança enquanto que o perfil de log-verossimilhança precisa da otimização sucessiva para um conjunto de valores do parâmetro de interesse. Considerando que o número de parâmetros no modelo de regressão é

elevado, intervalos individuais baseados em perfil de verossimilhança são computacionalmente caros.

5 Conclusão

Neste artigo apresentamos uma analise de dados apresentando subdispersão. Como abordagens de analise exploramos os modelos de regressão de Poisson, contagem Gama e o modelo semi-paramétrico Quasi-Poisson. A comparação foi feita através da análise de um conjunto de dados, referente a um experimento em casa de vegetação com plantas de algodão submetidas à diferentes níveis de desfolha artificial e estágio fenológico.

Os modelos contagem gama e Quasi-Poisson levaram as mesmas conclusões quanto a efeito dos fatores experimentais. O modelo Poisson mostrou-se mais conservador não sendo capaz de identificar alguns fatores experimentais como significativos, para o conjunto de dados analisado. Este modelo também apresentou erros padrões maiores que o do modelo contagem gama, além de bandas de predição mais largas, mostrando seu menor poder em resumir a informação contida nos dados através do modelo ajustado. O conjunto de dados analisado indica que na presença de subdispersão o modelo de Poisson é inadequado e pode levar a conclusões erradas sobre efeitos de fatores experimentais, ou covariáveis de interesse em modelos de regressão.

O modelo contagem gama apresentou resultados bastante satisfatórios quando comparado a abordagem semi-paramétrica mais flexível, porém com incoveniente dado sua estrutura semi-paramétrica, principalmente com relação a obtenção da distribuição de probabilidade para as contagens que é impossível de ser obtida. Além disso, generalizações como modelar a dispersão é bastante direta pelo modelo contagem Gama, outra vantagem evidente desta abordagem.

Dado o pouco uso do modelo contagem gama para a análise de dados de contagens em agronômia, foram investigados alguns aspecto da inferência nesta classe de modelos pelo

método da máxima verossimilhança. Foi verificado que a aproximação quadrática para o estimador de α parâmetro relevante do modelo contagem gama é bastante satisfatória, além disso, também verificamos que este parâmetro tem pouca influência nas estimativas pontuais dos parâmetros do modelo de regressão, ficando este parâmetro responsável por estabilizar as estimativas de variâncias dos parâmetros de regressão, que são em geral super estimadas quando usa-se a distribuição de Poisson, que não considera a subdispersão.

Deixamos como futuras agendas de pesquisa na análise de dados subdispersas comparações via simulação para verificar precisamente o efeito da má especificação da distribuição da variável resposta na presença de dados subdispersos, tanto na escolha entre modelos como na condução de testes de hipóteses aspectos relevantes sobre o fenômeno em estudo. Outro ponto para pesquisa é a inclusão de covariáveis para explicar a subdispersão, ou seja, modelar também o parâmetro α como função de covariáveis. E por fim, a inclusão de efeitos aleatórios pode também ser de grande valia em modelos para contagens subdispersas, em diversos experimentos onde várias medidas são feitas nas mesmas unidades experimentais, a variabilidade entre as unidades inflaciona a variabilidade total, porém quando controlada pela inclusão de efeitos aleatórios a variância restante pode não ser bem descrita pelo modelo Poisson, decorrente da sub ou superdispersão, nestas situações o modelo contagem Gama pode ser uma opção interessante.

Referências

Richard H. Byrd. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 35(5):773, 1995.

Anderson Miguel da Silva, Paulo Eduardo Degrande, Marcos Gino Fernandes, Renato Suekane, and Walmes Marques Zeviani. Impacto de diferentes níveis de desfolha artificial nos estádios fenológicos do algodoeiro. *Revista de Ciências Agrárias*, 35(1):163–172, 2012.

Abdel H El Shaarawi, Rong Zhu, and Harry Joe. Modelling species abundance using the Poisson-Tweedie family. *Environmetrics*, 22(2):152–164, March 2011.

Ursula Gonzales-Barron and Francis Butler. Characterisation of within-batch and between-batch variability in microbial counts in foods using Poisson-gamma and Poisson-lognormal regression models. *Food Control*, 22(8):1268–1278, August 2011.

Gary K Grunwald, Stephanie L Bruce, Luohua Jiang, Matthew Strand, and Nathan Rabinovitch. A statistical model for under or overdispersed clustered and longitudinal count data. *Biometrical Journal*, 53(4):578–594, July 2011.

Gary King. Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator. *American Journal of Political Science*, 33(3):762, August 1989.

Dominique Lord, Seth D Guikema, and Srinivas Reddy Geedipally. Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. *Accident; analysis and prevention*, 40(3):1123–34, May 2008.

Dominique Lord, Srinivas Reddy Geedipally, and Seth D Guikema. Extension of the application of conway-maxwell-poisson models: analyzing traffic crash data exhibiting underdispersion. *Risk analysis : an official publication of the Society for Risk Analysis*, 30(8): 1268–76, August 2010.

Blake McShane, Moshe Adrian, Eric T Bradlow, and Peter S Fader. Count Models Based on Weibull Interarrival Times. *Journal of Business and Economic Statistics*, 26(3):369–378, July 2008.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A*, 135(3):370–384, 1972.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

M. S Ridout and P. Besbeas. An empirical model for underdispersed count data. *Statistical Modelling*, 4(1):77–89, April 2004.

Nils Toft, Giles T. Innocent, Dominic J. Mellor, and Stuart W.J. Reid. The Gamma-Poisson model as a statistical method to determine if micro-organisms are randomly distributed in a food matrix. *Food Microbiology*, 23(1):90–94, February 2006.

Rainer Winkelmann. Duration Dependence and Dispersion in Count-Data Models. *Journal of Business & Economic Statistics*, 13(4):467–474, October 1995.

Rainer Winkelmann and Klaus Zimmermann. Count data models for demographic data. *Mathematical Population Studies*, 4(3):205–221, February 1994.

Rong Zhu and Harry Joe. Modelling heavy-tailed count data using a generalised Poisson-inverse Gaussian family. *Statistics & Probability Letters*, 79(15):1695–1703, August 2009.