Universidade Federal do Paraná



Extensões e Aplicações Modelo de Regressão Conway-Maxwell-Poisson para Modelagem de Dados de Contagem

Curitiba

Eduardo Elias Ribeiro Junior

Extensões e Aplicações Modelo de Regressão Conway-Maxwell-Poisson para Modelagem de Dados de Contagem

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório B do Curso de Graduação em Estatística da Universidade Federal do Paraná, como exigência parcial para obtenção do grau de Bacharel em Estatística.

Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística

Orientador: Prof. Dr. Walmes Marques Zeviani

Curitiba 2016

"Software is like sex: it's better when it's free"

— Linus Torvalds

"The numbers are where the scientific discussion should start, not end."
— Steven N. Goodman

Resumo

Variáveis aleatórias de contagem são de natureza discreta e representam o número de ocorrências de um evento em um domínio discreto ou contínuo. Para análise estatística dessas variáveis, o modelo de Poisson é amplamente utilizado. Porém, não são raras as situações de sub ou superdispersão, que inviabilizam o emprego deste modelo. Uma alternativa paramétrica é o modelo COM-Poisson que, com a adição de um parâmetro, contempla diferentes níveis de dispersão. Outras características bastantes frequentes em dados de contagem são frequência excessiva de valores zeros e estrutura de correlação entre observações, muitas vezes induzida pelo processo de casualização ou amostragem. Nesses casos os modelos adotados devem ser adaptados. Neste trabalho são exploradas as características da distribuição COM-Poisson e apresentados os modelos de regressão COM-Poisson de efeitos fixos, com modelagem para excesso de zeros e incluindo efeitos aleatórios. O emprego dos modelos COM-Poisson e suas extensões é ilustrado com aplicações onde seus resultados são comparados com as abordagens Poisson, Quasi-Poisson e Binomial Negativa (para casos de superdispersão) via níveis descritivos de testes de razão de verossimilhanças, critério de informação de Akaike e predições pontuais e intervalares. O ajuste dos modelos é feito via maximização da verossimilhança. Os resultados mostram que o modelo Poisson é de fato restritivo, com ajustes inadequados na maioria das aplicações. O modelo COM-Poisson, por sua vez, mostrou-se bastante flexível com resultados similares aos obtidos via abordagem semi-paramétrica Quasi-Poisson. As extensões propostas para o modelo COM-Poisson apresentaram resultados satisfatórios, sendo equivalentes as abordagens já consolidadas na literatura.

Palavras-chave: COM-Poisson; dados de contagem; subdispersão; superdispersão; excesso de zeros; efeitos aleatórios; Binomial Negativa; Quasi-Poisson

Lista de ilustrações

Figura 1 –	Ilustração de diferentes tipos de processos pontuais	17
Figura 2 –	Probabilidades pela distribuição Poisson para diferentes parâmetros	21
Figura 3 –	Probabilidades pela distribuição Binomial Negativa para diferentes	
	níveis de dispersão, fixando a média em 5	24
Figura 4 –	Relação Média e Variância na distribuição Binomial Negativa	24
Figura 5 –	Probabilidades pela distribuição COM-Poisson para diferentes parâ-	
	metros	26
Figura 6 –	Exemplos de casos particulares da distribuição COM-Poisson	27
Figura 7 –	Relação Média e Variância na distribuição COM-Poisson	28
Figura 8 –	Convergência da constante de normalização da COM-Poisson para	
	diferentes conjuntos de parâmetros	29
Figura 9 –	Ilustração de dados de contagem com excesso de zeros	30
Figura 10 –	Número de capulhos produzidos para cada nível de desfolha e estágio	
	fenológico (esquerda) e médias e variâncias das cinco repetições em	
	cada combinação de nível de desfolha e estágio fenológico (direita) .	36
Figura 11 –	Disposição das variáveis de contagem nº de estruturas reprodutivas,	
	nº de capulhos produzidos e nº de nós da planta observadas sob	
	diferentes dias de exposição à infestação de Mosca-branca	37
Figura 12 –	Disposição das variáveis número de grãos e número de vagens nos	
	diferentes níveis de adubação potássica e umidade do solo	38
Figura 13 –	Médias e variâncias amostrais das contagens de grão e vagens, avali-	
	adas no experimento com soja sob efeito umidade e adubação potássica	39
Figura 14 –	Dispersão entre o número total de ninfas de Mosca-branca nos folíolos	
	da soja e o número de dias após a primeira avaliação para as quatro	
	diferentes cultivares (esquerda)	40
Figura 15 –	Logarítmo neperiano do número de peixes capturados acrescido de	
	0,5 para as diferentes composições dos grupos (esquerda)	40
Figura 16 –	Dispersão do número de nematoides providos por uma alíquota da	
	solução de 1 g/ml de massa fresca diluída	41
Figura 17 –	Perfil de log-verossimilhança para o parâmetro extra da COM-Poisson,	
	estimado no modelo com o quinto preditor	49
Figura 18 –	Imagem da matriz de correlação entre os parâmetros do modelo	
	COM-Poisson	50
Figura 19 –	Curva dos valores preditos com intervalo de confiança de (95%) como	
	função do nível de desfolha e do estágio fenológico da planta	51

Figura 20 –	Perfis de log-verossimilhança para o parâmetro extra da COM-Poisson nos modelos para número de capulhos produzidos (esquerda), número de estruturas reprodutivas (central) e número de nós (direira).	53
Figura 21 –	Imagem da matriz de correlação entre os parâmetros do modelo COM-Poisson	54
Figura 22 –	Curva dos valores preditos com intervalo de confiança de (95%) como função dos dias de exposição a alta infestação de Mosca-branca considerando os modelos para o número de estruturas reprodutivas (esquerda), número de capulhos produzidos (centro) e número de nós (direita)	54
Figura 23 –	Convergência das constantes de normalização para cada indivíduo no modelo para o número de vagens viáveis (esquerda) e para o número de grãos produzidos (direita)	56
Figura 24 –	Perfis de log-verossimilhança para o parâmetro de precisão da COM- Poisson nos modelos para número de vagens viáveis por parcela (esquerda) e número grãos de soja por parcela (direira)	57
Figura 25 –	Imagem da matriz de correlação entre os parâmetros do modelo COM-Poisson ajustados ao número de vagens por parcela	58
Figura 26 –	Imagem da matriz de correlação entre os parâmetros do modelo COM-Poisson ajustados ao número de grãos por parcela	59
Figura 27 –	Valores preditos com intervalos de confiança (95%) como função do nível de adubação com potássio e do percentual de umidade do solo para cada variável de interesse mensurada (número de vagens e número de grãos por parcela)	60
Figura 28 –	Convergência das constantes de normalização para cada indivíduo (direita) e perfil de log-verossimilhança para o parâmetro extra da COM-Poisson (esquerda) no modelo para o número de ninfas de	
Figura 29 –	Mosca-branca	61
Figura 30 –	Valores preditos com intervalos de confiança (95%) em função das	63
Figura 31 –	Cultivares de soja e da data de avaliação da planta	64
Figura 32 –	número de crianças e pessoas no grupo e a presença de um campista Perfis de verossimilhança dos parâmetros estimados no modelo COM-	67
Figura 33 –	Poisson Misto	69 71

Figura 34 – Perfis de verossimilhança dos parâmetros estimados no modelo COM-	
Poisson Misto	71

Lista de tabelas

Tabela 1 – Distribuições de probabilidades para dados de contagem com indica-	
ção das características contempladas	20
Tabela 2 – Médias e variâncias amostras das contagens avaliadas no experimento	
de capulhos de algodão sob efeito de Mosca-Branca	37
Tabela 3 – Medidas de ajuste para avaliação e comparação entre preditores e	
modelos ajustados	48
Tabela 4 – Estimativas dos parâmetros e razões entre as estimativa e erro padrão	
para os três modelos em estudo	49
Tabela 5 – Medidas de ajuste para avaliação e comparação entre preditores e	
modelos ajustados	52
Tabela 6 – Medidas de ajuste para avaliação e comparação entre preditores e	
modelos ajustados ao número de vagens e ao número de grão por	
parcela	57
Tabela 7 – Medidas de ajuste para avaliação e comparação entre preditores e	
modelos ajustados	62
Tabela 8 – Medidas de ajuste para avaliação e comparação de preditores e mo-	
delos com componente de barreira ajustados	65
Tabela 9 – Estimativas dos parâmetros e razões entre as estimativa e erro padrão	
para os três modelos em estudo	66
Tabela 10 – Medidas de ajuste para avaliação e comparação entre preditores e	
modelos ajustados	69
Tabela 11 – Estimativas dos parâmetros e razões entre as estimativa e erro padrão	
para os três modelos em estudo	70

Sumário

1	INTRODUÇÃO	15
2	MODELOS PARA DADOS DE CONTAGEM	19
2.1	Modelo Poisson	20
2.1.1	Estimação via Quase-Verossimilhança	22
2.2	Modelo Binomial Negativo	23
2.3	Modelo COM-Poisson	25
2.4	Modelos para excesso de zeros	29
2.5	Modelos de efeitos aleatórios	31
3	MATERIAL E MÉTODOS	35
3.1	Materias	35
3.1.1	Conjuntos de dados	35
3.1.1.1	Capulhos de algodão sob efeito de desfolha artificial	35
3.1.1.2	Produtividade de algodão sob efeito de infestação de Mosca-branca	36
3.1.1.3	Produtividade de soja sob efeito de umidade do solo e adubação potássica	38
3.1.1.4	Ocorrência de ninfas de Mosca-branca em lavoura de soja	39
3.1.1.5	Peixes Capturados por Pescadores em um Parque Estadual	39
3.1.1.6	Número de nematoides em raizes de feijoeiro	41
3.1.2	Recursos computacionais	42
3.2	Métodos	42
4	RESULTADOS E DISCUSSÃO	47
4.1	Análise de dados de capulhos de algodão sob efeito de desfolha	47
4.2	Análise de dados de capulhos de algodão sob efeito de Mosca-Branca	52
4.3	Análise de produção de soja sob efeito de umidade e adubação potássica	55
4.4	Análise de ninfas de mosca-branca em lavoura de soja	60
4.5	Análise de captura de peixes em um parque estadual	64
4.6	Análise de dados de reprodução de nematoides em cultivares de feijoeiro	68
4.7	Discussões	72
5	CONSIDERAÇÕES FINAIS	75

REFERÊNCIAS				 	 	 		 77
APÊNDICES								81
APÊNDICE A -	PROG	RAMA	S R	 		 		 83

1 Introdução

Em diversas áreas do conhecimento é comum o interesse em i) compreender o relacionamento entre variáveis de interesse e características de uma amostra e ii) realizar predições por meio de modelos estatísticos ajustados por dados de uma amostra. A teoria de modelos de regressão sustentam muitas das pesquisas na área de Estatística aplicada.

Os modelos de regressão, na sua forma univariada e usual, consistem no estabelecimento de uma equação matemática que relaciona a média de uma variável aleatória de interesse (variável resposta) com as demais variáveis observadas (covariáveis). Nesta metodologia considera-se uma distribuição de probabilidades para a variável resposta condicionada as covariáveis cuja média está associada a uma preditor que acomoda os efeitos das covariáveis.

Pode-se destacar o modelo linear normal como o de uso predominante dentre os disponíveis para análises estatísticas aplicadas. Esse modelo estabelece que a variável resposta condicional as covariáveis tem distribuição Normal de média descrita por um preditor linear das covariáveis. Todavia, não são raras as situações em que a variável resposta é uma contagem, assumindo valores inteiros não negativos. Variáveis aleatórias de contagem, de forma geral, representam o número de ocorrências de um evento em um domínio específico que pode ser contínuo, como um intervalo de tempo ou espaço, ou discreto, como indivíduos ou grupos.

A análise de dados de contagem pelo modelo linear normal produz estimativas que contêm erros padrões inconsistentes e podem produzir predições negativas para o número de eventos (KING, 1989). Uma alternativa adotada durante muitos anos, e ainda aplicada, é encontrar alguma forma de transformação da variável resposta a fim de atender aos pressupostos do modelo de regressão normal. Contudo essa abordagem dispõe de resultados insatisfatórios, pois i) dificulta a interpretação dos resultados, ii) não contempla a natureza da variável (ainda serão um conjunto discreto de valores, só que em outra escala) iii) não contempla a relação média e variância, característica de dados de contagem e iv) o uso da transformação logarítmica é problemática quando há contagens nulas.

Diante do problema diferentes abordagens foram propostas, contudo destaca-se o trabalho apresentado por Nelder e Wedderburn (1972) que introduz a teoria dos modelos lineares generalizados (MLG's). Esta nova classe de modelos flexibilizou a distribuição condicional permitindo que outras distribuições pertencentes à família exponencial fossem consideradas para a distribuição da variável resposta. Tal família

contempla as distribuições Poisson, Binomial, Gama entre outras bem conhecidas na literatura, além da própria distribuição Normal.

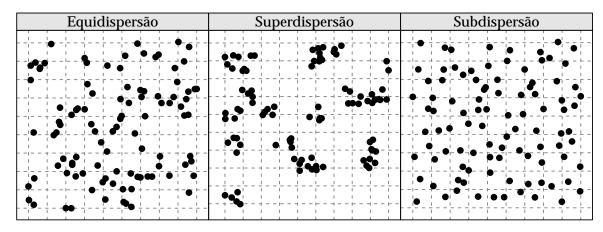
Com os MLG's a modelagem de dados passou a ser mais fiel a natureza da variável resposta, principalmente no que diz respeito ao seu suporte. Nesse contexto, a análise de variáveis aleatórias de contagem, que têm suporte nos conjunto dos números naturais, foi enriquecida expressivamente.

Para análise estatística dessas variáveis, o modelo probabilístico de Poisson, já consolidado na literatura é amplamente utilizado. Este modelo possui apenas um parâmetro, denotado por λ , que representa a média e também a variância, o que implica em uma relação identidade ($\lambda = E(Y) = V(Y)$). Essa propriedade, chamada de equidispersão, é uma particularidade do modelo Poisson que pode não ser adequada a diversas situações. Quando aplicado sob negligência desta suposição, o modelo Poisson apresenta erros padrões inconsistentes para as estimativas dos parâmetros e por consequência, para toda função desses parâmetros (WINKELMANN, 1995; WINKELMANN, 1994).

O caso de superdispersão, quando a variância é maior que a média, é o mais comum e existe uma variedade de métodos para análise de dados assim. A superdispersão pode ocorrer pela ausência de covariáveis importantes, excesso de zeros, diferentes amplitudes de domínio (offset) não consideradas, heterogeneidade de unidades amostrais, entre outros (RIBEIRO JR et al., 2012). Para tais casos, uma abordagem é a adoção de modelos com efeitos aleatórios que capturam a variabilidade extra com a adoção de um ou mais termos de efeito aleatório. Um caso particular do modelo Poisson de efeitos aleatórios, muito adotado no campo aplicado da Estatística, ocorre quando considera-se a distribuição Gama para os efeitos aleatórios, nessa situação há expressão fechada para a função de probabilidade marginal, que assume a forma Binomial Negativa.

Outra manifestação de fuga da suposição de equidispersão é a subdispersão, situação menos comum na literatura. Os processos que reduzem a variabilidade das contagens, abaixo do estabelecido pela Poisson, não são tão conhecidos quanto os que produzem variabilidade extra. Pela mesma razão, são poucas as abordagens descritas na literatura capazes de tratar subdispersão, uma vez que efeitos aleatórios só capturam a variabilidade extra. Cita-se os modelos de quasi-verossimilhança como a abordagem mais utilizada. Todavia não é possível descrever uma distribuição de probabilidades para a variável resposta nessa abordagem, pois a modelagem é baseada apenas nos dois primeiros momentos da distribuição condicional (PAULA, 2013).

A figura 1 ilustra, em duas dimensões, a ocorrência de equi, super e subdispersão respectivamente. Nesta figura cada ponto representa a ocorrência de um evento e cada parcela, delimitada pelas linhas pontilhadas, representa a unidade (ou domínio) na qual tem-se o número de eventos (como variável aleatória). O painel da esquerda representa



Fonte: Elaborado pelo autor.

Figura 1 – Ilustração de diferentes tipos de processos pontuais. Da direita para esquerda têm-se processos sob padrões aleatório, aglomerado e uniforme.

a situação de dados de contagem equidispersos, nesse cenário as ocorrências da variável aleatória se dispõem aleatoriamente. No painel central o padrão já se altera, tem-se a representação do caso de superdispersão. Nesse cenário formam-se aglomerados que deixam parcelas com contagens muito elevadas e parcelas com contagens baixas. Uma possível causa deste padrão se dá pelo processo de contágio (e.g. contagem de casos de uma doença contagiosa, contagem de frutos apodrecidos). No terceiro e último painel ilustra-se o caso de subdispersão, em que as ocorrências se dispõem uniformemente no espaço. Agora as contagens de ocorrências nas parcelas variam bem pouco. Ao contrário do caso superdisperso uma causa provável seria o oposto de contágio, a repulsa, ou seja, uma ocorrência causa a repulsa de outras ocorrências em seu redor (e.g. contagem de árvores, contagem de animais territoriais ou que disputam por território).

Uma alterativa paramétrica que contempla os casos de equi, super e subdispersão é a adoção de uma distribuição mais flexível para a variável resposta condicional as covariáveis. Conway e Maxwell (1962), antes da formalização dos MLG's, propuseram uma distribuição denominada COM-Poisson (nome em em homenagem aos seus autores Richard W. Conway, William L. Maxwell, Conway-Maxwell-Poisson) que generaliza a Poisson com a adição de mais um parâmetro, denotado por ν , que torna a razão de probabilidades sucessivas não linear, contemplando os casos de sub e superdispersão (SHMUELI et al., 2005).

Uma característica bastante relevante é que a COM-Poisson possui como casos particulares as distribuições Poisson, Geométrica e Binomial. Portanto, empregando a COM-Poisson como distribuição condicional de um modelo de regressão, a imposição de equidispersão não precisa ser satisfeita. Tal flexibilidade, considerando o amplo uso do modelo Poisson, significa que a COM-Poisson pode ser aplicada nessas situações e

será especialmente importante naquelas onde há fuga da equidispersão.

Assim como no modelo COM-Poisson vários aspectos do COM-Poisson podem ser estendidos. Por exemplo, há situações em que o delineamento do experimento sugere uma estrutura de covariância entre observações induzidas por um processo hierárquico de casualização ou amostragem. São casos assim os experimentos em parcelas subdivididas e experimentos com medidas repetidas ou longitudinais. Tais estruturas estabelecem modelos com efeitos não observáveis que agem em grupos experimentais e isso pode ser incorporado no modelo de regressão COM-Poisson com a inclusão de efeitos aleatórios. Da mesma forma, excesso de zeros pode ser introduzido a essa distribuição da mesma maneira que ocorre para o modelo Poisson, através de truncamento (modelos Hurdle) ou inflação (modelos de mistura) (SELLERS; RAIM, 2016). Estas extensões do modelo COM-Poisson ainda não são bem consolidadas na literatura e são escassas suas aplicações. Uma constatação do fato é que não há implementações destas extensões nos principais softwares estatísticos.

Na literatura brasileira, aplicações do modelo COM-Poisson são escassas. Foram encontradas apenas aplicações na área de Análise de Sobrevivência, mais especificamente em modelos com fração de cura (RIBEIRO, 2012; BORGES, 2012). Portanto, o presente trabalho visa colaborar com a literatura estatística brasileira i) apresentando e explorando o modelo de regressão COM-Poisson para dados de contagem, ii) estendendo as aplicações desse modelo para situações específicas como inclusão de efeitos aleatórios e modelagem de excesso de zeros, iii) discutindo os aspectos inferenciais por meio de análise de dados reais e iv) disponibilizando os recursos computacionais, em formato de pacote R, para ajuste dos modelos apresentados. Nas aplicações optou-se também pela análise via modelos já disponíveis para as situações estudadas.

O trabalho é organizado em cinco capítulos. Esse primeiro capítulo visa enfatizar as características das variáveis aleatórias de contagem e suas lacunas que podem ser complementadas na análise estatística dessas variáveis. O capítulo 2 é dedicado a revisão bibliográfica dos modelos estatísticos empregados a análise de dados de contagem. Nesse capítulo os modelos Poisson, Binomial Negativo, COM-Poisson, as abordagens para excesso de zeros e a estrutura dos modelos de efeitos aleatórios são apresentados. No capítulo 3 são apresentos os conjuntos de dados a serem analisados e os métodos para ajuste e comparação dos modelos. O capítulo 4 traz os os principais resultados da aplicação e comparação dos modelos estatísticos com ênfase nas discussões sob aspectos inferenciais empíricos. Finalmente no capítulo 5 são apresentadas as considerações finais obtidas desse trabalho e listados algumas possíveis linhas de pesquisa para estudos futuros.

2 Modelos para dados de contagem

Métodos para inferência em dados de contagem estão bem aquém da quantidade disponível para dados contínuos. Destaca-se o modelo log-linear Poisson como o modelo mais utilizado quando se trata de dados de contagem. Porém, não raramente os dados de contagens apresentam variância superior ou inferior à sua média. Esses são os casos de super ou subdispersão já enunciados no capítulo 1, que quando ocorrem inviabilizam o uso da distribuição Poisson.

Nos casos de fuga da equidispersão algumas abordagens não paramétricas são empregadas. Nesse contexto, são alternativas os métodos de estimação via quase-verossimilhança, estimação robusta dos erros padrões (estimador "sanduíche") e estimação dos erros padrões via reamostragem ("bootstrap") (HILBE, 2014). Desses métodos detalha-se, brevemente, somente o método de estimação via função de quase-verossimilhança na seção 2.1.1.

No contexto paramétrico, pesquisas recentes trazem modelos bastante flexíveis à fuga de equidispersão no campo da Estatística aplicada, veja Sellers e Shmueli (2010), Zeviani et al. (2014), Lord, Geedipally e Guikema (2010). Na tabela 1 são listadas as distribuições de probabilidades consideradas por Winkelmann (2008) e Kokonendji (2014) e as características de dados de contagem que são contempladas. Nota-se que a Poisson na verdade é um caso particular, pois é a única das distribuições listadas que contempla somente a característica de equidispersão, ainda observa-se um conjunto maior de distribuições para os casos de superdispersão com relação os casos de subdispersão. Embora este grande número de distribuições exista para lidar com os casos de fuga de equidispersão, são poucos os pacotes estatísticos que as disponibilizam como alternativas para ajuste de modelos de regressão para dados de contagem.

Dos modelos paramétricos, o Binomial Negativo aparece em destaque com implementações já consolidadas nos principais *softwares* estatísticos e frequentes aplicações nos casos de superdispersão. Na seção 2.2 detalhes da construção desses modelos são apresentados. Dos demais modelos derivados das distribuições listadas na tabela 1 este trabalho abordará somente o modelo COM-Poisson, que é apresentado com detalhes na seção 2.3.

Um outro fenômeno que é frequente em dados de contagem é a ocorrência excessiva de zeros. Esse fenômeno sugere a modelagem de dois processos geradores de dados, o gerador de zeros extra e o gerador das contagens. Existem ao menos duas abordagens pertinentes para estes casos que são os modelos de mistura e os modelos condicionais. Na abordagem por modelos de mistura a variável resposta é modelada

Distribuição	Contempla a característica de			
	Equidispersão	Superdispersão	Subdispersão	
Poisson	√			
Binomial Negativa	\checkmark	\checkmark		
Inverse Gaussian Poisson	\checkmark	\checkmark		
Compound Poisson	\checkmark	\checkmark		
Poisson Generalizada	\checkmark	\checkmark	\checkmark	
Gamma-Count	\checkmark	\checkmark	\checkmark	
COM-Poisson	\checkmark	\checkmark	\checkmark	
Katz	\checkmark	\checkmark	\checkmark	
Poisson Polynomial	\checkmark	\checkmark	\checkmark	
Double-Poisson	\checkmark	\checkmark	\checkmark	
Lagrangian Poisson	\checkmark	\checkmark	\checkmark	

Tabela 1 – Distribuições de probabilidades para dados de contagem com indicação das características contempladas

Fonte: Elaborado pelo autor.

como uma mistura de duas distribuições, no trabalho de Lambert (1992), uma mistura da distribuição Bernoulli com uma distribuição de Poisson ou Binomial Negativa. Considerando os modelos condicionais, também chamados de modelos de barreira (RIDOUT; DEMETRIO; HINDE, 1998), tem-se que a modelagem da variável resposta é realizada em duas etapas. A primeira refere-se ao processo gerador de contagens nulas e a segunda ao gerador de contagens não nulas. Nesse trabalho a modelagem de excesso de zeros se dará somente via modelos de barreira. A seção 2.4 é destinada a um breve detalhamento desta abordagem.

Neste capítulo também é abordada a situação da inclusão de efeitos aleatórios na seção 2.5. Em análise de dados de contagem a inclusão desses efeitos permitem acomodar variabilidade extra e incorporar a estrutura amostral do problema como em experimentos com medidas repetidas ou longitudinais e experimentos em parcelas subdivididas.

2.1 Modelo Poisson

A Poisson é uma das principais distribuição de probabilidades discretas. Com suporte nos inteiros não negativos, uma variável aleatória segue um modelo Poisson se sua função massa de probabilidade for

$$\Pr(Y = y \mid \lambda) = \frac{\lambda^{y} e^{-\lambda}}{y!} \qquad y = 0, 1, 2, \dots$$
 (2.1)

em que $\lambda > 0$ representa a taxa de ocorrência do evento. Uma particularidade já

2.1. Modelo Poisson 21

destacada desta distribuição é que $E(X) = V(X) = \lambda$. Isso torna a distribuição Poisson bastante restritiva. Na figura 2 são apresentadas as distribuições Poisson para diferentes parâmetros, note que devido a propriedade E(X) = V(X) contagens maiores também são mais dispersas.

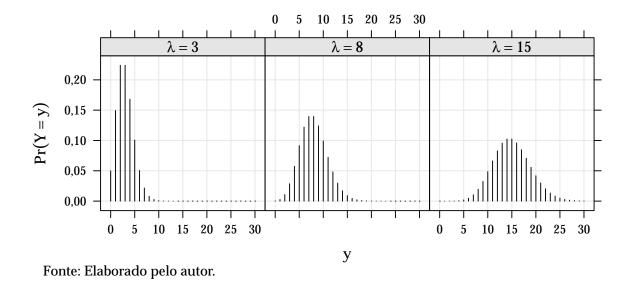


Figura 2 – Probabilidades pela distribuição Poisson para diferentes parâmetros.

Uma propriedade importante da distribuição Poisson é sua relação com a distribuição Exponencial. Essa relação estabelece que se os tempos entre a ocorrência de eventos se distribuem conforme modelo Exponencial de parâmetro λ a contagem de eventos em um intervalo de tempo t tem distribuição Poisson com média λt . A distribuição *Gamma-Count*, citada na tabela 1, estende esta propriedade do processo adotando a distribuição Gama para os tempos entre eventos tornando a distribuição da contagem decorrente mais flexível (WINKELMANN, 1995; ZEVIANI et al., 2014).

Outra propriedade que decorre da construção do modelo Poisson é sobre a razão entre probabilidades sucessivas, $\frac{\Pr(Y=y-1)}{\Pr(Y=y)} = \frac{y}{\lambda}$. Essa razão é linear em y e tem sua taxa de crescimento ou decrescimento como $\frac{1}{\lambda}$. Os modelos Katz e COM-Poisson se baseiam na generalização dessa razão de probabilidades a fim de flexibilizar a distribuição de probabilidades.

A utilização do modelo Poisson na análise de dados se dá por meio do modelo de regressão Poisson. Seja Y_i variáveis aleatórias condicionalmente independentes, dados as covariáveis X_i , $i=1,2,\ldots,n$. O modelo de regressão log-linear Poisson, sob a teoria dos MLG's é definido como

$$Y_i \mid X_i \sim \text{Poisson}(\mu_i)$$

 $\log(\mu_i) = X_i \beta$ (2.2)

em que $\mu_i > 0$ é a média da variável aleatória $Y_i \mid X_i$ que é calculada a partir do vetor $\beta \in \mathbb{R}^p$.

O processo de estimação do vetor β é baseado na maximização da verossimilhança, que nas distribuições pertencentes à família exponencial, os MLG's, é realizada via algoritmo de mínimos quadrados ponderados iterativamente, ou, do inglês *Iteractive Weighted Least Squares - IWLS* (NELDER; WEDDERBURN, 1972).

2.1.1 Estimação via Quase-Verossimilhança

Wedderburn (1974) propôs uma forma de estimação a partir de uma função biparamétrica, denominada quase-verossimilhança. Suponha y_i observações independentes com esperanças μ_i e variâncias $V(\mu_i)$, em que V é uma função positiva e conhecida. A função de quase-verossimilhança é expressa como

$$Q(\mu_i \mid y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2 V(\mu_i)} dt$$
 (2.3)

Na expressão 2.3 a função de quase-verossimilhança é definida a partir da especificação de μ_i , $V(\mu_i)$ e σ^2 . O processo de estimação via maximização dessa função compartilha, do método baseado na maximazação da verossimilhança, as mesmas estimativas para μ_i , porém a dispersão de y_i , $V(y_i) = \theta V(\mu_i)$ é corrigida pelo parâmetro adicional σ^2 .

Assim os problemas com a fuga da suposição de equidispersão podem ser superados quando a estimação por máxima quase-verossimilhança é adotada. Porém um resultado dessa abordagem é que

$$-E\left(\frac{\partial^{2}Q(\mu\mid y)}{\partial\mu^{2}}\right) \leq -E\left(\frac{\partial^{2}\ell(\mu\mid y)}{\partial\mu^{2}}\right) \tag{2.4}$$

ou seja a informação a respeito de μ quando se conhece apenas σ^2 e $V(\mu)$, a relação entre média e variância, é menor do que a informação quando se conhece a distribuição da variável resposta, dada pela log-verossimilhança $\ell(\mu \mid y)$. Além disso ressalta-se que, de forma geral, não é possível descrever uma distribuição de probabilides para Y somente com as especificações de σ^2 e $V(\mu)$.

Em modelos de regressão, $g(\mu_i) = X\beta$ e $V(\mu_i)$ definem a função de quaseverossimilhança. Nessa abordagem são estimados os parâmetros β e σ^2 . A estimativa do vetor β pode ser obtidas pelo algoritmo *IWLS*. Usando as funções quase-escore e matriz de quase-informação chega-se ao mesmo algoritmo de estimação dado no caso Poisson, que não depende de σ^2 . O parâmetro σ^2 é estimado separadamente, pós estimação dos β 's. Um estimador usual é o baseado na estatística χ^2 de Pearson.

$$\hat{\sigma^2} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$
 (2.5)

2.2 Modelo Binomial Negativo

Uma das principais alternativas paramétricas para dados de contagem superdispersos é a distribuição Binomial Negativa. A função massa de probabilidade da distribuição Binomial Negativa pode ser deduzida de um processo hierárquico de efeitos aleatórios onde se assume que

$$Y \mid b \sim \text{Poisson}(b)$$

$$b \sim \text{Gama}(\mu, \theta)$$
(2.6)

A função massa de probabilidade decorrente da estrutura descrita em 2.7 é deduzida integrando os efeitos aleatórios. Considere $f(y \mid b)$ como a função massa de probabilidade da distribuição Poisson (vide expressão em 2.1) e $g(b \mid \mu, \phi)$ a função densidade da distribuição Gama ¹

$$\Pr(Y = y \mid \mu, \theta) = \int_{0}^{\infty} f(y \mid b)g(b \mid \mu, \theta)db$$

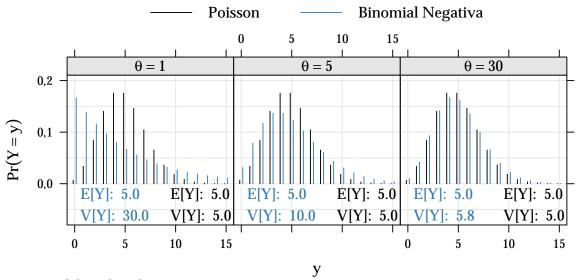
$$= \frac{\theta^{\theta}}{y!\mu^{\theta}\Gamma(\theta)} \int_{0}^{\infty} e^{-b(1+\theta/\mu)}b^{y+\theta-1}db$$

$$= \frac{\Gamma(\theta+y)}{\Gamma(y+1)\Gamma(\theta)} \left(\frac{\mu}{\mu+\theta}\right)^{y} \left(\frac{\theta}{\mu+\theta}\right)^{\theta} \qquad y = 0, 1, 2, \cdots$$
(2.7)

com $\mu>0$ e $\theta>0$. Esse é um caso particular de um modelo de efeito aleatório cuja integral tem solução analítica e por consequência o modelo marginal tem forma fechada. Outro caso que se baseia no mesmo princípio é o modelo *Inverse Gaussian Poisson*, que como o nome sugere adota a distribuição Inversa Gaussiana para os efeitos aleatórios. Na figura 3 são apresentadas as distribuições Binomial Negativa para diferentes parâmetros θ em comparação com a distribuição Poisson equivalente em locação. Note que quanto menor o parâmetro θ , maior a dispersão da distribuição. Isso introduz uma propriedade importante desse modelo, para $\theta\to\infty$ a distribuição reduz-se a Poisson.

Os momentos média e variância da distribuição Binomial Negativa são expressos como $E(Y) = \mu$ e $V(Y) = \mu + \mu^2/\sigma^2$. Pelas expressões fica evidente a característica

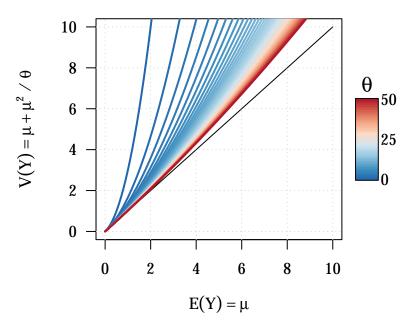
O desenvolvimento detalhado da integral pode ser visto em Paula (2013, pág. 303-305). Obs.: A função densidade do modelo Gama está parametrizada para que μ represente a média da distribuição.



Fonte: Elaborado pelo autor.

Figura 3 – Probabilidades pela distribuição Binomial Negativa para diferentes níveis de dispersão, fixando a média em 5.

da Binomial Negativa de acomodar somente superdispersão, pois E(Y) é menor que V(Y) para qualquer σ^2 . Percebe-se também que quanto maior o parâmetro σ^2 mais E(Y) se aproxima de V(Y), e no limite, quando $\sigma^2 \to \infty$, E(Y) = V(Y) fazendo com que a distribuição Binomial Negativa se reduza a Poisson.



Fonte: Elaborado pelo autor.

Figura 4 – Relação Média e Variância na distribuição Binomial Negativa.

2.3. Modelo COM-Poisson 25

A relação funcional entre média e variância é ilustrada na figura 4 onde são apresentadas as médias e variâncias para μ entre 0 e 10 e θ entre 0 e 50. O comportamento dessa relação proporciona um maior flexibilidade à distribuição em acomodar superdispersão, uma característica importante exibida nesta figura é que para a Binomial Negativa se aproximar a Poisson em contagens altas o θ deve ser extremamente grande.

O emprego do modelo Binomial Negativo em problemas se regressão ocorre de maneira similar aos MLG's, com exceção de que a distribuição só pertence a família exponencial de distribuições se o parâmetro θ for conhecido e assim o processo sofre algumas alterações. Primeiramente, assim como na Poisson, defini-se $g(\mu_i) = X\beta$, comumente utiliza-se a função $g(\mu_i) = \log(\mu_i)$. Desenvolvendo a log-verossimilhança e suas funções derivadas, função escore e matriz de informação de Fisher, mostrase que matriz de informação é bloco diagonal caracterizando a ortogonalidade dos parâmetros β de locação e θ de dispersão. Deste fato decorre que a estimação dos parâmetros pode ser realizada em paralelo, ou seja, estima-se o vetor β pelo método de IWLS e posteriormente o parâmetro θ pelo método de Newton-Raphson, faz-se os dois procedimentos simultaneamente até a convergência das estimativas.

2.3 Modelo COM-Poisson

A distribuição de probabilidades COM-Poisson foi proposta por Conway e Maxwell (1962), em um contexto de filas e generaliza a Poisson em termos da razão de probabilidades sucessivas, como será visto adiante. Seja Y uma variável aleatória COM-Poisson, então sua função massa de probabilidade é

$$Pr(Y = y \mid \lambda, \nu) = \frac{\lambda^y}{(y!)^{\nu} Z(\lambda, \nu)} \qquad y = 0, 1, 2, \dots$$
 (2.8)

em que $\lambda > 0$, $\nu \ge 0$ e $Z(\lambda, \nu)$ é uma constante de normalização, calculada para que de fato 2.8 seja uma função massa de probabilidade ($\sum_{i=1}^{\infty} \Pr(Y = y) = 1$). $Z(\lambda, \nu)$ é definida como se segue

$$Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^{\nu}}$$
 (2.9)

O fato que torna a distribuição COM-Poisson mais flexível é a razão entre probabilidades sucessivas

$$\frac{\Pr(Y=y-1)}{\Pr(Y=y)} = \frac{y^{\nu}}{\lambda} \tag{2.10}$$

que se caracteriza não necessariamente linear em y, diferentemente da Poisson, o que permite caudas mais pesadas ou mais leves à distribuição (SELLERS; SHMUELI, 2010). Na figura 5 são apresentadas as distribuições COM-Poisson para diferentes valores de λ e ν , em contraste com as equivalentes, em locação, distribuições Poisson. Nessa figura pode-se ver a flexibilidade desse modelo, pois i) contempla o caso de subdispersão mesmo em contagens baixas (E(Y)=3, painel a esquerda), a distribuição permite caudas pesadas e consequentemente uma dispersão extra Poisson, ii) contempla subdispersão mesmo em contagens altas, onde na Poisson tem-se variabilidade na mesma magnitude, na COM-Poisson pode-se ter caudas mais leves concentrando as probabilidades em torno da média (painel a direita) e iii) tem como caso particular a Poisson quando o parâmetro $\nu=1$ (painel central).

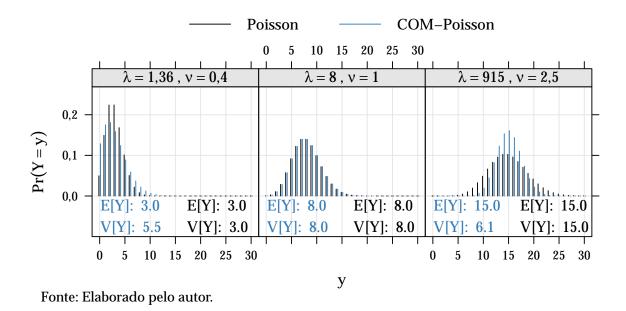


Figura 5 – Probabilidades pela distribuição COM-Poisson para diferentes parâmetros.

Uma das vantagens do modelo COM-Poisson é que possui, além da Poisson quando $\nu=1$, outras distribuições bem conhecidas como casos particulares. Esses casos particulares ocorrem essencialmente devido a forma assumida pela série infinita $Z(\lambda,\nu)$. Quando $\lambda=1$, $Z(\lambda,\nu=1)=e^{\lambda}$ e substituindo na expressão 2.8, tem-se a distribuição Poisson resultante. Quando $\nu\to\infty$, $Z(\lambda,\nu)\to 1+\lambda$ e a distribuição COM-Poisson se aproxima de uma distribuição Bernoulli com $P(Y=1)=\frac{\lambda}{1+\lambda}$. E quando $\nu=0$ e $\lambda<1$ $Z(\lambda,\nu)$ é uma soma geométrica que resulta em $(1-\lambda)^{-1}$ e a expressão 2.8 se resume a uma distribuição Geométrica com $P(Y=0)=(1-\lambda)$ (SHMUELI et al., 2005). Os três respectivos casos particulares citados são ilustrados na figura 6, onde os parâmetros foram escolhidos conforme restrições para redução da distribuição.

Um inconveniente desse modelo é que os momentos média e variância não tem forma fechada. Sendo assim, devem ser calculados a partir da definição

2.3. Modelo COM-Poisson 27

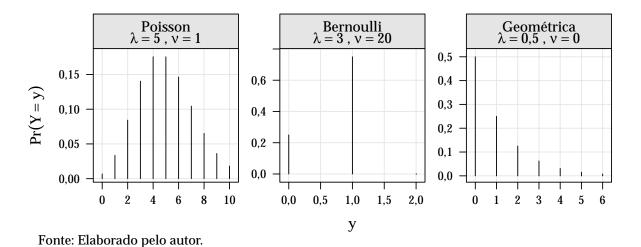


Figura 6 – Exemplos de casos particulares da distribuição COM-Poisson.

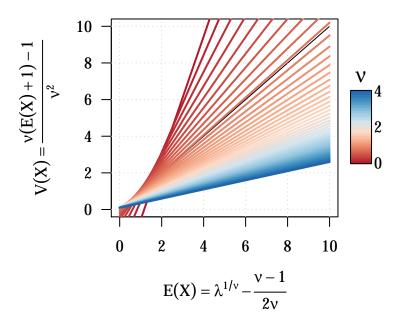
$$E(Y) = \sum_{y=0}^{\infty} y \cdot p(y)$$
 e $V(Y) = \sum_{y=0}^{\infty} y^2 \cdot p(y) - E^2(Y)$

Shmueli et al. (2005), a partir de uma aproximação para $Z(\lambda, \nu)$, apresenta uma forma aproximada para os momentos da distribuição

$$E(Y) \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu}$$
 e $V(Y) \approx \frac{\lambda^{1/\nu}}{\nu}$ (2.11)

os autores ressaltam que essa aproximação é satisfatória para $\nu \leq 1$ ou $\lambda > 10^{\nu}$. Na figura 7 é representada a relação média e variância aproximada pelas expressões em 2.11. Percebe-se que a relação é praticamente linear entre média e variância, Sellers e Shmueli (2010) descrevem que essa pode ser relação pode, ainda, ser aproximada por $\frac{1}{\nu}E(Y)$. Dessas aproximações, bem como das visualizações em 5, 6 e 7 deduz-se que o parâmetros ν , ou $\frac{1}{\nu}$, controla a precisão da distribuição, sendo ela equidispersa quando $\nu = 1$, superdispersa quando $\nu < 1$ e subdispersa quando $\nu > 1$.

Embora o modelo COM-Poisson não tenha expressão fechada para a média da distribuição pode-se utilizá-lo como modelo associado a distribuição condicional da variável resposta de contagem. Isso é feito incorporando um preditor linear em λ , que mesmo não representando a média, está associado com a locação da distribuição, ou seja, modela-se a média indiretamente nessa abordagem. O modelo de regressão é definido com as variáveis aleatórias condicionalmente independentes Y_1, Y_2, \ldots, Y_n , dado o vetor de covariáveis $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ seguindo um modelo COM-Poisson de parâmetros $\lambda_i = e^{X_i\beta}$, $i = 1, 2, \ldots, n$ e ν comum a todas as observações. Na expressão



Fonte: Elaborado pelo autor.

Figura 7 – Relação Média e Variância na distribuição COM-Poisson.

2.2 o modelo é devidamente formulado, conforme a notação de MLG's

$$Y_i \mid X_i \sim \text{COM-Poisson}(\lambda_i, \nu)$$

 $\eta(E(Y_i \mid X_i)) = \log(\lambda_i) = X_i \beta$ (2.12)

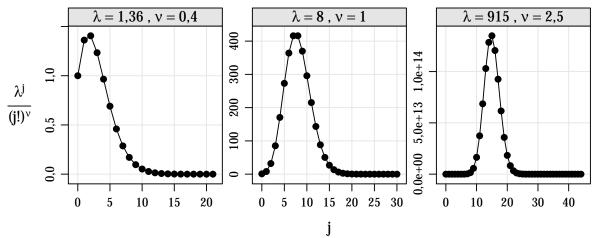
O algoritmo para estimação do conjunto de parâmetros $\Theta=(\nu,\beta)$ do modelo é baseado na maximização da log-verossimilhança, que decorrente da especificação em 2.12 é

$$\ell(\nu, \beta \mid \underline{y}) = \sum_{i}^{n} y_{i} \log(\lambda_{i}) - \nu \sum_{i}^{n} \log(y!) - \sum_{i}^{n} \log(Z(\lambda_{i}, \nu))$$
 (2.13)

e então as estimativas de máxima verossimilhança são

$$\hat{\Theta} = (\hat{v}, \hat{\beta}) = \underset{(v, \beta)}{\operatorname{arg max}} \ \ell(v, \beta \mid \underline{y})$$

Para avaliação da log-verossimilhança em 2.13 a constante de normalização $Z(\lambda, \nu)$, conforme definida em 2.9 é calcula para cada observação o que potencialmente torna o processo de estimação lento. Uma ilustração do número de incrementos considerados para cálculo da constante $Z(\lambda, \nu)$ é apresentada na figura 8. Nesta ilustração



Fonte: Elaborado pelo autor.

Figura 8 – Convergência da constante de normalização da COM-Poisson para diferentes conjuntos de parâmetros.

foram utilizados os mesmos parâmetros definidos em 5 e o número de incrementos considerados para convergência 2 . de $Z(\lambda, \nu)$ foram 22, 31, 45 nos primeiro, segundo e terceiro painéis respectivamente.

Detalhes computacionais do algoritmo de maximização e manipulações algébricas para eficiência na avaliação da log-verossimilhança no modelo COM-Poisson são discutidos na seção 3.2.

2.4 Modelos para excesso de zeros

Problemas com excesso de zeros são comuns em dados de contagem. Caracterizase como excesso de zeros casos em que a quantidade observada de contagens nulas supera substancialmente aquela esperada pelo modelo de contagem adotado. No caso do modelo Poisson a fração de zeros é $e^{-\lambda}$.

As contagens nulas que geram o excesso de zeros podem ser explicadas de duas formas distintas. A primeira denomina-se de zeros estruturais, quando a ocorrência de zero se dá pela ausência de determinada característica na população e a segunda, que zeros amostrais que ocorrem segundo um processo gerador de dados de contagem (e.g processo Poisson). Por exemplo, considerando o número de dias que uma família consome um determinado produto, tem-se aquelas famílias que não consomem o produto (zeros estruturais) e as demais famílias que consomem o produto, porém não o consumiram no intervalo de tempo considerado no estudo (zeros amostrais). Assim, de forma geral são dois processos geradores de dados em uma variável aleatória de

Adotou-se como critério de convergência a iteração j tal que $\lambda^j/(j!)^{\nu} < 0,00001$

contagem com excessivos zeros.

Em geral, quando dados de contagem apresentam excessos de valores zero também apresentarão superdispersão. Todavia, essa dispersão pode ser exclusivamente devido ao excesso de zeros e assim os modelos alternativos já apresentados não terão um bom desempenho. Uma ilustração deste fato é apresentada na figura 9, em que foram simulados dados com excesso de zeros e ajustado um modelo COM-Poisson. Em ambos os casos o modelo não se ajustou adequadamente, indicando que os excessos de zeros devem ser abordados de forma diferente.

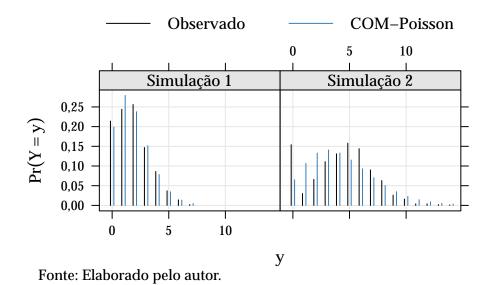


Figura 9 – Ilustração de dados de contagem com excesso de zeros.

Hilbe (2014, capítulo 7) discute sobre a interpretação e modelagem de dados de contagem com excesso de zeros. Para essa situação as duas principais abordagens são i) os modelos de mistura (LAMBERT, 1992), também chamados de inflacionados, em inglês *Zero Inflated Models* e ii) os modelos condicionais (RIDOUT; DEMETRIO; HINDE, 1998), também chamados de modelos de barreira, em inglês *Hurdle Models*. Neste trabalho somente a abordagem via modelos condicionais será considerada. A função massa de probabilidade do modelo Hurdle é

$$\Pr(Y = y \mid \pi, \Theta_c) = \begin{cases} \pi & \text{se } y = 0, \\ (1 - \pi) \frac{\Pr(Z = z \mid \Theta_c)}{1 - \Pr(Z = 0 \mid \Theta_c)} & \text{se } y = 1, 2, \dots \end{cases}$$
 (2.14)

em que $0 < \pi < 1$, representa a probabilidade de ocorrência de zeros e $\Pr(Z = z \mid \Theta_c)$ a função massa de probabilidade de uma variável aleatória de contagem Z, como a Poisson ou a Binomial Negativa.

Da especificação em 2.14, os momentos média e variância são obtidos usando as definições $E(Y) = \sum_{y=1}^{\infty} y \cdot \Pr(Y = y)$ e $V(Y) = \sum_{y=1}^{\infty} y^2 \cdot \Pr(Y = y) - E^2(Y)$

$$E(Y) = \frac{E(Z)(1-\pi)}{1-\Pr(Z=0)}$$
 e $V(Y) = \frac{1-\pi}{1-\Pr(Z=0)} \left[E(Z) \frac{(1-\pi)}{1-\Pr(Z=0)} \right]$

Para a inclusão de covariáveis, caracterizando um problema de regressão, dado que o modelo tem dois processos modela-se ambos como se segue

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = G_i \gamma \qquad e \qquad \begin{array}{c} Z_i \sim D(\mu_i, \phi) \\ g(\mu_i) = X_i \beta \end{array}$$
 (2.15)

com $i=1,2,\ldots,n$, G_i e X_i as covariáveis da i-ésima observação consideradas para explicação da contagens nulas e não nulas respectivamente, $D(\mu_i,\phi)$ uma distribuição de probabilidades considerada para as contagens não nulas que pode conter ou não um parâmetro ϕ adicional, se Poisson, $D(\mu_i,\phi)$ se resume a Poisson(μ_i) e $g(\mu_i)$ uma função de ligação, nos casos Poisson e Binomial Negativa considera-se $\log(\mu_i)$. O que está implícito na formulação 2.15 é que para a componente que explica a geração de zeros está sendo considerada a distribuição Bernoulli de parâmetro π_i , contudo podese utilizar distribuições censuradas a direita no ponto y=1 para estimação desta probabilidade, como explicam Zeileis, Kleiber e Jackman (2007).

2.5 Modelos de efeitos aleatórios

Nas seções anteriores os modelos que flexibilizam algumas suposições do modelo Poisson, basicamente permitindo casos não equidispersos e modelando conjuntamente um processo gerador de zeros extra foram explorados. Contudo, uma suposição dos modelos de regressão para dados de contagem vistos até aqui é que as variáveis aleatórias Y_1, Y_2, \ldots, Y_n são condicionalmente independentes, dado o vetor de covariáveis. Porém não são raras as situações em que essa suposição não se mostra adequada. Ribeiro (2012) cita alguns exemplos:

- as observações podem ser correlacionadas no espaço,
- as observações podem ser correlacionadas no tempo,
- interações complexas podem ser necessárias para modelar o efeito conjunto de algumas covariáveis,
- heterogeneidade entre indivíduos ou unidades podem não ser suficientemente descrita por covariáveis.

Nessas situações pode-se estender a classe de modelos de regressão com a adição de efeitos aleatórios que incorporam termos baseados em variáveis não observáveis (latentes) ao modelo, permitindo assim acomodar uma variabilidade, que pode ser ou não estruturada, não prescrita pelo modelo. De forma geral a especificação dos modelos de efeitos aleatórios segue uma especificação hierárquica

$$Y_{ij} \mid b_i, X_{ij} \sim D(\mu_{ij}, \phi)$$

$$g(\mu_{ij}) = X_{ij}\beta + Z_i b_i$$

$$b \sim K(\Theta_h)$$
(2.16)

para $i=1,2,\ldots,m$ (grupos com efeitos aleatórios comuns) e $j=1,2,\ldots,n$ (observações) com $D(\mu_{ij},\phi)$, uma distribuição considerada para as variáveis resposta condicionalmente independentes, $g(\mu_{ij})$ uma função de ligação conforme definida na teoria dos MLG's, X_{ij} e Z_i os vetores conhecidos que representam os efeitos das covariáveis de interesse e os termos que definem os grupos considerados como aleatórios, b_i uma quantidade aleatória provida de uma distribuição $K(\Theta_b)$. Nesses modelos uma quantidade aleatória é somada ao preditor linear, diferentemente dos modelos de efeitos fixos e a partir desta quantidade é possível induzir uma estrutura de dependência entre as observações.

Como são duas quantidades aleatórias no modelo, $Y \mid X$ e b, a verossimilhança para um modelo de efeito aleatório é dada integrando-se os efeitos aleatórios

$$\mathcal{L}(\beta, \phi, \Theta_b \mid \underline{y}) = \prod_{i=1}^m \int_{\mathbb{R}^q} \left(\prod_{j=1}^{n_i} f_D(y_{ij}, \mu, b_i) \right) \cdot f_K(b \mid \Theta_b) db_i$$
 (2.17)

Na avaliação da verossimilhança é necessário o cálculo de *m* integrais de dimensão *q*. Para muitos casos essa integral não tem forma analítica sendo necessários métodos numéricos de intergração, que são discutidos na seção 3.2. As estimativas de máxima verossimilhança são

$$\hat{\Theta} = (\hat{\beta}, \hat{\Theta_b}) = \underset{(\beta, \Theta_b)}{\operatorname{arg max}} \log(\mathcal{L}(\beta, \phi, \Theta_b \mid \underline{y}))$$

Em modelos de efeitos mistos é comum adotar como distribuição para os efeitos aleatórios uma Normal q-variada com média 0 e matriz de variância e covariâncias Σ , ou seja, na especificação $2.16 \text{ K}(\Theta_b) = NMV_q(0,\Sigma)$. Para estes casos os principais métodos de aproximação da integral tem desempenhos melhores (BATES et al., 2015).

Como mencionado anteriormente modelos de efeitos aleatórios são candidatos a modelagem de dados superdispersos. Quando não há uma estrutura de delineamento experimental ou observacional pode-se incluir efeitos aleatórios ao nível de observação (e então m=n, ou seja, os vetores Y e b tem mesma dimensão). Casos particulares de modelos de efeitos aleatórios, onde o efeito aleatório é adicionado ao nível de observação são o modelo Binomial Negativo e o *Inverse Gaussian Model*, em ambos os casos a integral, definida em 2.17 tem solução analítica e consequentemente a marginal em Y forma fechada.

3 Material e Métodos

Essa seção é destinada a apresentação dos conjuntos de dados analisados no trabalho e descrição dos recursos computacionais e métodos utilizados na análise. Na seção 3.1.1 seis conjuntos de dados com diferentes características são apresentados. Os recursos computacionais utilizados são descritos na seção 3.1.2. Na última seção desse capítulo, 3.2, são apresentados os métodos para ajuste, avaliação e comparação dos modelos propostos.

3.1 Materias

3.1.1 Conjuntos de dados

A seguir são apresentados os seis conjuntos de dados utilizados para avaliar o desempenho dos modelos COM-Poisson. Os dados em estudo são, quase em sua totalidade, resultantes de experimentos agronômicos com delineamentos balanceados, o que é uma característica vantajosa para avaliação do desempenho do modelo COM-Poisson quando empregado a análise desses dados.

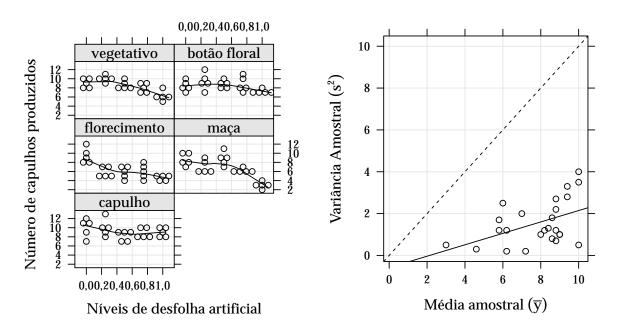
A apresentação dos conjuntos segue a ordem de 1) descrição do experimento ou estudo em destaque, 2) definição das variáveis e suas unidades de medidas e 3) descrição de suas características, potencialmente contempladas por modelos alternativos ao Poisson.

3.1.1.1 Capulhos de algodão sob efeito de desfolha artificial

Experimento conduzido sob delineamento inteiramente casualizado com cinco repetições em casa de vegetação com plantas de algodão *Gossypium hirsutum* submetidas a diferentes níveis de desfolha artificial de remoção foliar (0, 25, 50, 75, 100%), em combinação com o estágio fenológico no qual a desfolha foi aplicada (vegetativo, botão floral, florecimento, maça, capulho). A unidade experimental foi um vaso com duas plantas onde avaliou-se o número de capulhos produzidos ao final da ciclo cultura (SILVA et al., 2012). O experimento contou com 125 observações das quais têm-se as informações das variáveis número de capulhos de algodão produzidos (ncap), nível de desfolha de remoção foliar (des) e estágio fenológico das planta na unidade experimental (est).

Esse conjunto de dados já fora publicado sob a motivação da característica de subdispersão, utilizando o modelo *Gamma-Count* (ZEVIANI et al., 2014). Na figura 10, são apresentados os dados do experimento. À esquerda apresenta-se a disposição

das cinco observações em cada tratamento (combinação de nível de desfolha e estágio fenológico do algodão) e à direita um gráfico descritivo cruzando médias e variâncias amostrais calculadas em cada tratamento, onde a linha pontilhada representa a característica de equidispersão, média igual a variância. Em todos os tratamentos obteve-se a média menor que a variância apontando evidência de subdispersão.



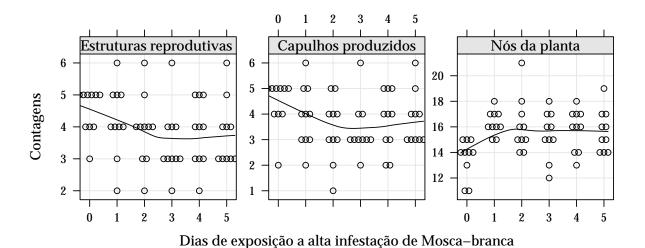
Fonte: Traduzido de Zeviani et al. (Figura 2)

Figura 10 – Número de capulhos produzidos para cada nível de desfolha e estágio fenológico (esquerda) e médias e variâncias das cinco repetições em cada combinação de nível de desfolha e estágio fenológico (direita).

3.1.1.2 Produtividade de algodão sob efeito de infestação de Mosca-branca

Experimento conduzido na Universidade Federal da Grande Dourados (UFGD) em 2007, cujo objetivo foi avaliar os impactos da exposição de plantas à alta infestação de Mosca-branca *Bemisia tabaci* em componentes de produção do algodão (MARTELLI et al., 2008). No experimento, plantas de algodão foram expostas à alta infestação da praga por diferentes períodos, 0, 1, 2, 3, 4, 5 dias onde avaliou-se o número de capulhos produzidos (ncapu), o número de estruturas reprodutivas (nerep) e o número de nós (nnos), como variáveis de interesse que representam a produtividade do cultivo de algodão. A condução do estudo deu-se via delineamento inteiramente casualizado com cinco vasos contendo duas plantas, para cada período de exposição.

Na figura 11 a disposição de cada uma das variáveis aleatórias de contagem número de estruturas reprodutivas, número de capulhos produzidos e número de nós da planta para os diferentes períodos em que as plantas estiveram sob alta infestação de 3.1. *Materias* 37



Fonte: Elaborado pelo autor.

Figura 11 – Disposição das variáveis de contagem nº de estruturas reprodutivas, nº de capulhos produzidos e nº de nós da planta observadas sob diferentes dias de exposição à infestação de Mosca-branca.

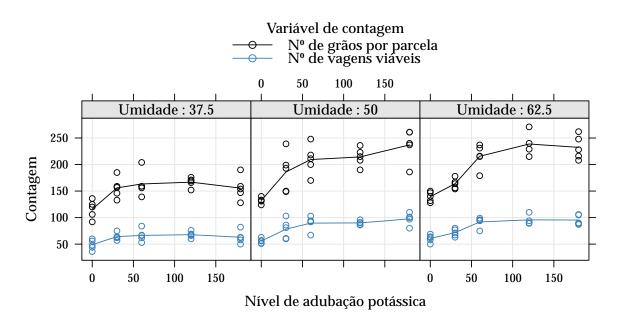
Mosca-branca é apresentada. Para todas as contagens parece haver um comportamento subdisperso. A indicação de subdispersão também se observa na tabela 2, onde as médias e variâncias amostrais calculadas com as dez observações nos seis períodos de exposição à infestação de Mosca-branca são exibidas. Em todos os casos observa-se as variâncias amostrais substancialmente menores que respectivas médias, ainda a manifestação de subdispersão é mais expressiva na variável número de nós da planta. Portanto, nesse experimento modelos alternativos ao Poisson devem ser empregados, pois a suposição de equidispersão é violada.

Tabela 2 – Médias e variâncias amostras das contagens avaliadas no experimento de capulhos de algodão sob efeito de Mosca-Branca

Dias de	N. Es	struturas	N. C	apulhos	N. Nós		
Exposição	média	variância	média	variância	média	variância	
0	4,50	0,50	4,40	0,93	13,60	2,27	
1	4,20	1,29	3,90	1,43	16,30	0,90	
2	3,90	1,21	3,40	1,60	16,10	4,54	
3	3,50	1,17	3,40	1,16	15,40	3,38	
4	3,80	1,07	3,70	1,34	15,80	2,62	
5	3,80	1,07	3,80	1,07	15,70	2,68	

Fonte: Elaborado pelo autor.

3.1.1.3 Produtividade de soja sob efeito de umidade do solo e adubação potássica

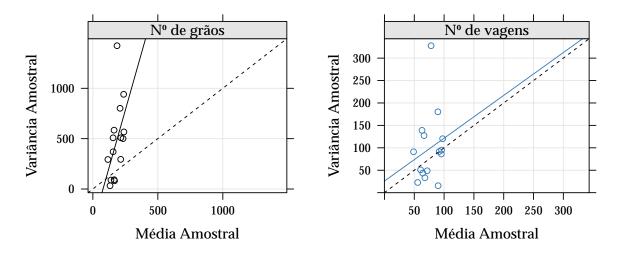


Fonte: Elaborado pelo autor.

Figura 12 – Disposição das variáveis número de grãos e número de vagens nos diferentes níveis de adubação potássica e umidade do solo.

Na figura 12 é apresentada a dispersão das contagens nas combinações das covariáveis umidade do solo e adubação potássica. As duas variáveis de contagem avaliadas no experimento apresentam níveis de dispersão distintos, essa característica fica explícita na figura 13, em que é exibida a dispersão entre médias e variâncias amostrais para cada uma das variáveis. Para o número de grãos por parcela, com contagens mais elevadas, as variâncias amostrais são, quase em sua totalidade, superiores as médias caracterizando uma evidência de superdispersão. Já para o número de vagens por parcela as médias e variâncias são, em média, próximas, o que indica que a suposição de equidispersão é razoável.

3.1. *Materias* 39



Fonte: Elaborado pelo autor.

Figura 13 – Médias e variâncias amostrais das contagens de grão e vagens, avaliadas no experimento com soja sob efeito umidade e adubação potássica.

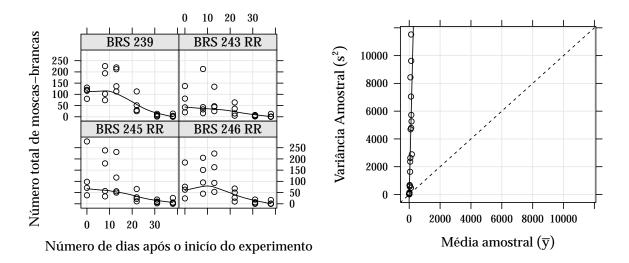
3.1.1.4 Ocorrência de ninfas de Mosca-branca em lavoura de soja

Nesse experimento também envolvendo a cultura de soja e a praga Mosca-branca, foram avaliadas plantas de quatro diferentes cultivares de soja, BRS 245 RR, BRS 243 RR, BRS 246 RR, BRS 239, contabilizando o número de ninfas de mosca-branca nos folíolos dos terços superior, médio e inferior das plantas em seis datas, 11/12/09, 19/12/09, 24/12/09, 02/01/10, 11/01/10, 18/01/10 dentre os 38 dias de estudo . O experimento foi conduzido em casa de vegetação sob o delineamento de blocos casualizados para controle de variação local (SUEKANE, 2011).

As contagens da praga para cada cultivar em cada uma das datas de avaliação, representadas pelos dias decorridos após a primeira avaliação, em 11/12/09 são mostradas na figura 14 a esquerda. As contagens são muito altas e dispersas, principalmente nas quatro primeiras avaliações. A direita uma descrição no nível de dispersão da variável de contagem é apresentada. Esse é um conjunto de dados extremamente superdisperso. Os pontos, que representam as médias e variâncias em cada combinação de cultivares de soja e dias após a primeira avaliação, estão todos acima da reta identidade (de equidispersão) com variâncias em torno de 1.000 vezes maiores que as respectivas médias.

3.1.1.5 Peixes Capturados por Pescadores em um Parque Estadual

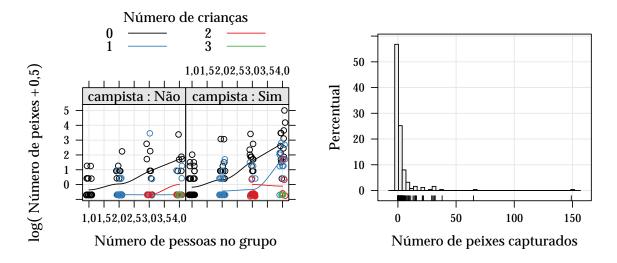
Diferentemente dos demais, esse é um estudo observacional feito por biólogos com interesse em modelar o número de peixes capturados por grupos de pescadores visitantes em um Parque Estadual (UCLA, 2015). Nesse estudo tem-se como informações



Fonte: Elaborado pelo autor

Figura 14 – Dispersão entre o número total de ninfas de Mosca-branca nos folíolos da soja e o número de dias após a primeira avaliação para as quatro diferentes cultivares (esquerda). Relação entre as médias e as variâncias amostrais do número de ninfas nesse experimento (direita).

a respeito dos grupos de visitantes, o número de pessoas e de crianças e se há ou não a presença de campista. Um fato interessante deste dado é que nem todos os grupos de visitantes praticaram pescaria, portanto, nesses grupos o número de peixes capturado será zero.



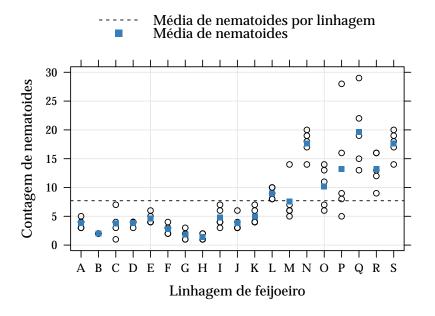
Fonte: Elaborado pelo autor.

Figura 15 – Logarítmo neperiano do número de peixes capturados acrescido de 0,5 para as diferentes composições dos grupos (esquerda). Histograma do número de peixes capturados por grupo (direita).

3.1. Materias 41

Na figura 15 é evidente o excesso de contagens zero. No gráfico à esquerda tem-se a disposição das contagens, transformadas por $\log(y_i|x_i+0.5)$. É característica marcante no gráfico a grande quantidade de pontos dispostos no primeiro valor do eixo y, $-0.693 = \log(0.5)$. Embora seja um gráfico marginal, não considerando as covariáveis de cada contagem, a direita um histograma da variável resposta é realizado e percebe-se novamente a grande quantidade de valores nulos, ao todo 56.8% dos dados são contagens nulas. Portanto nesse problema, claramente modelos alternativos que acomodem excesso de zeros se fazem necessários.

3.1.1.6 Número de nematoides em raizes de feijoeiro



Fonte: Elaborado pelo autor.

Figura 16 – Dispersão do número de nematoides providos por uma alíquota da solução de 1 g/ml de massa fresca diluída.

Esse último conjunto de dados explorado no trabalho, é resultado de um experimento em casa de vegetação que estudou a reprodução de nematoides em cultivares de feijoeiro. No experimento, o solo de vasos com duas plantas de feijão foi inicialmente contaminado com nematoides e as raizes das duas plantas por vaso foram, ao final do experimento, lavadas, trituradas, peneiradas e diluídas e, a partir de alíquotas dessa solução, contou-se o número de nematoides. Como denominador da contagem tem-se a razão entre a massa fresca de raizes (em gramas) por parcela e o volume de água (em milímetros) utilizado para diluir essa quantidade 1 .

Cedido para fins acadêmicos por Andressa Cristina Zamboni Machado, pesquisadora do Instituto Agronômico do Paraná (IAPAR), e pelo técnico agrícola do IAPAR Santino Aleandro da Silva

Na figura 16 a dispersão das contagens de nematoides em aliquotas da solução de uma grama de massa fresca de raiz por um milímetro de água para cada linhagem é exibida. As contagens para cada uma das linhagens se distribuem em torno do perfil médio (linha pontilhada). Um detalhe interesse desse conjunto de dados é que o efeito das linhagens pode ser considerado aleatório em certas fazes do programa de melhoramento genético. Portanto, pode-se interpretar as linhagens escolhidas como um sorteio aleatório dentre uma população de linhagens de feijoeiro. Assim, modelos com efeitos aleatórios a nível de linhagem são capazes de representar as características distintas de cada linhagem por meio de uma distribuição de probabilidades.

3.1.2 Recursos computacionais

O software R, versão 3.3.0, é utilizado tanto para a preparação e apresentação dos dados quanto para ajuste dos modelos e apresentação de resultados. Pacotes auxiliares utilizados no trabalho são: MASS (7.3.45) para ajuste e inferências dos modelos Binomial Negativo, bbmle (versão 1.0.18) para estimação via máxima verossimilhança das funções implementadas para o modelo COM-Poisson , psc1 (1.4.9) para ajuste dos modelos Poisson e Binomial Negativo com componente de barreira para modelagem de excesso de zeros e 1me4 (versão 1.1.12) para ajuste dos modelos Poisson com efeitos aleatórios normais. Para apresentação gráfica dos resultados os pacotes 1attice (0.20.33), 1atticeExtra (0.6.28) e corrplot (0.73) são exaustivamente utilizados. Finalmente, para elaboração do relatório, mesclando códigos em R e escrita na linguagem de marcação LATEX, o pacote knitr (1.12.3) é requerido.

Destaca-se nesse trabalho que todas as funções implementadas para ajuste e inferência dos modelos de regressão COM-Poisson estão disponíveis, em formato de um pacote R, cmpreg, no endereço https://github.com/JrEduardo/cmpreg. No apêndice A o emprego do pacote na análise de um conjunto de dados exibido no trabalho é ilustrado com códigos R.

3.2 Métodos

A estimação dos parâmetros do modelo de regressão COM-Poisson de efeitos fixos é realizada maximizando uma forma reparametrizada da log-verossimilhança, definida na expressão 2.13, via algoritmo numérico de otimização *BFGS*. O parâmetro extra da COM-Poisson, ν tem suporte nos reais positivos, restringindo o espaço paramétrico de busca do otimizador, o que é numericamente indesejável. Para deixar o domínio de busca nos reais reparametrou-se o modelo com o parâmetro $\phi = \log(\nu)$, como

3.2. Métodos 43

 $0<
u<\infty$ então $-\infty<\phi<\infty$. Sob a reparametrização a função a ser maximizada é

$$\ell(\phi, \beta \mid \underline{y}) = \sum_{i}^{n} y_{i} \log(\lambda_{i}) - e^{\phi} \sum_{i}^{n} \log(y!) - \sum_{i}^{n} \log(Z(\lambda_{i}, \phi))$$
(3.1)

em que $\lambda_i = e^{X_i \beta}$, com X_i o vetor $(x_{i1}, x_{i2}, \dots x_{ip})$ de covariáveis da i-ésima observação, e $(\beta, \phi) \in \mathbb{R}^{p+1}$.

O ajuste do modelo é realizado sob ϕ . Portanto as inferências decorrentes são sobre esse parâmetro. Todavia pode-se retornar para parametrização original utilizando a função inversa em valores pontuais ou método delta para funções de ϕ . Nesse trabalho as inferências são realizadas sob o parâmetro ϕ . Para esse parâmetro as interpretações são como se segue

$$\phi < 0 \Rightarrow$$
 Superdispersão $\phi = 0 \Rightarrow$ Equidispersão $\phi > 0 \Rightarrow$ Subdispersão

ou seja, possui a interpretação de um parâmetro de precisão.

A partir dessa reparametrização a condução de testes de hipóteses é facilitada. Uma vez que $\phi=0$, representa o caso particular em que a COM-Poisson se reduz a Poisson, a estatística

$$TRV = 2 \cdot (\ell_{CMP} - \ell_P) \sim \chi_1^2$$

sendo ℓ_{CMP} e ℓ_P as log-verossimilhanças maximizadas dos modelos COM-Poisson e Poisson com mesmo preditor linear respectivamente, se refere ao teste de razão de verossimilhanças para $H_0: \phi = 0$, ou de forma mais apelativa, ao teste sobre a equivalência dos modelos COM-Poisson e Poisson.

A partir da definição em 2.14, para incluir um componente de barreira no modelo COM-Poisson, acomodando excesso de zeros, adota-se para $\Pr(Z=z\mid\Theta_c)$ a distribuição COM-Poisson (2.8) resultando em

$$\Pr(Y = y \mid \pi, \phi, \lambda) = \begin{cases} \pi & \text{se } y = 0, \\ (1 - \pi) \frac{\lambda^{y}}{(y!)^{e^{\phi}} Z(\lambda, \phi)} \left(1 - \frac{1}{Z(\lambda, \phi)} \right)^{-1} & \text{se } y = 1, 2, \dots \end{cases}$$
(3.2)

Para modelos de regressão com componente de barreira, são incorporados preditores lineares em π , $\underline{\pi} = \frac{\exp(G\gamma)}{1+\exp(G\gamma)}$ e λ , $\underline{\lambda} = \exp(X\beta)$ e a verossimilhança desse modelo

toma a forma

$$\mathcal{L}(\phi, \beta, \gamma \mid \underline{y}) = \mathbb{1}[\underline{\pi}] \cdot (1 - \mathbb{1}) \left[(1 - \underline{\pi}) \left(\frac{\underline{\lambda}^{y}}{(y!)^{e^{\phi}} Z(\underline{\lambda}, \phi)} \right) \left(1 - \frac{1}{Z(\underline{\lambda}, \phi)} \right) \right]$$
(3.3)

em que $\mathbbm{1}$ é uma função indicadora para y=0. Os argumentos $\hat{\phi}$, $\hat{\beta}$ e $\hat{\gamma}$, que maximizam o logaritmo neperiano da função 3.3 serão as estimativas de máxima verossimilhança do modelo COM-Poisson com componente de barreira.

Uma outra extensão proposta para o modelo COM-Poisson é a inclusão de efeitos aleatórios a fim de modelar a estrutura experimental ou observacional de um conjunto de dados. Este trabalho restringe-se a inclusão de efeitos aleatórios Normais, ou seja, $b \sim \text{Normal}(0, \Sigma)$, que são incorporados sob a forma $\underline{\lambda} = X\beta + Zb$ conforme especificação em 2.16. Assim, considerando a distribuição COM-Poisson para a variável resposta condicionada as covariáveis e os efeitos aleatórios, a verossimilhança pode ser escrita como

$$\mathcal{L}(\phi, \Sigma, \beta \mid \underline{y}) = \prod_{i=1}^{m} \int_{\mathbb{R}^{q}} \left(\prod_{j=1}^{n_{i}} \frac{\underline{\lambda}^{y}}{(y!)^{e^{\phi}} Z(\underline{\lambda}, \phi)} \right) \cdot (2\pi)^{q/2} |\Sigma| \exp\left(-\frac{1}{2} b^{t} \Sigma^{-1} b\right) db_{i} \quad (3.4)$$

sendo m o número de grupos que compartilham do mesmo efeito aleatório, q o número de efeitos aleatórios (intercepto aleatório, inclinação e intercepto aleatórios, etc.) e n_i o número de observações no i-ésimo grupo. A integração em 3.4, necessária para a avaliação da verossimilhança não tem forma analítica. Utiliza-se a aproximação de Laplace da forma como apresentada em Ribeiro Jr et al. (2012, pág. 141) para aproximação dessa integral. A estimação dos parâmetros é realizada via maximização da $\log(\mathcal{L}(\phi,\Sigma,\beta\mid\underline{y}))$ com métodos numéricos de otimização. Ressalta-se que esse é um procedimento computacionalmente intensivo, pois a cada iteração do algoritmo de maximização, m aproximações de Laplace para integrais de dimensão q são realizadas. Ainda, quando considerada a distribuição COM-Poisson para a variável resposta condicionalmente independente, tem-se também o cálculo de n_m constantes normalizadoras $Z(\lambda,\phi)$ (2.9) para cada m grupo em cada iteração do algoritmo de otimização. Com toda essa estrutura hierárquica, procedimentos computacionais realizados a cada estágio são potencialmente instáveis numericamente.

Para comparação entre os modelos COM-Poisson e demais modelos listados no capítulo 2 utiliza-se essencialmente o valor maximizado da log-verossimilhança e o critério de informação de Akaike (AIC) definido como

$$AIC = 2(k - \ell(\Theta_k, y))$$
(3.5)

3.2. *Métodos* 45

sendo k o número de parâmetros e $\ell(\Theta_k, \underline{y})$ a logverossimilhança maximizada do modelo definido pelo conjunto Θ_k de parâmetros. Nas análises compara-se também, os níveis descritivos nos testes de razão de verossimilhanças entre modelos encaixados. Nos modelos de regressão de efeitos fixos os valores preditos pelos modelos COM-Poisson e demais alternativas pertinentes são exibidos graficamente com bandas de confiança.

Para maximização numérica das logverossimilhanças dos modelos de regressão COM-Poisson e suas extensões utiliza-se um método de otimização quasi-Newton bastante popular, denominado *BFGS* (NOCEDAL; WRIGHT, 1995). As informações do vetor gradiente (derivadas de primeira e matriz hessiana (derivadas de segunda ordem) são obtidos numericamente via aproximação de diferenças finitas.

4 Resultados e Discussão

Nesse capítulo são apresentados os resultados e discussões da aplicação modelos de regressão COM-Poisson ajustados aos dados apresentados na seção 3.1.1 comparando-as com abordagens já utilizadas na Estatística aplicada. As primeiras seis seções são destinadas a apresentação das análises estatísticas de cada conjunto de dados citado. Na seção 4.7 discussões gerais sobre os resultados dos modelos COM-Poisson empregados nas análises são realizadas.

4.1 Análise de dados de capulhos de algodão sob efeito de desfolha

Diante da estrutura do experimento apresentada na seção 3.1.1.1 foram propostos, por Zeviani et al. (2014), cinco preditores crescentes em complexidade que testam aspectos interesses sobre os fatores experimentais. Abaixo os cinco preditores considerados são descritos.

```
Preditor 1: g(\mu) = \beta_0

Preditor 2: g(\mu) = \beta_0 + \beta_1 \text{def}

Preditor 3: g(\mu) = \beta_0 + \beta_1 \text{def} + \beta_2 \text{def}^2

Preditor 4: g(\mu) = \beta_0 + \beta_{1j} \text{def} + \beta_2 \text{def}^2

Preditor 5: g(\mu) = \beta_0 + \beta_{1j} \text{def} + \beta_{2j} \text{def}^2
```

onde j varia nos níveis de estágio fenológico da planta (1: vegetativo, 2: botão floral, 3: florescimento, 4: maça, 5: capulho) e $g(\mu)$ uma função de ligação. A proposta desses preditores foi realizada de forma aninhada a fim de facilitar a condução de testes de hipóteses. O modelo 1 contêm somente o intercepto, e 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 segunda ordem. O modelo 4, apresenta o efeito de desfolha linear mudando de acordo com o estágio de crescimento (interação entre o efeito linear de desfolha e estágio), e por fim o modelo 5 não somente o efeito de primeira ordem muda com o estágio de crescimento, mais também o efeito de segunda ordem (interação entre o efeito de primeira e segunda ordem de desfolha e estágio).

A seguir são ajustados os modelos Poisson e COM-Poisson como alternativas paramétricas à análise de dados e como alternativa semi-paramétrica a estimação via quasi-verossimilhança Poisson. Na tabela 3 os resultados dos três modelos ajustados aos cinco preditores são apresentados. O modelo COM-Poisson apresentou melhor ajuste dentre todos os preditores considerados quando comparado ao Poisson, indicado pelas

maiores log-verossimilhanças e menores AIC's.

Tabela 3 – Medidas de ajuste para avaliação e comparação entre preditores e modelos ajustados

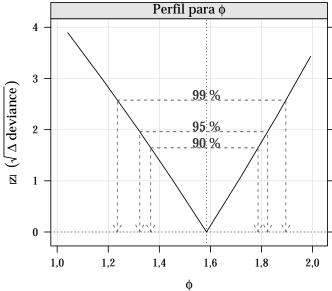
Poisson	np	ℓ	AIC	2(diff ℓ)	diff np	$P(>\chi^2)$		
Preditor 1	1	-279,93	561,87					
Preditor 2	2	-272,00	548,00	15,86	1	6,81E-05		
Preditor 3	3	-271,35	548,71	1,29	1	2,56E-01		
Preditor 4	7	-258,67	531,35	25,36	4	4,26E-05		
Preditor 5	11	-255,80	533,61	5,74	4	2,19E-01		
COM-Poisson	np	ℓ	AIC	$2(\text{diff }\ell)$	diff np	$P(>\chi^2)$	$\hat{\phi}$	$P(>\chi^2)$
Preditor 1	2	-272,48	548,96				0,551	1,13E-04
Preditor 2	3	-257,46	520,93	30,03	1	4,25E-08	0,794	6,97E-08
Preditor 3	4	-256,09	520,18	2,75	1	9,73E-02	0,816	3,29E-08
Preditor 4	8	-220,20	456,40	71,78	4	9,54E-15	1,392	1,75E-18
Preditor 5	12	-208,25	440,50	23,90	4	8,38E-05	1,585	1,80E-22
Quase-Poisson	np	deviance	AIC	F	diff np	P(> <i>F</i>)	$\hat{\sigma}^2$	$P(>\chi^2)$
Preditor 1	1	75,51					0,567	3,66E-04
Preditor 2	2	59,65		34,21	1	4,17E-08	0,464	5,13E-07
Preditor 3	3	58,36		2,81	1	9,62E-02	0,460	3,66E-07
Preditor 4	7	33,00		22,77	4	5,89E-14	0,278	9,15E-16
Preditor 5	11	27,25		5,96	4	2,18E-04	0,241	3,57E-18

np, número de parâmetros, diff ℓ , diferença entre log-verossimilhanças, F, estatística F baseada nas quasi-deviances, diff np, diferença entre o np.

Fonte: Elaborado pelo autor.

As estimativas dos parâmetros extras ϕ e σ^2 dos modelos COM-Poisson e Quasi-Poisson respectivamente, também são apresentadas na tabela 3 e indicam subdispersão ($\phi > 0$ e $\sigma^2 < 1$). Note que, mesmo não considerando covariáveis, preditor 1, a hipótese de equidispersão foi rejeitada pelo modelos COM-Poisson e Quasi-Poisson. Isso se reflete nos níveis descritivos dos testes de razão de verossimilhanças realizados, em que o modelo Poisson, em discordância com os demais, não indicou significância do efeito quadrático por nível de desfolha, preditor 5, pois superestima a variabilidade do processo. Esses resultados estão de acordos com os apresentados por Zeviani et al. (2014), onde um modelo *Gamma-Count* foi ajustado, destaca-se a similaridade entre as medidas de ajuste dos modelos COM-Poisson e *Gamma-Count*. Os valores das logverossimilhanças maximizadas nos dois modelos difere somente nas casas decimais, para todos os preditores.

Na figura 17 a avaliação do parâmetro ϕ do modelo COM-Poisson com efeito de desfolha artificial de primeira e segunda ordem para cada estágio fenológico, via verossimilhança perfilhada é apresentada. O valor zero, que representa a não necessidade de um modelo COM-Poisson está dentro dos limites de confiança de 90, 95 e



Fonte: Elaborado pelo autor.

Figura 17 – Perfil de log-verossimilhança para o parâmetro extra da COM-Poisson, estimado no modelo com o quinto preditor.

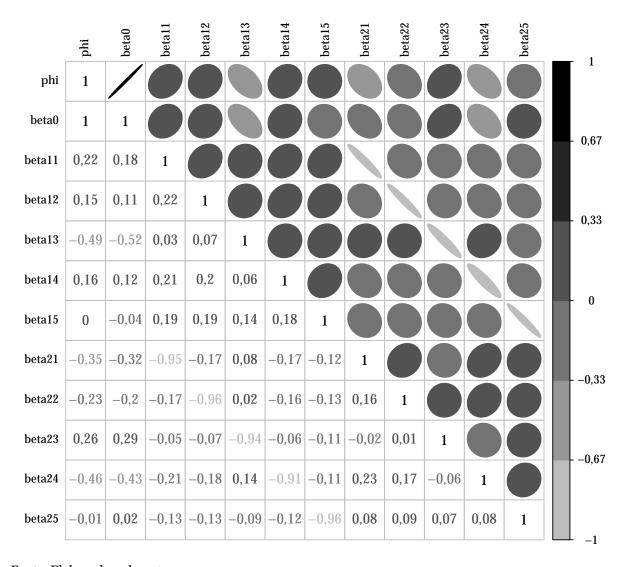
até 99%. A simetria do perfil de verossimilhança também é algo para se destacar, pois neste caso intervalos do tipo Wald (computacionalmente mais fáceis), via aproximação quadrática da verossimilhança, podem ser construídos, muito embora os construídos via perfil de log-verossimilhança sejam preferíveis. Em concordância com a figura, o teste de hipóteses via razão de verossimilhanças para $H_0: \phi = 0$, rejeitou a hipótese nula com um nível de significância muito próximo a zero, tabela 3.

Tabela 4 – Estimativas dos parâmetros e razões entre as estimativa e erro padrão para os três modelos em estudo

	Poiss	on	Quasi-Po	oisson	COM-Poisson		
Parâmetro	Estimativa	Est/EP	Estimativa	Est/EP	Estimativa	Est/EP	
σ^2 , ϕ	_	_	0,24	_	1,58	12,42	
eta_0 .	2,19	34,57	2,19	70,42	10,90	7,76	
β_{11}	0,44	0,85	0,44	1,73	2,02	1 <i>,</i> 77	
β_{12}	0,29	0,57	0,29	1,16	1,34	1,21	
eta_{13}	-1,24	-2,06	-1,24	-4,19	<i>-5,</i> 75	-3,89	
eta_{14}	0,36	0,64	0,36	1,31	1,60	1,30	
eta_{15}	0,01	0,02	0,01	0,04	0,04	0,03	
eta_{21}	-0,81	-1,38	-0,81	-2,81	-3,72	-2,78	
β_{22}	-0,49	-0,86	-0,49	<i>-</i> 1 <i>,</i> 75	-2,26	-1,80	
β_{23}	0,67	0,99	0,67	2,01	3,13	2,08	
β_{24}	-1,31	-1,95	-1,31	-3,97	-5,89	-3,66	
β_{25}	-0,02	-0,04	-0,02	-0,07	-0,09	-0,08	

Fonte: Elaborado pelo autor.

As estimativas dos efeitos lineares e quadráticos de desfolha artificial, conforme notação do preditor 5, são apresentadas na tabela 4 para os modelos Poisson, Quasi-Poisson e COM-Poisson. Para os modelos Poisson e Quasi-Poisson as estimativas são idênticas, por construção 2.1, o que difere são as magnitudes dessas estimativas em comparação com seu erro padrão, que no caso Quasi-Poisson é corrigido pelo parâmetro σ^2 . Considerando o modelo COM-Poisson as estimativas são notavelmente diferentes, pois o preditor linear é construído em λ , da expressão 2.8, e este parâmetro não descreve, diretamente, a média da distribuição. Sendo assim as estimativas do COM-Poisson não podem ser comparadas com as demais estimativas. Contudo, a magnitude desses efeitos com relação ao efeito padrão sim. E neste caso os modelos Quasi-Poisson e COM-Poisson levam as mesmas conclusões.

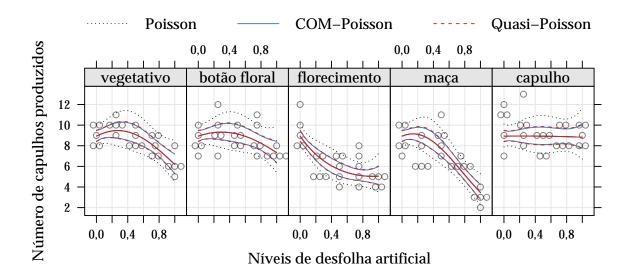


Fonte: Elaborado pelo autor.

Figura 18 – Imagem da matriz de correlação entre os parâmetros do modelo COM-Poisson.

As covariâncias entre as estimativas dos parâmetros do modelo COM-Poisson são apresentadas, na escala da correlação, na figura 18. Destaca-se nessa figura a forte correlação do parâmetro de precisão ϕ com os β 's da regressão. Embora seja uma representação empírica, observada a esse particular conjunto de dados, nota-se a não ortogonalidade na matriz de informação observada, o que implica que inferências sobre os β 's são condicionais a ϕ . Esse comportamento dos modelos COM-Poisson é recorrente, como será visto nos demais conjuntos de dados.

Essa característica de não ortogonalidade da matriz de informação observada teve de ser levada em consideração para cálculo dos valores preditos, uma vez que a informação sobre a incerteza das estimativas contida na matriz de variâncias e covariâncias não pôde ser marginalizada para os β 's, que efetivamente são utilizados para cálculo de $\hat{\lambda}_i$ e consequentemente $\hat{\mu}_i$. Portanto, para cálculo dos valores preditos utiliza-se a matriz de variâncias e covariâncias condicionada a ϕ , conforme Ferreira (2011, teorema 3.6, pág. 123). Essa é uma prática tomada também para cálculo dos valores preditos nos demais conjunto de dados.



Fonte: Elaborado pelo autor.

Figura 19 – Curva dos valores preditos com intervalo de confiança de (95%) como função do nível de desfolha e do estágio fenológico da planta.

As médias com intervalos de confiança calculadas com os modelos COM-Poisson e Quasi-Poisson são praticamente idênticas, conforme pode ser visto na figura 19. Contudo, destaca-se que o modelo COM-Poisson é totalmente paramétrico permitindo representar uma distribuição, calculando probabilidades, o que não é possível com a formulação Quasi-Poisson. Ainda nota-se claramente que o modelo Poisson é inadequado a esse conjunto de dados e que inferências a partir deste seriam incorretas.

4.2 Análise de dados de capulhos de algodão sob efeito de Mosca-Branca

Nesse conjunto de dados também há indícios de subdispersão para as três variáveis de interesse mensuradas no estudo, conforme apresentado na seção 3.1.1.2. Para cada contagem procedeu-se com o ajuste dos modelos Poisson, Quasi-Poisson e COM-Poisson adotando os preditores:

Preditor 1: $g(\mu) = \beta_0$

Preditor 2: $g(\mu) = \beta_0 + \beta_1 \text{dexp}$

Preditor 3: $g(\mu) = \beta_0 + \beta_1 \text{dexp} + \beta_2 \text{dexp}^2$

sendo dexp a variável dias de exposição à alta infestação de mosca-branca. Assim os preditores 1, 2, 3 representam efeito nulo, linear e quadrático dos dias de exposição, respectivamente.

Tabela 5 – Medidas de ajuste para avaliação e comparação entre preditores e modelos ajustados

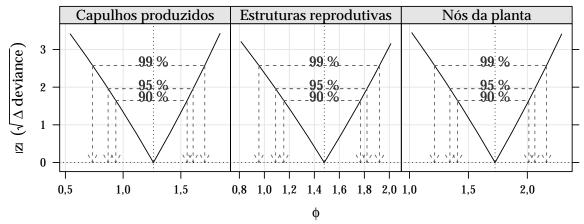
		Poisson	l	C	OM-Pois	son	Quasi-Poisson	
np	ℓ	AIC	$P(>\chi^2)$	ℓ	AIC	$P(>\chi^2)$	deviance	P(> F)
Núme	ro de capu	lhos prod	uzidos	-				
1	-105,27	212,55		-92,05	188,09		20,80	
2	-105,03	214,05	4,83E-01	-91,31	188,62	2,25E-01	20,31	2,23E-01
3	-104,44	214,88	2,78E-01	-89,47	186,95	5,52E-02	19,13	6,16E-02
Núme	ro de estru	turas rep	rodutivas	_				
1	-104,74	211,49		-86,41	176,82		16,23	
2	-104,27	212,54	3,32E-01	-84,59	175,18	5,66E-02	15,29	6,19E-02
3	-104,06	214,12	5,16E-01	-83,73	175,47	1,90E-01	14,87	2,07E-01
Núme	ro de nós c	la planta		_				
1	-143,79	289,59		-120,58	245,16		12,69	
2	-143,48	290,95	4,25E-01	-119,03	244,06	7,87E-02	12,05	7,39E-02
3	-142,95	291,89	3,04E-01	-116,27	240,54	1,88E-02	11,00	2,23E-02

np, número de parâmetros. Fonte: Elaborado pelo autor.

Na tabela 5 são exibidas as medidas de ajuste dos modelos para as três variáveis resposta. Em todos os casos o modelo COM-Poisson apresentou maiores logverossimilhanças indicando um melhor ajuste, quando comparado ao Poisson, também indicado pelos os valores de AIC que ponderam a log-verossimilhança pelo número de parâmetros considerados no modelo. Para questões inferenciais novamente, há um desacordo entre os modelos paramétricos. Pelos modelos Poisson não há evidências para manutenção de nenhum efeito da variável número de dias sob infestação, em todos

os casos, ao passo que no modelo COM-Poisson tem-se evidências do efeito quadrático quando considerado o modelo para o número de nós da planta (nível descritivo de 0,981) e o número de capulhos produzidos (nível descritivo de 0,945, na borda da região de significância, mas com uma diminuição do AIC em favor do efeito quadrático). Quando modelado o número de estruturas reprodutivas o modelo COM-Poisson também não indicou efeito quadrático, contudo o efeito linear de dexp pode ser discutido uma vez que a significância do TRV foi de 0,055 e o AIC apresentou um pequeno aumento com relação ao modelo nulo. Considera-se nas demais inferências os preditores com efeitos linear, para o número de estruturas reprodutivas e quadrático, para o número de capulhos produzidos e número de nós da planta.

A especificação do modelo via Quasi-Verossimilhança Poisson obteve níveis descritivos mais conservadores para a rejeição da hipótese nula que o modelo COM-Poisson. Contudo, para escolha de preditores as mesmas tendências apontadas pelo COM-Poisson foram seguidas.



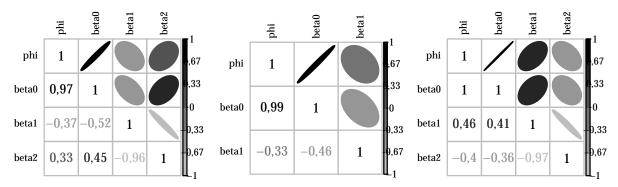
Fonte: Elaborado pelo autor.

Figura 20 – Perfis de log-verossimilhança para o parâmetro extra da COM-Poisson nos modelos para número de capulhos produzidos (esquerda), número de estruturas reprodutivas (central) e número de nós (direira).

Para avaliação do parâmetro ϕ da COM-Poisson nos três modelos considerados, intervalos de confiança construídos sob perfilhamento da verossimilhança são exibidos na figura 20. Para nenhum dos modelos o valor de $\phi=0$ esteve dentro dos limites de confiança de 90, 95 e 99%. Os valores estimados dos parâmetros nos modelos para número de capulhos, número de estruturas reprodutivas e número de nós da planta foram de 1,263, 1,479, 1,726 respectivamente, indicando subdispersão em todos os casos.

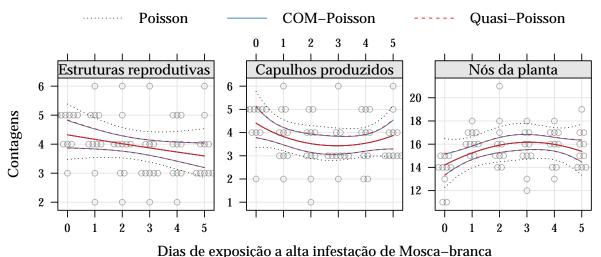
Na figura 21 são representadas as matrizes de covariâncias (via correlações) entre as estimativas dos modelos para número de capulhos, à esquerda, número de estruturas reprodutivas, ao centro e número de nós da plantas, à direita. A forte correlação entre

o parâmetro de precisão ϕ e β_0 (principalmente) também foi observada no ajuste do modelo para esses conjuntos de dados.



Fonte: Elaborado pelo autor.

Figura 21 – Imagem da matriz de correlação entre os parâmetros do modelo COM-Poisson. (Esquerda) Modelo para o número de capulhos por parcela, (centro) para o número de estruturas reprodutivas e (direita) para o número de nós por parcela.



Dias de exposição à aita infestação de Mosca-Dranc

Fonte: Elaborado pelo autor.

Figura 22 – Curva dos valores preditos com intervalo de confiança de (95%) como função dos dias de exposição a alta infestação de Mosca-branca considerando os modelos para o número de estruturas reprodutivas (esquerda), número de capulhos produzidos (centro) e número de nós (direita).

Finalmente a representação gráfica na figura 22 mostra os valores preditos pelos modelos Poisson, COM-Poisson e Quasi-Poisson com intervalos de confiança para média com 95% de confiança. Assim como na análise realizada na seção 4.1, os valores preditos com bandas de confiança obtidos dos modelos COM-Poisson e Quasi-Poisson, são praticamente idênticos levando as mesmas interpretações.

Com esse segundo exemplo de subdispersão, em que três contagens foram realizados em um único experimento. A flexibilidade do modelo COM-Poisson no que tange à característica de subdispersão ganha destaque, uma vez que seus resultados (predições pontuais e intervalares e testes de hipóteses para comparação de modelos) se equivalem a uma abordagem semi-paramétrica.

4.3 Análise de produção de soja sob efeito de umidade e adubação potássica

Nesse experimento apresentado em 3.1.1.3, mais de uma variável de interesse em forma de contagem é mensurada e pela descrição dos dados características relacionadas a dispersão da contagem são distintas em ambas (equidispersão e superdispersão). Dos modelos apresentados no capítulo 2, o Poisson, COM-Poisson, Binomial-Negativo são as alternativas paramétricas avaliadas e o Quasi-Poisson é tomado como a alternativa semi-paramétrica. As variáveis de interesse números de grãos de soja e de vagens viáveis foram contabilizados por unidade experimental (vaso com duas plantas) e estão sob o efeito, controlado, de duas covariáveis, níveis de adubação potássica (0, 30, 60, 120, 180 mg dm⁻³) e níveis de umidade do solo (37.5, 50, 62.5 % do volume total dos poros), que foram considerados na análise como fatores com 5 e 3 níveis respectivamente. Ainda têm-se, pela condução do experimento, o efeito relacionado a blocagem realizada, foram cinco blocos utilizados para controle de variação local. Os preditores considerados são

Preditor 1:
$$\eta_1 = g(\mu_{ijk}) = \beta_0 + \tau_i + \gamma_j + \delta_k$$

Preditor 1: $\eta_2 = g(\mu_{ijk}) = \beta_0 + \tau_i + \gamma_j + \delta_k + \alpha_{jk}$

em que τ_i é o efeito do i-ésimo bloco, i=1: bloco II, 2: bloco III, 3: bloco IV e 4: V; γ_j o efeito do j-ésimo nível de umidade aplicado, j=1: 50% e 2: 62,5%; δ_k o efeito do k-ésimo nível de adubação potássica, k=1: 30, 2: 60, 3: 20 e 4: 180 mg dm $^{-3}$ e α_{jk} o efeito da interação entre o j-ésimo nível de umidade do solo e o k-ésimo nível de adubação potássica. Assim no modelo mais completo, com interação, são 19 parâmetros de locação a serem estimados.

Para ajuste dos modelos COM-Poisson nesse exemplo o tempo computacional foi ligeiramente mais demorado (em torno de 10s para os quatro modelos considerando as duas contagens e os dois preditores). Isso se deve ao fato das contagens serem altas (variando entre 92 e 271 para o número de grãos e 36 e 110 para o número de vagens) e superdispersas (ϕ < 0). Nesse cenário os incrementos da constante normalizadora $Z(\lambda_i, \nu = \exp(\phi))$, expressão 2.9, convergem para 0 mais lentamente.

Na figura 23 são exibidos os termos dessa constante para cada observação nos modelos mais complexos (com interação) para o número de vagens e para o número

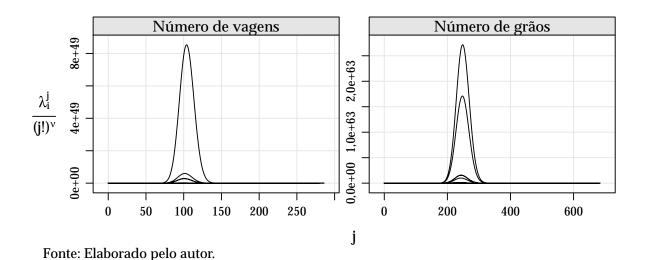


Figura 23 – Convergência das constantes de normalização para cada indivíduo no modelo para o número de vagens viáveis (esquerda) e para o número de grãos produzidos (direita)

de grãos. O critério de convergência adotado foi de $\lambda^j/(j!)^\nu < 1 \times 10^{-3}$. No modelo para número de vagens o maior valor para a constante foi de 2,048 × 10⁵¹, soma de 287 termos, calculados para a observação 10, que teve o maior valor estimado para o parâmetro λ , $\hat{\lambda}=5,286$. Nesse o modelo o parâmetro ϕ foi estimado em 0,129. Já no modelo para o número de grãos foram necessários 685 termos que somados resultaram em 1,391 × 10⁶⁵, maior constante calculada. Isso também se deu na observação 10 que para este modelo, com $\hat{\phi}=-0,518$, teve um parâmetro λ estimado em 3,287.

Medidas de qualidade de ajuste calculadas sob os modelos Poisson, COM-Poisson, Binomial Negativo e Quasi-Poisson são apresentadas na tabela 6. Considerando a variável resposta número de vagens viáveis, não há indícios de afastamento da equidispersão indicados i) pelos parâmetros extras dos modelos alternativos ao Poisson, em que estimativas $\hat{\phi}$, $\hat{\theta}$ e $\hat{\sigma^2}$ estão próximas dos valores 0, ∞ e 1, que compreendem o caso particular Poisson nos modelos COM-Poisson, Binomial Negativo e Quasi-Poisson respectivamente, ii) pelas log-verossimilhanças dos modelos paramétricos que resultaram em valores muito próximos, iii) pelos valores de AIC que foram menores nos modelos Poisson, mostrando que não há ganho expressivo quando estimados os parâmetros extra dos modelos alternativos. Os p-valores associados ao TRV entre os modelos COM-Poisson e Poisson com preditores 1 e 2 foram de 0,671, 0,446, evidenciando a não fuga de equidispersão dos dados. Na figura 24 à esquerda são apresentados os intervalos de confiança baseados no perfil de verossimilhança para ϕ , no modelo COM-Poisson com efeito de interação, como esses intervalos contém o valor da hipótese nula 0, o modelo COM-Poisson pode ser reduzido ao Poisson. Para avaliação dos preditores, novamente tem-se um caso de valores na borda de significância. Nas análises que a seguir o modelo

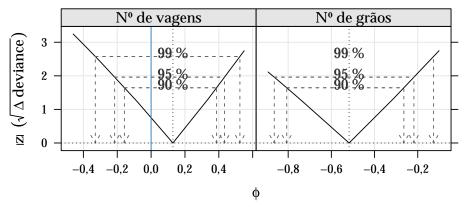
mais completo com a interação entre adubação e umidade é considerado.

Tabela 6 – Medidas de ajuste para avaliação e comparação entre preditores e modelos ajustados ao número de vagens e ao número de grão por parcela

			Númer	o de vagen	s	Número de grãos			
РО	np	ℓ	AIC	$P(>\chi^2)$		ℓ	AIC	$P(>\chi^2)$	
η_1 η_2	11 19	-266,69 -259,62	555,38 557,23	7,79E-02		-343,16 -321,67	708,33 681,34	8,83E-07	
CP	np	ℓ	AIC	$P(>\chi^2)$	$\hat{\phi}$	ℓ	AIC	$P(>\chi^2)$	$\hat{\phi}$
η_1 η_2	12 20	-266,60 -259,33	557,20 558,65	6,85E-02	-6,75E-02 1,29E-01	-326,61 -315,64	677,21 671,29	5,06E-03	-8,17E-01 -5,18E-01
BN	np	ℓ	AIC	$P(>\chi^2)$	$\hat{ heta}$	ℓ	AIC	$P(>\chi^2)$	$\hat{ heta}$
η_1 η_2	12 20	-266,69 -259,62	557,37 559,23	7,82E-02	4,59E+03 1,03E+06	-326,54 -315,39	677,07 670,77	4,39E-03	1,42E+02 2,61E+02
QP	np	ℓ	AIC	$P(>\chi^2)$	$\hat{\sigma^2}$	ℓ	AIC	$P(>\chi^2)$	$\hat{\sigma^2}$
η_1 η_2	11 19	79,43 65,28		1,87E-01	1,28E+00 1,20E+00	167,71 124,72		3,00E-02	2,71E+00 2,29E+00

np, número de parâmetros, PO, Poisson, CP, COM-Poisson, BN, Binomial Negativo, QP, Quasi-Poisson.

Fonte: Elaborado pelo autor.



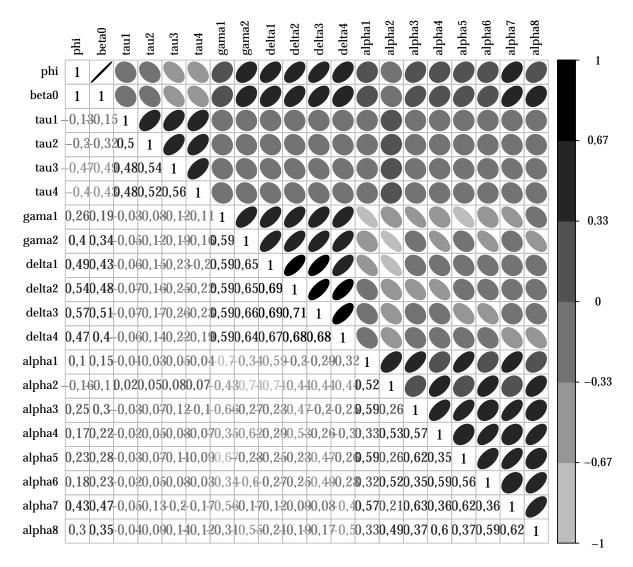
Fonte: Elaborado pelo autor.

Figura 24 – Perfis de log-verossimilhança para o parâmetro de precisão da COM-Poisson nos modelos para número de vagens viáveis por parcela (esquerda) e número grãos de soja por parcela (direira).

No fragmento direito da tabela 6 são apresentados os resultados para os modelos que ajustam os efeitos para o número de grãos por parcela. Neste caso há evidências de superdispersão, pois as estimativas dos parâmetros ϕ e σ^2 foram menor que zero e maior que 1 respectivamente. Os valores de AIC se apresentam menores e as avaliações da logverossimilhança no ponto máximo maiores para os modelos paramétricos alternativos

ao Poisson. Ainda a evidência sobre o efeito de interação para essa variável resposta é maior. Na figura 24 à direita, a verossimilhança perfilhada em ϕ é apresentada com indicação dos intervalos de confiança e estes não contém o valor zero.

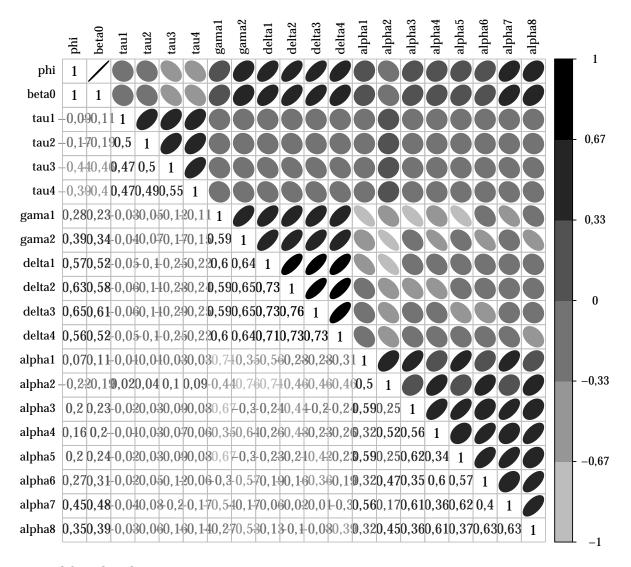
A visualização das covariâncias entre as estimativas dos parâmetros no modelo COM-Poisson para o número de vagens por parcela é feita na figura 25 e para o número de grãos por parcela na figura 26. Em ambos os casos a correlação entre os parâmetros de locação (β 's) e dispersão (ϕ) ganha destaque.



Fonte: Elaborado pelo autor.

Figura 25 – Imagem da matriz de correlação entre os parâmetros do modelo COM-Poisson ajustados ao número de vagens por parcela.

Na figura 27 são apresentadas as médias calculadas com intervalos de confiança 95% sob os modelos Poisson, COM-Poisson, Binomial-Negativo e Quasi-Poisson, considerando efeito de interação entre os níveis de umidade do solo e adubação potássica.



Fonte: Elaborado pelo autor.

Figura 26 – Imagem da matriz de correlação entre os parâmetros do modelo COM-Poisson ajustados ao número de grãos por parcela.

Tomou-se o efeito médio de bloco, uma vez que esse efeito aditivo não é de interesse prático.

Para a contagem do número de vagens, observa-se os intervalos com comprimento muito parecidos, ligeiramente menores para o caso COM-Poisson e Binomial Negativo. Para a contagem do número de grão por parcela, um caso superdisperso, percebe-se que o modelo Poisson nos leva a uma falsa precisão, uma vez que os intervalos são menores não por se ajustar melhor aos dados, mas sim por subestimar a variabilidade do processo. Para as formulações alternativas, obteve-se intervalos de confiança para média menores nos modelos paramétricos quando comparados com o semi-paramétrico Quasi-Poisson, isso é razoável, pois nos Quasi-Poisson somente a

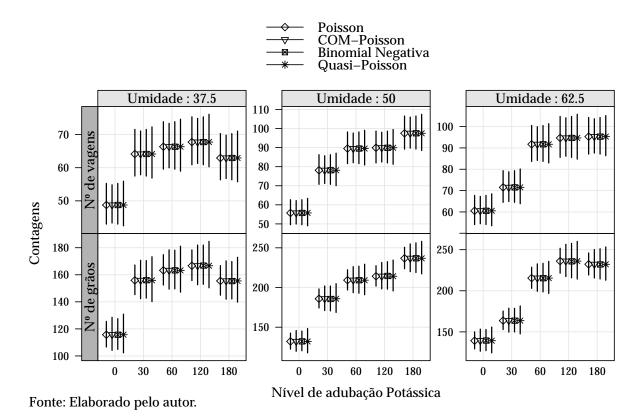


Figura 27 – Valores preditos com intervalos de confiança (95%) como função do nível de adubação com potássio e do percentual de umidade do solo para cada variável de interesse mensurada (número de vagens e número de grãos por parcela).

especificação de dois momentos é feita, enquanto que nos paramétricos especifica-se a distribuição completa, ganhando informação (ver equação 2.4). De forma geral os intervalos sob os modelos COM-Poisson e Binomial Negativa são maiores, porém fiéis a variabilidade inerente ao processo.

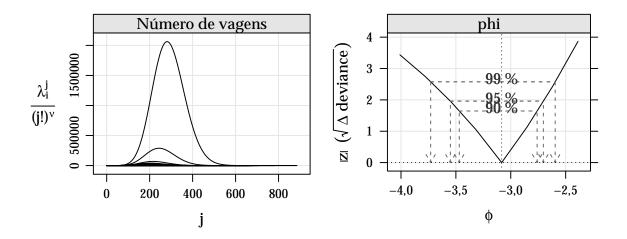
4.4 Análise de ninfas de mosca-branca em lavoura de soja

Neste experimento também há fortes indícios de superdispersão, conforme visto na seção 3.1.1.4. Assim os modelos Poisson, COM-Poisson, Binomial Negativo e Quasi-Poisson serão aplicados. A variável em estudo é a contagem da quantidade de ninfas de Mosca-branca nos folíolos de plantas de soja, ao longo dos dias nas diferentes cultivares. Como o experimento foi conduzido sob delineamento de blocos casualizados, os efeitos de bloco são considerados no modelo. As covariáveis serão tratadas como fator, assim como na aplicação anterior, com seis níveis para o número de dias decorridos a partir da primeira avaliação e quatro níveis para o fator cultivar de soja. Os preditores em comparação são:

Preditor 1: $\eta_1 = g(\mu_{ijk}) = \beta_0 + \tau_i + \gamma_j + \delta_k$

Preditor 1:
$$\eta_2 = g(\mu_{ijk}) = \beta_0 + \tau_i + \gamma_j + \delta_k + \alpha_{jk}$$

em que τ_i é o efeito do i-ésimo bloco, i=1: bloco II, 2: bloco III, 3: bloco IV e 4: V; γ_j o efeito da j-ésima cultivar , j=1: BRS 243 RR, 2: BRS 245 RR e 3: BRS 246 RR; δ_k o efeito do k-ésimo nível do número de dias após o início do experimento, k=8, 13, 22, 31 e 38 dias e α_{jk} o efeito da interação entre a j-ésima cultivar e o k-ésimo nível do número de dias após o início do experimento. A avaliação do efeito de interação é de interesse prático, pois informa se há um padrão distinto na quantidade de ninfas ao longo do tempo entre as cultivares. No modelo com interação, 27 parâmetros de locação a devem ser estimados.



Fonte: Elaborado pelo autor.

Figura 28 – Convergência das constantes de normalização para cada indivíduo (direita) e perfil de log-verossimilhança para o parâmetro extra da COM-Poisson (esquerda) no modelo para o número de ninfas de Mosca-branca.

Assim como na aplicação superdispersa apresentada na seção 4.3, nesse exemplo tem-se um cenário com contagens altas (variando entre 92 e 271) e ainda superdispersas (parâmetros ϕ estimados próximos à -3). Isso torna a convergência da função $Z(\lambda_i, \nu = \exp(\phi))$ demorada e o valor dessa constante, que normaliza a densidade, é altíssimo para a maioria das observações. Considerando o modelo com interação, podese visualizar os termos, que somados compõem a constante Z, para cada observação, à direira da figura 28. Para a observação 45 tem-se o maior valor calculado da constante Z, 3,785 × 108. Para obtenção deste valor 886 termos foram necessários, conforme exibido no eixo x do gráfico.

Em problemas com contagens altas e comportamento muito superdisperso a obtenção da constante Z pode se tornar proibitiva computacionalmente, devido à *overflow* (valores que ultrapassam o limite de capacidade de cálculo da máquina) e consequentemente o modelo COM-Poisson não se ajusta.

Nesse exemplo, os modelos COM-Poisson convergiram e seus resultados são exibidos na tabela 7 em conjunto com os resultados do ajuste dos modelos Poisson, Binomial Negativo e Quasi-Poisson. Todas as estimativas dos parâmetros extras nos modelos concorrentes ao Poisson, $\hat{\phi}$, $\hat{\theta}$ e $\hat{\sigma}^2$ indicam expressivamente a superdispersão os dados. Em benefício dos modelos alternativos ao Poisson tem-se todas as medidas apresentadas indicando uma substancial melhora de ajuste quando flexibilizado o modelo. Destaque para a magnitude dessas evidências, em que, por exemplo, o AIC obtido dos modelos alternativos é em torno de 0,47 vezes o obtido do Poisson.

Tabela 7 – Medidas de ajuste para avaliação e comparação entre preditores e modelos ajustados

Poisson	np	ℓ	AIC	$2(\text{diff }\ell)$	diff np	$P(>\chi^2)$	
Preditor 1	12	-922,98	1869,96				
Preditor 2	27	-879,23	1812,46	87,50	15	2,90E-12	
COM-Poisson	np	ℓ	AIC	$2(\text{diff }\ell)$	diff np	$P(>\chi^2)$	$\hat{\phi}$
Preditor 1	13	-410,44	846,89				-3,08
Preditor 2	28	<i>-</i> 407,15	870,30	6,59	15	9,68E-01	-2,95
Binomial Neg.	np	ℓ	AIC	$2(\text{diff }\ell)$	diff np	$P(>\chi^2)$	$\hat{ heta}$
Preditor 1	13	-406,16	838,31				3,44
Preditor 2	28	-400,55	857,10	11,21	15	7,38E-01	3,99
Quase-Poisson	np	deviance	AIC	F	diff np	P(>F)	$\hat{\sigma}^2$
Preditor 1	12	1371,32					17,03
Preditor 2	27	1283,82		0,31	15	9,93E-01	19,03

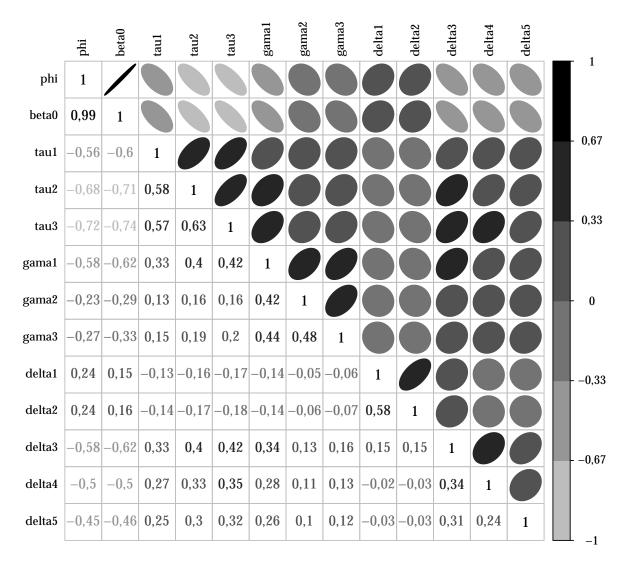
np, número de parâmetros, diff ℓ , diferença entre log-verossimilhanças, F, estatística F baseada nas quasi-deviances, diff np, diferença entre o np.

Fonte: Elaborado pelo autor.

Para tomada de decisão, observa-se que o modelo Poisson é claramente inadequado. Para avaliação dos preditores, na tabela 7, o modelo Poisson indica (com uma significância inferior a 1×10^{-10}) que há efeito de interação entre os dias decorridos da primeira avaliação e as cultivares ao passo que, nos modelos alternativos, esse efeito é marcadamente não significativo. Essa discordância se deve, conforme já discutido, ao fato de o modelo Poisson subestimar a variabilidade por sua restrição de equidispersão. Assim, com variâncias menores, qualquer efeito acrescido ao modelo passará por significativo.

Enfatizando a superdispersão indicada pelo modelo COM-Poisson e considerando o preditor de efeitos aditivos, tem-se o perfil de verossimilhança para o parâmetro ϕ apresentado na figura 28. Pode-se observar que os limites inferiores dos intervalos de confiança de 90, 95 e 99% estão muito distantes do valor 0, sob o qual os modelos Poisson e COM-Poisson são equivalentes. Outra característica desse gráfico é a leve

assimetria à esquerda, indicando que haverá imperfeições para inferências baseadas na aproximação quadrática da verossimilhança.

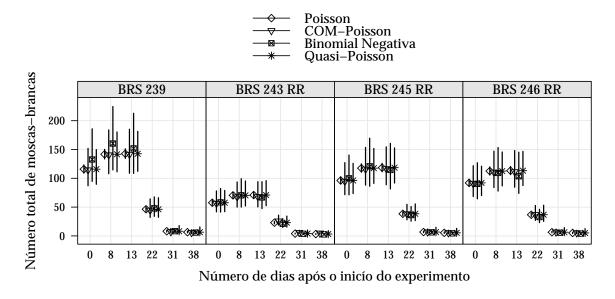


Fonte: Elaborado pelo autor.

Figura 29 – Imagem da matriz de correlação entre os parâmetros do modelo COM-Poisson.

As covariâncias entre os efeitos estimados pelo modelo COM-Poisson também são apresentadas, conforme descrição do preditor 1 na figura 29, sob a escala de correlação. Similarmente as análises anteriores observa-se a alta correlação entre $\hat{\phi}$ e os demais parâmetros de regressão. A soma dos valores absolutos das correlações observadas entre $\hat{\phi}$ e as demais estimativas é de 7,059 e a média 0,543.

As médias com intervalos de confiança calculadas para cada combinação dos níveis de dias após a primeira avaliação e cultivar de soja considerando os modelos Poisson, COM-Poisson, Binomial-Negativo e Quasi-Poisson, são apresentadas na figura



Fonte: Elaborado pelo autor.

Figura 30 – Valores preditos com intervalos de confiança (95%) em função das cultivares de soja e da data de avaliação da planta.

30. Para o efeito de bloco foi considerado o efeito médio, para uma correta comparação. Pode-se observar que o intervalo de confiança descrito pelo modelo Poisson é quase imperceptível quando comparados aos demais, mostrando novamente que seu uso é inadequado a esses dados. Já para as outras alternativas não tivemos um comportamento padrão em todas as cultivares. Os intervalos pelo modelos Quasi-Poisson e COM-Poisson foram muito similares em todos os casos e os intervalos pelo modelo Binomial Negativo mais amplos. Um fato interessante é que não necessariamente as estimativas pontuais da média desses modelos alternativos serão iguais, isso ocorre, por construção, somente para nos modelos Poisson e Quasi-Poisson, esse exemplo ilustra na prática a constatação desse fato. Para o modelo Binomial Negativo tivemos médias visivelmente superiores que os demais para a cultivar BRS 239. Para o modelo COM-Poisson as estimativas pontuais são visivelmente iguais as do modelo Poisson.

4.5 Análise de captura de peixes em um parque estadual

Nesse exemplo ilustra-se a análise de um estudo observacional em que aparentemente há uma quantidade excessiva de contagens nulas (veja a seção 3.1.1.5). O estudo tem por objetivo a modelagem do número de peixes capturados por grupos de visitantes em um Parque Estadual. As covariáveis mensuradas foram (np), o número de pessoas no grupo, (nc), o número de crianças e (ca) variável binária que indica a presença ou não de um campista no grupo.

Como já antecipado pela visualização e apresentação dos dados, modelos es-

truturados de forma convencional, que pressupõe apenas um processo estocástico na geração de dados, não se ajustaram adequadamente. A seguir a alternativa de inclusão de um efeito de barreira para acomodar a quantidade excessiva de valores zero é apresentada. Os modelos Poisson, Binomial Negativo e COM-Poisson sob esta estruturação são ajustados e comparados.

O número de peixes capturados é modelado em duas partes, as contagens nulas e as não nulas, conforme descrito na seção 2.4. Abaixo define-se os preditores considerados para as duas partes

Preditor 1:
$$g(\mu) = \beta_0 + \beta_1 ca + \beta_2 np$$

$$logit(\pi) = \gamma_0 + \gamma_1 ca + \gamma_2 np + \gamma_3 nc$$

$$g(\mu) = \beta_0 + \beta_1 ca + \beta_2 np + \beta_3 nc + \beta_4 (np \cdot nc)$$

$$logit(\pi) = \gamma_0 + \gamma_1 ca + \gamma_2 np + \gamma_3 nc + \gamma_4 (np \cdot nc)$$

sendo $g(\mu)$ e $logit(\pi)$ as funções de ligação que relacionam os preditores lineares com as médias dos modelos para contagens não nulas e contagens zero respectivamente. Os preditores lineares foram propostos de forma aninhada. No primeiro considera-se os efeitos aditivos de todas as covariáveis mensuradas para a parte das contagens nulas e efeitos aditivos do número de pessoas e de crianças para a parte das contagens não nulas. No segundo tem-se os efeitos aditivos de todas as covariáveis acrescido do efeito de interação entre o número de pessoas e de crianças para ambas as partes do modelo.

Tabela 8 – Medidas de ajuste para avaliação e comparação de preditores e modelos com componente de barreira ajustados

Poisson	np	ℓ	AIC	$2(diff \ell)$	diff np	$P(>\chi^2)$	
Preditor 1	7	-857,48	1728,96				
Preditor 2	10	-744,58	1509,17	225,79	3	1,12E-48	
Binomial Negativo	np	ℓ	AIC	$2(\text{diff }\ell)$	diff np	$P(>\chi^2)$	$\hat{ heta}$
Preditor 1	8	-399,79	815,58				0,20
Preditor 2	11	-393,72	809,44	12,14	3	6,91E-03	0,37
COM-Poisson	np	ℓ	AIC	$2(\text{diff }\ell)$	diff np	$P(>\chi^2)$	$\hat{\phi}$
Preditor 1	8	-409,85	835,71				-8,77
Preditor 2	11	-402,30	826,59	15,12	3	1,72E-03	-3,77

np, número de parâmetros, diff ℓ , diferença entre log-verossimilhanças, F, estatística F baseada nas quasi-deviances, diff np, diferença entre o np.

Fonte: Elaborado pelo autor.

Na tabela 8 as medidas de ajuste dos modelos Poisson, Binomial Negativo e COM-Poisson são apresentadas para comparação dos resultados. Observa-se pelas logverossimilhanças maximizadas que o modelo Poisson não se ajustou adequadamente

quando comparado aos demais. Isso se deve ao fato discutido na seção 2.4, que mesmo modelando os zeros pode-se ter diferentes níveis de dispersão para as contagens nulas. Nesse exemplo as contagens não nulas são superdispersas, conforme visto pelas estimativas dos parâmetros extras do modelo Binomial Negativo e COM-Poisson. Indicado pelos níveis descritivos dos TRV's aplicados nos modelos encaixados há evidências de que o modelo com efeitos de interação é distinto do modelo com efeitos aditivos definidos no preditor 1.

As estimativas dos parâmetros para cada especificação de modelos são exibidas na tabela 9. Observe, primeiramente, que as estimativas dos parâmetros γ_i , i=0,1,2,3,4 são idênticas, independentemente do modelo adotado. Esse resultado é esperado, pois na construção dos modelos com componente de barreira, a modelagem da parte que contempla os valores zero é realizada via distribuição Bernoulli com parâmetro $\pi = \text{logit}(Z\gamma)$. As diferenças entre os modelos ocorre na distribuição considerada para a parte das contagens não nulas.

Tabela 9 – Estimativas dos parâmetros e razões entre as estimativa e erro padrão para os três modelos em estudo

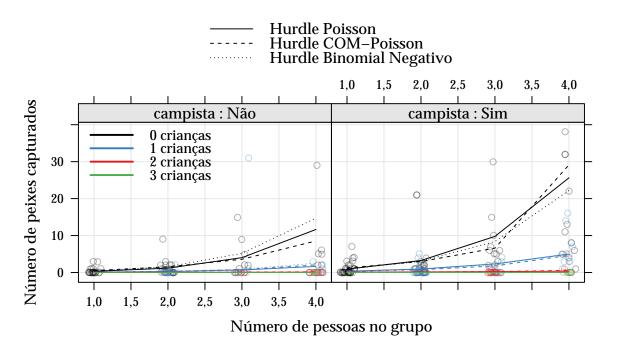
	Poiss	on	Binomial N	legativo	COM-Po	oisson
Parâmetro	Estimativa	Est/EP	Estimativa	Est/EP	Estimativa	Est/EP
φ, θ			0,37	-2,08	-3,77	-9,52
β_0	-1,01	-5,44	-1,75	-2,90	-0,62	-29,74
β_1	0,74	7,88	0,41	1,23	0,10	29,20
β_2	0,89	18,55	1,05	6,41	0,14	21,86
β_3	0,49	1,11	-0,06	-0,05	-0,33	-17,53
eta_4	-0,45	-3,69	-0,32	-0,90	0,04	33,41
γ_0	-2,58	-5,08	-2,58	-5,08	-2,59	-5,09
γ_1	0,98	3,00	0,98	3,00	1,00	3,04
γ_2	1,25	5,60	1,25	5,60	1,26	5,61
γ_3	-0,93	-1,05	-0,93	-1,05	-0,93	-1,06
γ_4	-0,41	-1,41	-0,41	-1,41	-0,41	-1,41

Fonte: Elaborado pelo autor.

Nos efeitos estimados para a parte da modelagem dos valores não nulos têm-se algumas diferenças consideráveis. Destaca-se que o valor das estimativas dos modelos Poisson e Binomial Negativo são comparáveis entre si, pois modelam a média da distribuição, mas não comparáveis com as estimativas do modelo COM-Poisson, pois este modela um parâmetro que não representa, diretamente, a média. Contudo, independente da distribuição o sinal dos efeitos deve ser o mesmo. Isso não ocorre nas estimativas dos parâmetros β_3 , positiva no modelo Poisson e negativa nos demais e β_4 , positiva no modelo COM-Poisson e negativa nos demais. Porém, esses efeitos não tem impacto significativo para definição dos parâmetros das distribuições, conforme visto

na figura 31 que exibe as médias calculadas com base nas três formulações. A seguir uma discussão sobre os valores apresentados para os erros padrão dessas estimativas é feita.

Calculando a magnitude desses efeitos quando escalonados pelo seu erro padrão, obtido pelo negativo do inverso da matriz hessiana, há diferenças substanciais. O modelo COM-Poisson indica erros padrões das estimativas muito menores que os apresentados no modelo Binomial Negativo. Sob investigações do problema, encontrou-se que este resultado se deve por inconsistências no procedimento numérico para determinação da matriz hessiana por diferenças finitas no modelo COM-Poisson. Portanto, os erros padrão sob o modelo COM-Poisson apresentados estão incorretos.



Fonte: Elaborado pelo autor.

Figura 31 – Valores preditos do número de peixes capturados considerando o número de crianças e pessoas no grupo e a presença de um campista.

Embora tenha-se constatado problemas nos algoritmos numéricos para determinar a curvatura da log-verossimilhança, as estimativas pontuais são coerentes com os demais modelos, conforme visto na figura 31 onde são apresentadas as médias calculadas com base nos três modelos estudados. Observa-se em todos os modelos a mesma tendência.

Com esse exemplo ilustra-se a extensão do modelo COM-Poisson para acomodar excesso de zeros e ressalta-se que as contagens não nulas analisadas são superdispersas. Para esses casos a distribuição Binomial Negativa se apresenta como principal alternativa. Porém, em casos que as contagens não nulas se mostram subdispersas

não há opções prontamente disponíveis para análise e o modelo COM-Poisson com componente de barreira, conforme apresentado, se torna uma abordagem atrativa.

4.6 Análise de dados de reprodução de nematoides em cultivares de feijoeiro

Nessa última aplicação apresentada no trabalho a extensão dos modelos de contagem para inclusão de efeitos aleatórios é ilustrada. Os modelos em competição são o Poisson e o COM-Poisson com efeitos aleatórios. O conjunto de dados se refere ao número de nematoides em cultivares medidas em soluções sol compostas da massa fresca de raizes diluídas em água, mensuradas em gramas· ml⁻¹ conforme apresentado na seção 3.1.1.6. Considera-se para os modelos em competição, os seguintes preditores:

```
Preditor 2: g(\mu) = \beta_0 + b_j
Preditor 2: g(\mu) = \beta_0 + \beta_1 \log(\text{sol})_i + b_j
```

em que $i=1,2,\cdots$, 94 (número de observações) e j varia nos níveis da cultivar de feijão ($j=A,B,C,\cdots$, S representando o efeito aleatório, realização de uma variável aleatória Normal de média 0 e variância σ^2 . Assim, nos modelos propostos têm-se a variabilidade entre as cultivares explicada por uma distribuição Normal e a variabilidade dentro das cultivares explicada pela relação média variância descrita pelo modelo considerado, Poisson ou COM-Poisson.

O ajuste dos modelos com a inclusão de efeitos aleatórios requer a solução de uma integral que, em geral, é resolvida numericamente. Isso torna o procedimento de ajuste computacionalmente intensivo e bastante suscetível a problemas numéricos. Para o ajuste dos modelos COM-Poisson de efeitos mistos algumas iterações do algoritmo de estimação propuseram valores para os parâmetros que resultaram em somas $Z(\lambda_i,\phi)$ que não puderam ser representados pela máquina, *overflow*. Porém, o algoritmo dispõe de procedimentos que evitam sua interrupção, propondo novos valores mesmo quando a função objetivo não puder ser calculada, alcançando o máximo da log-verossimilhança. Para o modelo Poisson de efeito aleatório utilizou-se das programações em R providas pelo pacote 1me4 (BATES et al., 2015), que trabalham com matrizes esparsas para os efeitos aleatórios e otimização em linguagem de baixo nível, minimizando os problemas numéricos.

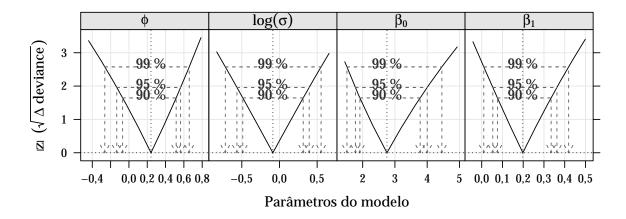
Os resultados do ajuste para avaliação e comparação dos modelos são apresentados na tabela 10. Os valores na tabela indicam que os modelos Poisson e COM-Poisson se ajustaram de forma equivalente, os valores da log-verossimilhança foram muito próximos. Essa equivalência também é apontada pelos AIC's, que foram maiores para nos modelos COM-Poisson e pelos níveis descritivos dos TRV's realizados sob a hipótese

 $H_0: \phi=0$, indicando que a adoção de um modelo com um parâmetro adicional não é justificado pelo pequeno acréscimo na log-verossimilhança. Com relação ao efeito do logaritmo da solução de massa fresca de raiz, há evidências apontando um efeito significativo para explicação do número de nematoides.

Tabela 10 – Medidas de ajuste para avaliação e comparação entre preditores e modelos ajustados

Poisson	np	ℓ	AIC	2(diff ℓ)	diff np	$P(>\chi^2)$		
Preditor 1 Preditor 2		-237,20 -234,66	478,40 475,32	5,07	1	2,43E-02		
COM-Poisson	np	ℓ	AIC	$2(\text{diff }\ell)$	diff np	$P(>\chi^2)$	$\hat{\phi}$	$P(>\chi^2)$

np, número de parâmetros, diff ℓ , diferença entre log-verossimilhanças, diff np, diferença entre o np. Fonte: Elaborado pelo autor.



Fonte: Elaborado pelo autor.

Figura 32 – Perfis de verossimilhança dos parâmetros estimados no modelo COM-Poisson Misto.

Permanecendo com o segundo preditor, com o efeito do logaritmo da solução, as estimativas dos parâmetros do modelo são apresentadas na tabela 11 em conjunto com seu erro padrão, calculado sob aproximação quadrática da verossimilhança, ou seja via inversão da matriz hessiana. Novamente, os resultados entre os modelos são similares. Lembre-se que, desta tabela o único resultado comparável diretamente é a razão entre estimativa e erro padrão do parâmetro β_1 . O parâmetro σ é a variância da distribuição dos efeitos aleatórios, que no modelo Poisson são somados aos efeitos fixos para composição de μ e na COM-Poisson para composição de λ . Outro resultado interessante dessa tabela é a estimativa do parâmetro ϕ da COM-Poisson, que positiva indica uma subdispersão moderada nesse conjunto de dados. Uma vantagem do modelo misto

COM-Poisson é que pode-se distinguir a variabilidade da contagem com a variabilidade do efeito do grupo no experimento. Nesse exemplo tem-se uma variabilidade do efeito aleatório maior, σ estimado no caso COM-Poisson maior que no caso Poisson, porém essa variabilidade extra capturada pelo efeito aleatório é compensada pela subdispersão capturada pelo parâmetro ϕ .

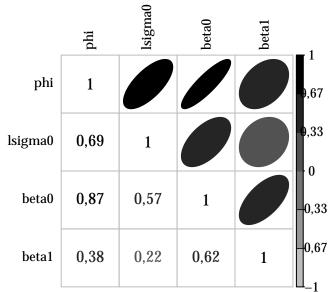
Tabela 11 – Estimativas dos parâmetros e r	azões entre as estimativa e erro padrão para
os três modelos em estudo	

		Poisson		COM-Poisson		
Parâmetro	Estimativa	E. Padrão	Est/EP	Estimativa	E. Padrão	Est/EP
σ	0,75			0,93		
eta_0	1,62	0,19	8,50	2,08	0,45	4,59
eta_1	1,27	0,56	2,29	1,58	0,68	2,33
ϕ				0,23	0,18	1,33

Como resultados complementares a tabela 11, tem-se os perfis de verossimilhança com intervalos de confianças de níveis 90, 95 e 99% apresentados na figura 32. Observa-se um comportamento razoavelmente simétrico para todos os parâmetros, apenas com uma assimetria levemente destacada para o parâmetro β_0 . Isso traz mais segurança na interpretação dos resultados baseados na aproximação quadrática da verossimilhança, que são de fácil obtenção pois só envolvem inversão de matrizes. No perfil de verossimilhança para o parâmetro ϕ , há mais uma evidência da equivalência entre os modelos Poisson e COM-Poisson, pois os intervalos contém o valor 0.

Conforme já observado anteriormente, no modelo COM-Poisson misto os parâmetros ϕ , da distribuição considerada para a variável de contagem condicional aos efeitos aleatórios e as covariáveis e σ , da distribuição considerada para os efeitos aleatórios são conjuntamente responsáveis pela explicação da variabilidade do processo em estudo. Na figura 33 apresentados as covariâncias entre os parâmetros do modelo, na escala de correlação, a fim de verificar, principalmente, a correlação entre σ e ϕ . Observa-se que, conforme esperado, estes parâmetros apresentam uma forte correlação e ainda que esta é positiva, pois as contagens são superdispersas, ainda que não de forma acentuada. Nota-se também que a característica de não ortogonalidade entre os parâmetros de locação e ϕ se mantém com a inclusão de efeitos aleatórios.

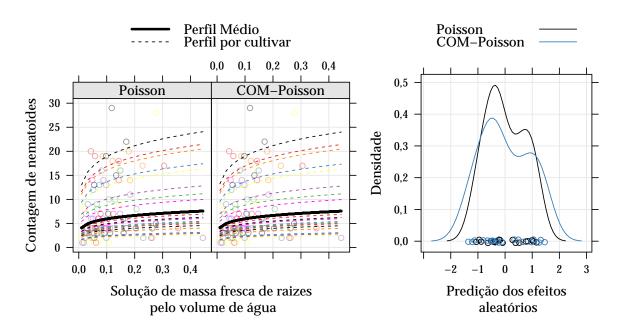
Na figura 34 são apresentados as predições do efeito aleatório em cada modelo, à direita e as contagem preditas para cada cultivar e para o comportamento médio, à esquerda. A distribuição empírica dos efeitos aleatórios, gráfico à direita, está de acordo com os parâmetros estimados para σ , vistos na tabela 11. Têm-se a ordenação dos efeitos aleatórios idêntica em ambos os modelos, porém valores mais dispersos no caso COM-Poisson. Devido ao parâmetro adicional ϕ do modelo COM-Poisson, que indica subdispersão, tem-se os valores preditos por esse modelo muito similares aos



Fonte: Elaborado pelo autor.

Figura 33 – Imagem da matriz de correlação entre os parâmetros do modelo COM-Poisson.

preditos pelo modelo Poisson, conforme observa-se no gráfico à direita da figura 34. A soma das diferenças ao quadrado, entre valores preditos pelos dois modelos foi de 1,17, o que mostra que ambos os modelos levam ao mesmo resultado.



Fonte: Elaborado pelo autor.

Figura 34 – Perfis de verossimilhança dos parâmetros estimados no modelo COM-Poisson Misto.

Nessa aplicação ilustra-se a extensão do modelo COM-Poisson para inclusão de efeitos aleatórios. Nesse caso a análise se deu a um experimento em que as contagens, condicionadas aos efeitos aleatórios, se apresentaram de forma equidispersa, indicada pelo modelo COM-Poisson, e os resultados entre os modelos COM-Poisson e Poisson foram equivalentes.

4.7 Discussões

Nos quatro primeiros conjuntos de dados, em que modelou-se as contagens via modelos de regressão de efeitos fixos, observou-se resultados dos modelos COM-Poisson equivalentes a abordagem semi-paramétrica via quasi-verossimilhança, quanto a significância dos efeitos e predição com bandas de confiança. Porém ressalta-se que na abordagem por quasi-verossimilhança, com a especificação de apenas dois momentos, i) não se pode representar a distribuição de probabilidades da variável em estudo, ii) a informação a respeito da média é igual ou inferior a uma abordagem totalmente paramétrica e iii) extensões como a modelagem de excesso de zeros e modelagem do parâmetro de dispersão não são imediatas. Nos casos de superdispersão explorou-se também os resultados dos modelos baseados na distribuição Binomial Negativa e nessa abordagem tem-se o inconveniente de somente a característica de superdispersão ser contemplada. Nos estudos de caso os modelos Binomial Negativo proporcionaram resultados, com relação a significância dos efeitos, equivalentes ao COM-Poisson e Quasi-Poisson. Porém, em um dos estudos de caso com acentuada superdispersão, os valores preditos pontuais e intervalares nessa abordagem diferiram dos modelos COM-Poisson e Quasi-Poisson, isso devido a forma da relação média e variância dessa distribuição, figura 4.

Nas extensões propostas para o modelo COM-Poisson obteve-se resultados satisfatórios. No caso da inclusão de um componente de barreira para modelagem de excesso de zeros, os resultados dos testes de razão de verossimilhanças para testar a significância dos efeitos foram equivalentes ao modelo Hurdle Binomial Negativo assim como as estimativas pontuais dos valores preditos. Ainda nessa aplicação, não foi possível a obtenção dos erros padrão das estimativas dos efeitos, baseados na matriz hessiana, devido a problemas numéricos na determinação dessa matriz. Para o caso estendido do modelo COM-Poisson em que acomoda-se efeitos aleatórios, os procedimentos computacionalmente intensivos que são empregados no algoritmo de estimação ganham destaque. A aplicação se deu a um experimento que apresentou contagens com um grau não significativo de subdispersão. Nessa aplicação os modelos em competição foram o Poisson e o COM-Poisson de efeitos mistos e todos os resultados em questões inferenciais foram equivalentes em ambos os modelos, com poder de teste maior para o modelo COM-Poisson.

4.7. Discussões 73

Nas aplicações, em geral, pode-se notar características que permearam a todos os modelos baseados na distribuição COM-Poisson. A primeira delas, e talvez a mais difícil de se contornar, é a determinação da constante de normalização, pois essa depende do parâmetro que está associado a um preditor linear assim deve-se calcular n constantes a cada iteração do algoritmo de estimação. Em casos de contagens altas e superdispersão o cálculo dessa constante é extremamente demorado. Outra característica que se manisfestou em todas as aplicações foi a não ortogonalidade entre os parâmetros de regressão e o parâmetro adicional ϕ , observada pelas correlações calculadas a partir da matriz hessiana. O que torna as inferências dependentes. Em pesquisas não relatadas nesse trabalho verificou-se que a reparametrização do parâmetro λ , adotando a aproximação para média contorna essa característica com o preço de se ter uma distribuição aproximada. Nas aplicações explorou-se também os perfis de verossimilhança para o parâmetro ϕ da COM-Poisson e o comportamento aproximadamente simétrico em todos casos induz que aproximações quadráticas da verossimilhança podem ter desempenhos satisfatórios.

5 Considerações Finais

Os objetivos nesse trabalho foram a exploração, extensão e aplicação da distribuição COM-Poisson na análise de dados de contagem cujo foram atendidos com a apresentação de seis aplicações dos modelos COM-Poisson à conjuntos de dados reais que exibem equidispersão, subdispersão, superdispersão, contagens altas, excesso de zeros e efeito aleatório, mostrando a flexibilidade do modelo COM-Poisson.

Das análises realizadas destaca-se a característica restritiva do modelo Poisson, que na maioria dos casos não se ajustou adequadamente devido a suposição de equidispersão. Para os modelos de regressão de efeitos fixos, os resultados obtidos com as abordagens via modelo COM-Poisson, Quasi-Poisson e Binomial Negativo (para os casos de superdispersão) foram bastante similares quanto a significância dos efeitos e predição com bandas de confiança. Resultados satisfatórios também foram obtidos para nos modelos COM-Poisson para modelagem de excesso de zeros e inclusão de efeitos aleatórios. Nessas extensões, há dificuldade computacional de ajuste dos modelos, principalmente devido ao cálculo das constantes de normalização, que mesmo nos modelos de efeitos fixos ainda são problemáticas.

Em todas as aplicações observou-se a não ortonalidade empírica na matriz hessiana, o que se mostra como característica da distribuição. Outra característica observada na análise de dados é a simetria nos perfis de verossimilhança para o parâmetro ϕ , indicando que aproximações quadráticas da verossimilhança podem ter bons desempenhos.

De forma geral, sugere-se a aplicação dos modelos COM-Poisson na análise de dados de contagem, pois devido a sua flexibilidade, seus resultados se equivalem a abordagem semi-paramétrica via quasi-verossimilhança, porém com todos os benefícios da inferência totalmente paramétrica.

Dado o escopo do trabalho foram vários os tópicos levantados para pesquisas futuras. Estudo de reparametrizações que tornem os parâmetros λ e ν ortogonais no modelo COM-Poisson podem ser de grande valia, pois tornaram as inferências entre eles independentes, além de possivelmente permitir a fatoração da verossimilhança com estimação concentrada. Para acelerar o algoritmo de estimação aproximações da constante normalização podem resultar em ajustes satisfatórios. Estudos de simulação para verificar a robustez do modelo à má especificação da distribuição da variável resposta. Implementação da modelagem de excesso de zeros via mistura de distribuições. Inclusão de efeitos aleatórios dependentes no modelo misto COM-Poisson. São algumas das muitas possibilidades para pesquisa envolvendo dados de contagem subdispersos ou superdispersos modelados com a distribuição COM-Poisson.

REFERÊNCIAS

BATES, D. M. et al. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, v. 67, p. 1–48, 2015. Disponível em: http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf. Citado 2 vezes nas páginas 32 e 68.

BORGES, P. *Novos modelos de sobrevivência com fração de cura baseados no processo da carcinogênese*. Tese (Doutorado) – Universidade Federal de São Carlos, 2012. Citado na página 18.

CONWAY, R. W.; MAXWELL, W. L. A queuing model with state dependent service rates. *Journal of Industrial Engineering*, v. 12, p. 132–136, 1962. Citado 2 vezes nas páginas 17 e 25.

FERREIRA, D. F. Estatística Multivariada. Editora UFLA, 2011. Citado na página 51.

HILBE, J. M. Modeling Count Data., 2014. 300 p. Citado 2 vezes nas páginas 19 e 30.

KING, G. Variance specification in event count models: from restrictive assumptions to a generalized estimator. *American Journal of Political Science*, v. 33, n. 3, p. 762–784, Disponível em: http://www.jstor.org/stable/2111071. Citado na página 15.

KOKONENDJI, C. C. Over- and Underdisperson Models. In: *Methods and Applications of Statistics in Clinical Trials: Planning, Analysis, and Inferential Methods.*, 2014. p. 506–526. Disponível em: https://lmb.univ-fcomte.fr/IMG/pdf/ch30{_}kokonendji2014. Citado na página 19.

LAMBERT, D. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, v. 34, n. 1, p. 1, feb 1992. Disponível em: http://www.jstor.org/stable/1269547?origin=crossref. Citado 2 vezes nas páginas 20 e 30.

LORD, D.; GEEDIPALLY, S. R.; GUIKEMA, S. D. Extension of the application of conway-maxwell-poisson models: Analyzing traffic crash data exhibiting underdispersion. *Risk Analysis*, v. 30, n. 8, p. 1268–1276, 2010. Citado na página 19.

MARTELLI, T. et al. *Influência do ataque de mosca-branca Bemisia tabaci Biotipo B, nos índices de produtividade do algodoeiro*. Uberlândia- MG: XXII Congresso Brasileiro de Entomologia, 2008. Citado na página 36.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, v. 135, p. 370–384, 1972. Citado 2 vezes nas páginas 15 e 22.

NOCEDAL, J.; WRIGHT, S. J. *Numerical optimization*. Springer, 1995. 636 p. Disponível em: https://books.google.com/books?id=epc5fX0lqRIC{&}pgi. Citado na página 45.

PAULA, G. A. *Modelos de regressão com apoio computacional*. IME-USP São Paulo, 2013. Disponível em: https://www.ime.usp.br/{~}giapaula/textoregressao.h>. Citado 2 vezes nas páginas 16 e 23.

RIBEIRO, A. M. T. *Distribuição COM-Poisson na análise de dados de experimentos de quimioprevenção do câncer em animais*. Dissertação (Mestrado) — Universidade Federal de São Carlos, 2012. Citado 2 vezes nas páginas 18 e 31.

RIBEIRO JR, P. J. et al. Métodos computacionais para inferência com aplicações em R. In: 20° *Simpósio Nacional de Probabilidade e Estatística.*, 2012. p. 282. Disponível em: http://leg.ufpr.br/doku.php/cursos:mcie. Citado 2 vezes nas páginas 16 e 44.

RIDOUT, M.; DEMETRIO, C. G.; HINDE, J. Models for count data with many zeros. *International Biometric Conference*, n. December, p. 1–13, 1998. Citado 2 vezes nas páginas 20 e 30.

SELLERS, K. F.; RAIM, A. A flexible zero-inflated model to address data dispersion. *Computational Statistics & Data Analysis*, Elsevier B.V., v. 99, p. 68–80, jul 2016. Disponível em: http://dx.doi.org/10.1016/j.csda.2016.01.007http://linkinghub.elsevier.com/retrieve/pii/S0167947316000165. Citado na página 18.

SELLERS, K. F.; SHMUELI, G. A flexible regression model for count data. *Annals of Applied Statistics*, v. 4, n. 2, p. 943–961, 2010. Citado 3 vezes nas páginas 19, 26 e 27.

SERAFIM, M. E. et al. Umidade do solo e doses de potássio na cultura da soja. *Revista Ciência Agronômica*, v. 43, n. 2, p. 222–227, jun 2012. Disponível em: http://www.scielo.br/scielo.php?script=sci| _arttext{&pid=S1806-66902012000200003{&lng=pt{&nrm>}}. Citado na página 38.

SHMUELI, G. et al. A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, v. 54, n. 1, p. 127–142, 2005. Citado 3 vezes nas páginas 17, 26 e 27.

SILVA, A. M. et al. Impacto de diferentes níveis de desfolha artificial nos estádios fenológicos do algodoeiro. *Revista de Ciências Agrárias*, v. 35, n. 1, p. 163–172, 2012. Disponível em: <a href="http://www.cabdirect.org/abstracts/20123299470.html;jsessionid="http://www.cabdirect.org/abstracts/20123299470.html;jsessionid="http://www.cabdirect.org/abstracts/20123299470.html;jsessionid="http://www.cabdirect.org/abstracts/20123299470.html;jsessionid="http://www.cabdirect.org/abstracts/20123299470.html;jsessionid="http://www.cabdirect.org/abstracts/20123299470.html;jsessionid="http://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstracts/20123299470.html;jsessionid="https://www.cabdirect.org/abstract

SUEKANE, R. *DISTRIBUIÇÃO ESPACIAL E DANO DE MOSCA-BRANCA Bemisia tabaci* (GENNADIUS, 1889) BIÓTIPO B NA SOJA. Dissertação (Mestrado) – Universidade Federal da Grande Dourados, 2011. Citado na página 39.

UCLA, S. C. G. *Data Analysis Examples*. 2015. Disponível em: http://www.ats.ucla.edu/stat/dae/. Citado na página 39.

WEDDERBURN, R. W. M. Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*, v. 61, n. 3, p. 439, 1974. Disponível em: http://www.jstor.org/stable/2334725?origin=crossref. Citado na página 22.

WINKELMANN, R. Duration Dependence and Dispersion in Count-Data Models. *Journal of Business & Economic Statistics*, v. 13, n. 4, p. 467–474, oct 1995. Disponível em: http://www.tandfonline.com/doi/abs/10.1080/07350015.1995.10524620. Citado 2 vezes nas páginas 16 e 21.

REFERÊNCIAS 79

WINKELMANN, R. *Econometric Analysis of Count Data*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. 342 p. Disponível em: http://link.springer.com/10.1007/978-3-540-78389-3. Citado na página 19.

WINKELMANN, R.; ZIMMERMANN, K. F. Count data models for demographic data. 1994. 205–221, 223 p. Citado na página 16.

ZEILEIS, A.; KLEIBER, C.; JACKMAN, S. Regression Models for Count Data in R. *Journal Of Statistical Software*, v. 27, n. 8, p. 1076–84, 2007. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed/21518631. Citado na página 31.

ZEVIANI, W. M. et al. The Gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics*, n. October, p. 1–11, 2014. Disponível em: http://dx.doi.org/10.1080/02664763.2014.922168>. Citado 5 vezes nas páginas 19, 21, 35, 47 e 48.



APÊNDICE A – Programas R

Todos os resultados apresentados são realizados com o *software* R, cujo códigos para ajuste dos modelos COM-Poisson de efeito fixo, aleatório e com componente de barreira foram disponibilizados em formato de pacote no endereço <github.com/jreduardo/tccPackage>. Nesse apêndice são apresentados os códigos, que utilizam as funções do pacote, para produzir os resultados da seção 4.2 (modelos de regressão de efeitos fixos). Todavia, os códigos que produzem os demais resultados apresentados no trabalho podem ser visualizados no complemento online

```
##-----
## Instalando o pacote tccPackage, elaborado no trabalho
library(devtools)
install_git("git@github.com: JrEduardo/tccPackage.git")
## Análise de dados apresentados na seção ... (v.a. número de nós)
## Carrega o pacote no workspace
library(tccPackage)
## Dados
data(cottonBolls2)
help(cottonBolls2)
## Preditores considerados
f1 <- nnos ~ 1
f2 <- nnos ~ dexp
f3 <- nnos \sim dexp + I(dexp^2)
## Ajustando os modelos Poisson
m1P.nnos <- glm(f1, data = cottonBolls2, family = poisson)
m2P.nnos <- glm(f2, data = cottonBolls2, family = poisson)
m3P.nnos <- glm(f3, data = cottonBolls2, family = poisson)
## Ajustando os modelos Quasi-Poisson
m1Q.nnos <- glm(f1, data = cottonBolls2, family = quasipoisson)</pre>
```

```
m2Q.nnos \leftarrow glm(f2, data = cottonBolls2, family = quasipoisson)
m3Q.nnos <- glm(f3, data = cottonBolls2, family = quasipoisson)
## Ajustando os modelos COM-Poisson
m1C.nnos <- cmp(f1, data = cottonBolls2, sumto = 30)
m2C.nnos <- cmp(f2, data = cottonBolls2, sumto = 30)
m3C.nnos <- cmp(f3, data = cottonBolls2, sumto = 30)
## TRV's entre modelos encaixados
anova(m1P.nnos, m2P.nnos, m3P.nnos, test = "Chisq")
anova(m1Q.nnos, m2Q.nnos, m3Q.nnos, test = "F")
anova(m1C.nnos, m2C.nnos, m3C.nnos)
## Estimativas e testes de Wald
summary(m3P.nnos)
summary(m3Q.nnos)
summary(m3C.nnos)
##-----
## Testando H0: phi = 0
cmptest(m1C.nnos, m2C.nnos, m3C.nnos)
## Matrix de variância e covariância da COM-Poisson
V <- vcov(m3C.nnos); V</pre>
cov2cor(V)
## Perfis de versossimilhança
prof <- profile(m3C.nnos)</pre>
plot(prof); confint(prof)
## Valores preditos
da <- data.frame(dexp = 0:5)</pre>
predict(m3C.nnos, da)
predict(m3C.nnos, da, interval = "confidence")
```