UNIVERSIDADE DE SÃO PAULO
ESCOLA SUPERIOR DE AGRICULTURA “LUIZ DE QUEIROZ”
PLANO DE DISSERTAÇÃO

PROGRAMA DE PÓS GRADUAÇÃO EM
“ESTATÍSTICA E EXPERIMENTAÇÃO AGRONÔMICA”

CURSO DE MESTRADO

EDUARDO ELIAS RIBEIRO JUNIOR

ORIENTADORA: CLARICE GARCIA BORGES DEMÉTRIO

“MODELOS FLEXÍVEIS PARA ANÁLISE DE DADOS DE CONTAGEM”

PIRACICABA
ESTADO DE SÃO PAULO – BRASIL
FEVEREIRO / 2018

Sumário

1 Introdução

A classe dos modelos lineares generalizados (GLM) foi introduzida por Nelder & Wedderburn (1972) contemplando modelos para análise de dados normais e não normais em uma mesma teoria. Wedderburn (1974) generalizou essa classe para os modelos de quase-verossimilhança, em que há apenas a suposição de primeiro e segundo momentos, ou seja, descreve-se apenas a relação entre a média e a variância da variável resposta. Embora a classe estendida para quase-verossimilhança seja mais flexível do que os GLM's, na maioria das situações não é possível recuperar a distribuição da variável resposta (Paula 2013).

Os modelos lineares generalizados podem ser ajustados por um algoritmo Newton-escore eficiente e há muitos softwares estatísticos que implementam facilidades para isso. Além disso, há excelentes contribuições na literatura da área como McCullagh & Nelder (1989), Venables & Ripley (2002) e Dobson & Barnett (2008). Assim, os modelos lineares generalizados se tornaram métodos proeminentes na análise de dados em estatística aplicada.

Apesar da flexibilidade dos modelos lineares generalizados, a relação média–variância, determinada pela função de variância $V(\mu_i)$, pode ser bastante restritiva, principalmente nos casos de dados binomiais, de contagens e de tempo até o evento (Molenberghs et al. 2010, 2017). Para contagens, por exemplo, o modelo Poisson tem função de variância $V(\mu_i) = \mu_i$, isto é, a esperança e a variância da distribuição são iguais para a i-ésima observação, característica conhecida como equidispersão. Porém, na prática, os dados podem apresentar características de superdispersão (média $<$ variância) e subdispersão (média $>$ variância), que tornam o modelo Poisson inadequado. Nesses casos, os coeficientes ainda podem ser estimados consistentemente, porém os erros padrões das estimativas são incorretos (Winkelmann & Zimmermann 1994, Hinde & Demétrio 1998). Em particular, um modelo Poisson ajustado a contagens superdispersas leva à subestimação dos erros padrões das estimativas; e para contagens subdispersas os erros padrões são superestimados.

O caso mais comum de falha da suposição de equidispersão e, consequentemente com um maior número de abordagens possíveis, é a superdispersão. Hinde & Demétrio (1998) discutem diversas razões que podem levar à variabilidade extra Poisson. Dentre elas, citam-se amostragem por aglomerados, heterogeneidade das unidades amostras, dados em nível agregado e correlação entre observações. 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 das causas da subdispersão pode ser a violação do processo Poisson, em que o tempo entre eventos pode não ser mais exponencialmente distribuído; esse processo motiva a classe dos modelos de dependência de duração (Winkelmann 1995). Outra possível causa da subdispersão está relacionada à obtenção de estatísticas de ordem da variável resposta, por exemplo tomar o máximo das contagens observadas (Steutel & Thiemann 1989).

Na Figura 1, são apresentados diferentes tipos de processos pontuais para ilustrar o processo Poisson em um espaço bidimensional. Cada ponto representa a ocorrência de um evento e cada quadrado, delimitado pelas linhas pontilhadas, representa a unidade (ou domínio) na qual se conta o número de eventos (como variável aleatória). A Figura 1(a), representa a situação de dados de contagem equidispersos. As ocorrências dos eventos se dispõem aleatoriamente. Na Figura 1(b), 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 desse padrão se dá pelo processo de contágio (e.g. contagem de casos de uma doença contagiosa, contagem de frutos apodrecidos). Na Figura 1(c), 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 da superdispersão, uma causa interpretável 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) ou ainda a competição (e.g. contagem de animais territoriais ou que disputam por território).

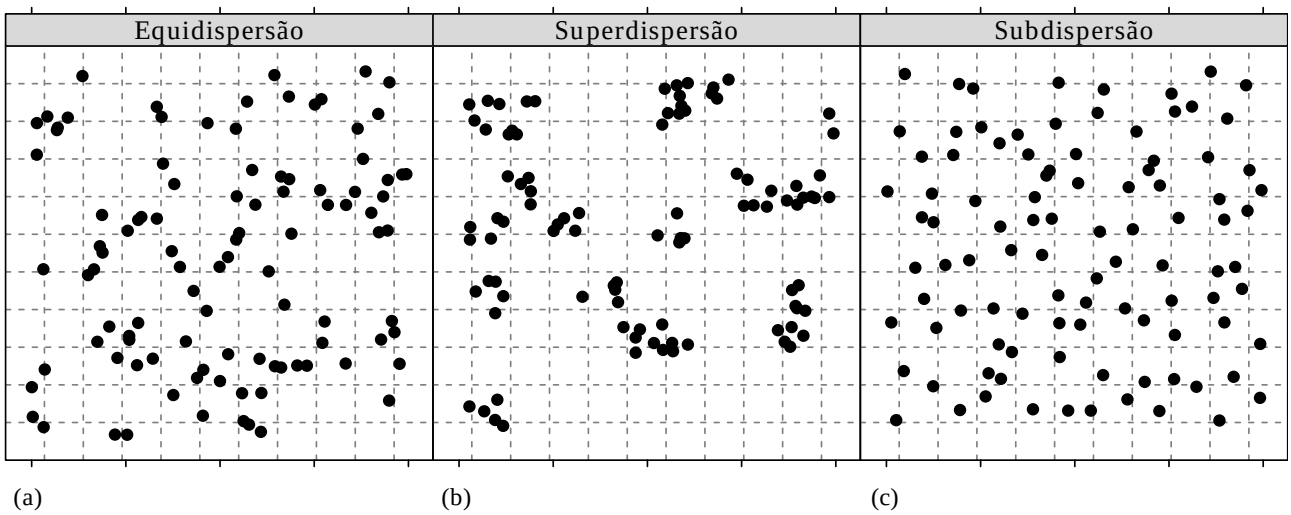


Figura 1 – Ilustração de diferentes tipos de processos pontuais. Da direita para esquerda têm-se processos sob padrões aleatório, aglomerado e uniforme.

As alternativas paramétricas para análise de dados na forma de contagens não equidipersas estão, geralmente, relacionadas às causas da não equidispersão. Para superdispersão, destacam-se os modelos mistos, que incluem efeitos aleatórios em nível de observação, considerando a heterogeneidade não observada. Um exemplo bem conhecido dessa prática é o modelo Poisson com efeitos aleatórios gama, que resulta no modelo binomial negativo. Porém, outras escolhas para a distribuição dos efeitos aleatórios podem ser tomadas como, por exemplo, o modelo Poisson-Tweedie (Bonat et al. 2018) e seus casos particulares Poisson-inversa Gaussiana (PIG) e Neyman-Type A assumem que os efeitos aleatórios são Tweedie, Gaussiano inverso e Poisson distribuídos, respectivamente.

Além dos modelos mistos, alguns modelos alternativos que modelam sub e superdispersão têm merecido atenção da comunidade estatística. Dentre eles destacam-se: o modelo *Gamma-Count* (Winkelmann 1995, Zeviani et al. 2014), que considera a distribuição gama para o tempo entre eventos; o modelo COM-Poisson (Shmueli et al. 2005, Sellers & Shmueli 2010), em que a razão de recorrência das probabilidades é não linear por meio da adição de um parâmetro de dispersão; e o modelo Poisson generalizado (Consul & Famoye 1992, Zamani & Ismail 2012), uma generalização biparamétrica do modelo Poisson.

Embora as distribuições citadas sejam flexíveis para lidar com diferentes níveis de dispersão, há situações em que o delineamento do experimento sugere uma estrutura de covariância entre observações induzida por um processo hierárquico de casualização ou amostragem e portanto, devem ser consideradas no modelo. São casos assim os experimentos em parcelas subdivididas e experimentos com medidas repetidas ou longitudinais. Tais estruturas estabelecem modelos com efeitos não observáveis e isso pode ser incorporado no modelo de regressão com a inclusão de efeitos aleatórios em nível de grupos experimentais; tais modelos são denominados modelos lineares generalizados mistos (GLMM) (Molenberghs & Verbeke 2005).

Para dados de contagem hierárquicos, o modelo GLMM Poisson assume que as contagens são distribuídas condicionalmente como Poisson e que os efeitos aleatórios são normalmente distribuídos. Embora esse modelo se adeque a diversas situações, casos em que as contagens condicionais não são equidispersas são pouco explorados na literatura. Molenberghs et al. (2007, 2010, 2017) apresentam uma nova família de modelos que consideram efeitos aleatórios normais, para contemplar a estrutura hierárquica e conjugados, para as contagens condicionalmente superdispersas. No entanto, modelos hierárquicos mais flexíveis, capazes de modelar diferentes níveis de dispersão das contagens condicionais aos efeitos aleatórios, não são consolidadas na literatura e são escassas suas aplicações.

2 Objetivos

2.1 Objetivo Geral

Contribuir para a área de análise de dados na forma de contagens, explorando distribuições probabilísticas que podem ser adotadas na estrutura aleatória de modelos de regressão estendidos para análise de dados hierárquicos e heterogêneos.

2.2 Objetivos Específicos

Como objetivos específicos, têm-se:

(i) Propor uma nova reparametrização para o modelo COM-Poisson, avaliar a flexibilidade do modelo, as propriedades dos estimadores e apresentar aplicações;

(ii) Revisar distribuições capazes de modelar diferentes níveis de dispersão, destacando suas diferenças usando-se estudos de caso e dados simulados;

(iii) Propor e avaliar as propriedades de um modelo linear generalizado duplo, em que o parâmetro de dispersão das distribuições para dados na forma de contagens também é modelado com covariáveis; e

(iv) Propor e avaliar as propriedades de um modelo linear generalizado misto, em que a distribuição para a variável aleatória de contagem, condicional aos efeitos aleatórios, é flexível para lidar com diferentes níveis de dispersão.

Todos os conjuntos de dados e as implementações computacionais utilizadas, serão disponibilizadas para que toda a pesquisa seja facilmente reproduzível. Como forma de divulgação científica, almeja-se a produção e publicação de artigos científicos em periódicos na área de análise de dados e modelagem estatística.

3 Estudos de caso

Nesse capítulo, são apresentados cinco estudos de caso, em que a variável resposta se apresenta como contagem e, portanto, podem ilustrar aplicações dos modelos a serem propostos.

3.1 Desfolha artificial em capulhos de algodão

No cultivo de plantas de algodão, algumas pragas, doenças, fitotoxicidade por substâncias químicas, granizo e certas injúrias mecânicas são os principais agentes de desfolha, causando prejuízos na produtividade. A fim de explorar os efeitos da redução da área foliar na cultura do algodão *Gossypium hirsutum*, Silva et al. (2012) conduziram um experimento fatorial 5 (níveis de desfolha) × 5 (estágios fenológicos) no delineamento inteiramente casualizados com cinco repetições e condições controladas em uma casa de vegetação. A variável de interesse foi o número de capulhos de algodão em cada unidade experimental (vaso com duas plantas). Na Figura 2, são apresentados os números de capulhos observados em cada combinação entre estágio fenológico e nível de desfolha.

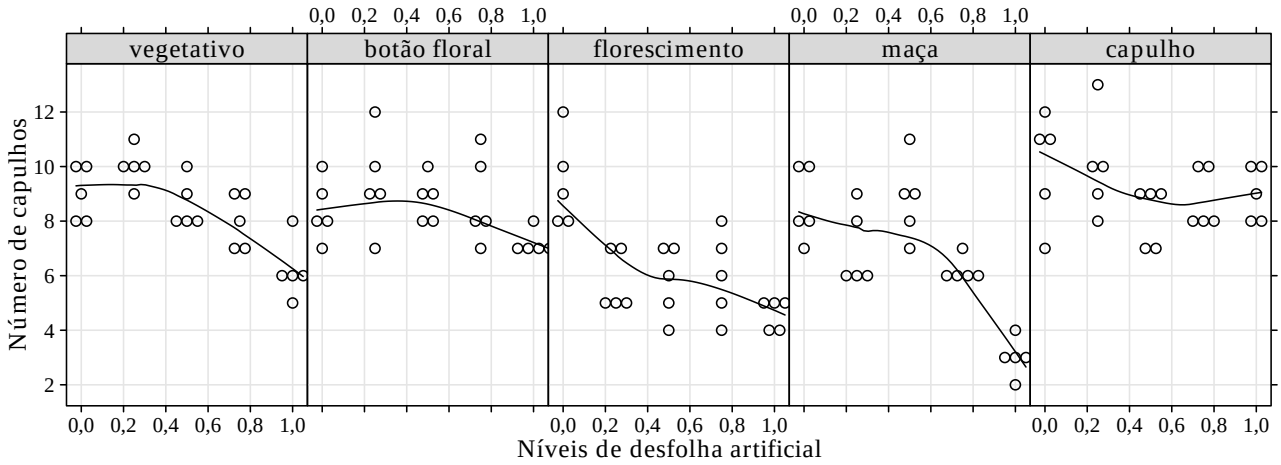


Figura 2 – Dispersão do número de capulhos observados para cada nível de desfolha artificial e estágio fenológico da planta. As linhas contínuas são curvas suaves estimadas pelo algoritmo *lowess*.

Diversas alternativas já foram propostas para análise desse conjunto de dados. Zeviani et al. (2014) utilizaram o modelo *Gamma-Count*, Huang (2017) e Ribeiro Jr et al. (2018) utilizaram reparametrizações do modelo COM-Poisson e Bonat et al. (2018) apresentaram os resultados da aplicação do modelo Poisson-Tweedie estendido. A notável relevância desse conjunto de dados é a característica de subdispersão em uma estrutura fatorial.

3.2 Dose potássica e umidade do solo na produtividade de soja

Os solos tropicais, normalmente pobres em potássio (K), quando cultivados com soja (*Glycine max L.*) demandam adubação potássica para obtenção de produtividades satisfatórias. Em um experimento fatorial (5×3) conduzido em casa de vegetação no delineamento de blocos casualizados, Serafim et al. (2012) avaliaram a influência de doses de potássio ($0, 30, 60, 120$ e 180 mg dm^{-3}) combinadas com diferentes níveis de umidade do solo ($37,5; 50$ e $62,5$ % do volume total de poros) no número de grãos e de vagens viáveis. Na Figura 3, são apresentados os valores contabilizados de cada variável de interesse, grãos de soja e vagens viáveis, para cada um dos tratamentos, combinação entre umidade do solo e adubação com potássio.

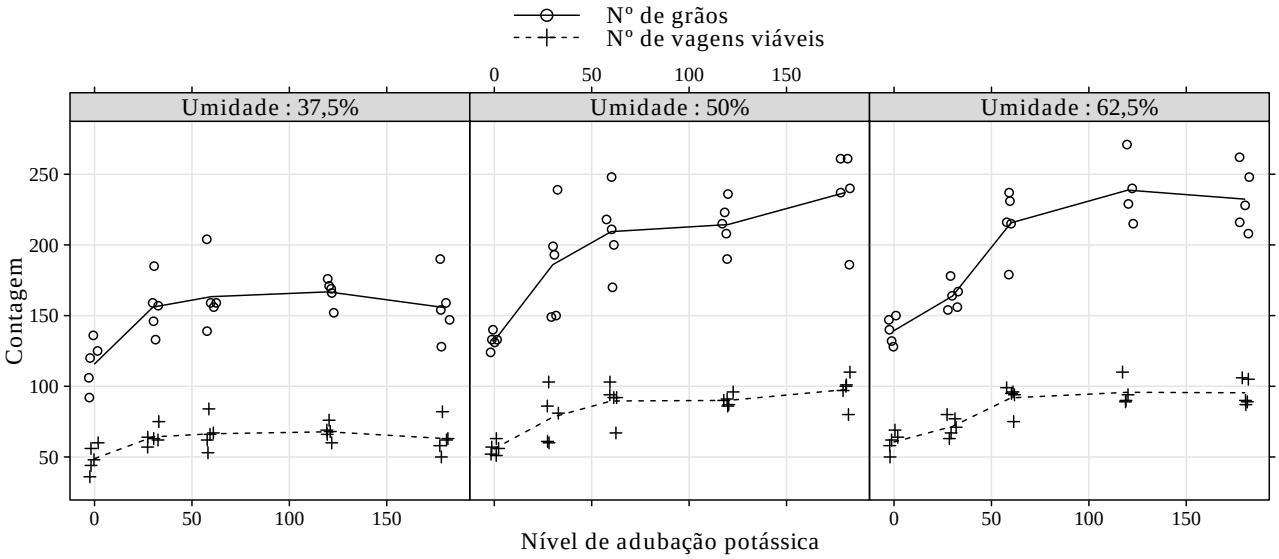


Figura 3 – Dispersão do número de grão e de vagens viáveis observado para cada nível de umidade do solo e de adubação potássica. As linhas contínuas ligam os pontos médios de cada tratamento.

Ribeiro Jr et al. (2018) analisaram o número de grãos por parcela com sua proposta de reparametrização do modelo COM-Poisson. A característica interessante desse conjunto é que há diferentes níveis de dispersão em cada contagem mensurada. Para o número de grãos por parcela, por exemplo, Ribeiro Jr et al. (2018) mostram que há uma variabilidade extra Poisson. Outra característica ainda não contemplada na análise desses dados é o comportamento não linear das contagens, em particular, modelos lineares com platô apresentam-se como bastante adequados, uma vez que se espera haver um particular nível, a partir do qual, não se tem mais influência de adubação potássica nas contagens.

3.3 Ensaio clínico em pacientes epilépticos

Epilepsia é uma doença que se manifesta por crises de perda da consciência, acompanhadas de convulsões, que ocorrem em intervalos irregulares de tempo. A manipulação de drogas para diminuir as convulsões epilépticas é um campo de interesse em medicina. Faught et al. (1996) apresentam um estudo aleatorizado, multicêntrico e duplo-cego para avaliação de uma nova droga anti-epiléptica (AED), topiramate, em comparação com um falso placebo, combinados de uma ou duas outras AEDs. A aleatorização dos pacientes ocorreu após um período de 12 semanas de estabilização. Devidamente aleatorizados, 45 pacientes foram medicados periodicamente com o placebo e 44 com a AED topiramate. Os pacientes foram acompanhados semanalmente, registrando-se o número de convulsões da respectiva semana. Na Figura 4, são apresentados os perfis do número de convulsões epilépticas para cada paciente, bem como os perfis médio e mediano. Como destacam Molenberghs et al. (2007), é interessante notar a forte assimetria à direita na distribuição do número de convulsões, a presença de valores extremos e a redução no número de pacientes acompanhados no decorrer do estudo. Além disso, observa-se que há excesso de contagens nulas (semanas em que não foram observadas convulsões epilépticas em um determinado paciente) que representam 33,122% dos dados observados.

Esse conjunto de dados é interessante, pois traz observações multivariadas para cada indivíduo, isto é, para cada i-ésimo indivíduo registram-se n_i observações (n_i, número de semanas em que o i-ésimo indivíduo foi acompanhado). Isto traz desafios para análise, que são frequentemente superados pela inclusão de efeitos aleatórios a fim de induzir correlação entre observações de um mesmo indivíduo. Molenberghs & Verbeke (2005) utilizam esse conjunto de dados como exemplo para a aplicação dos modelos lineares generalizados Poisson pela abordagem condicional (efeitos aleatórios) e marginal (equações de estimação generalizadas). Molenberghs et al. (2007) e Molenberghs et al. (2010) também utilizam esses dados como exemplo de aplicação de um modelo Poisson misto, porém os autores propõem a inclusão de duas fontes de efeitos aleatórios, para ajustar a superdispersão inerente à

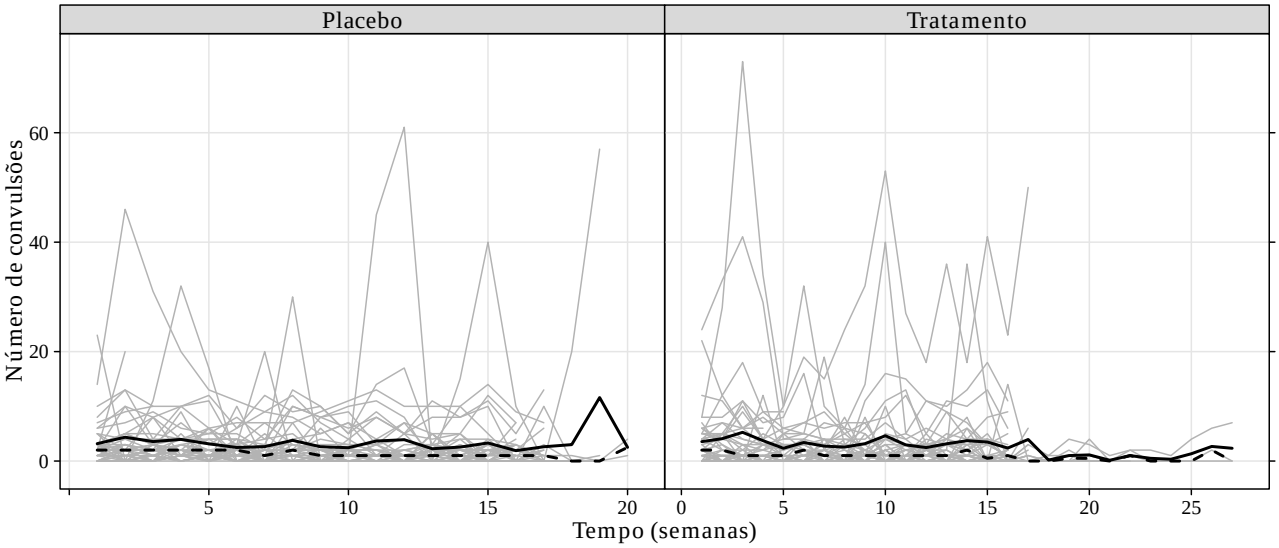


Figura 4 – Perfis do número de convulções de cada paciente dos grupos tratamento e placebo. As linhas em preto representam os perfis médio (linha contínua) e mediano (linha pontilhada).

variável de contagem, e para incorporar a estrutura longitudinal do estudo. Para análise desse conjunto de dados, pretende-se adotar uma distribuição mais flexível para o número de convulsões de cada indivíduo ao invés de incluir efeitos aleatórios em nível de observação.

3.4 Substratos alternativos para crescimento de bromélias

Xaxim é um substrato utilizado no cultivo de bromélias e orquídeas, cujo comercialização foi proibida em 2001. Desde então, há pesquisas em botânica para propor substratos alternativos ao Xaxim no cultivo de bromélias, orquídeas e outras epífitas. Esse conjunto de dados provém de um experimento aleatorizado em 4 blocos, cujo objetivo foi avaliar 5 diferentes recipientes de substratos alternativos para bromélias. Todos os tratamentos continham turfa e perlita e se diferenciavam no terceiro componente: casca de Pinus, casca de Eucaliptos, Coxim, fibra de coco e Xaxim. A variável de interesse foi o número de folhas por unidade experimental (pote com inicialmente 8 plantas), que foi registrado diariamente durante 6 dias após a plantação.

Na Figura 5(a), são apresentados os perfis individuais do número de folhas para cada bloco em cada tratamento e seus respectivos perfis médios. Nota-se um comportamento não linear (sigmoidal) dos perfis em cada tratamento e uma pouca variabilidade entre blocos em um mesmo tempo e tratamento. Essa última característica é destacada na Figura 5(b), onde os logaritmos das médias e variâncias são apresentados em um gráfico de dispersão. Para todas as combinações de tempo e tratamento, as variâncias do número de folhas obtidas dos quatro blocos são bastante menores do que as respectivas médias. Com isso, esse conjunto de dados apresenta alguns desafios para análise, pois não se encontraram na literatura, exemplos de análise de dados com as características de subdispersão em estudos longitudinais.

3.5 Infestação de mosca-branca em cultura de soja

A mosca-branca *Bemisia tabaci* é praga de diversas culturas sendo capaz de se alimentar de mais de 500 espécies de vegetais. Na cultura de soja, essa espécie de mosca-branca era pouco citada até a primeira década de 2000. Porém, períodos de seca e o uso intensivo de inseticidas e de fungicidas não selecionados afetaram seus inimigos naturais e, consequentemente, contribuíram para o aumento da importância econômica dessa praga em agronomia.

Sob essa motivação, Suekane et al. (2013) conduziram um experimento sob o delineamento

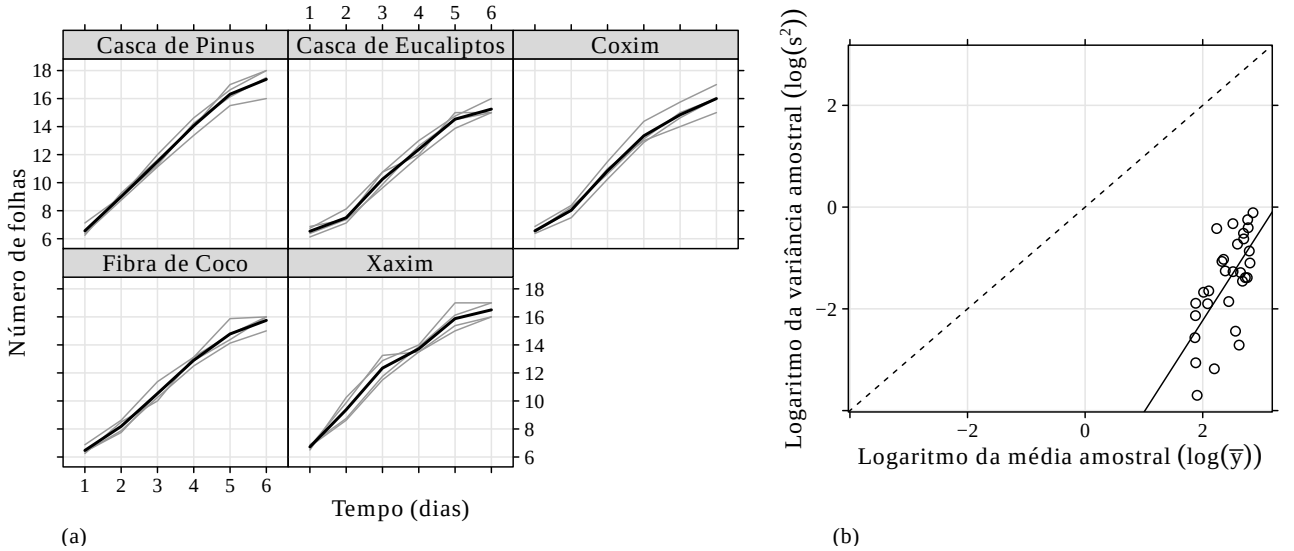


Figura 5 – (a) Perfis do número de folhas observadas para cada bloco em cada componente dos substratos alternativos, as linhas em preto representam os perfis médios. (b) Dispersão das médias e variâncias amostrais, na escala logarítmica, calculadas sobre os blocos em cada tempo e tratamento, a linha pontilhada representa a reta média = variância (equidispersão) e a contínua representa um ajuste de mínimos quadrados.

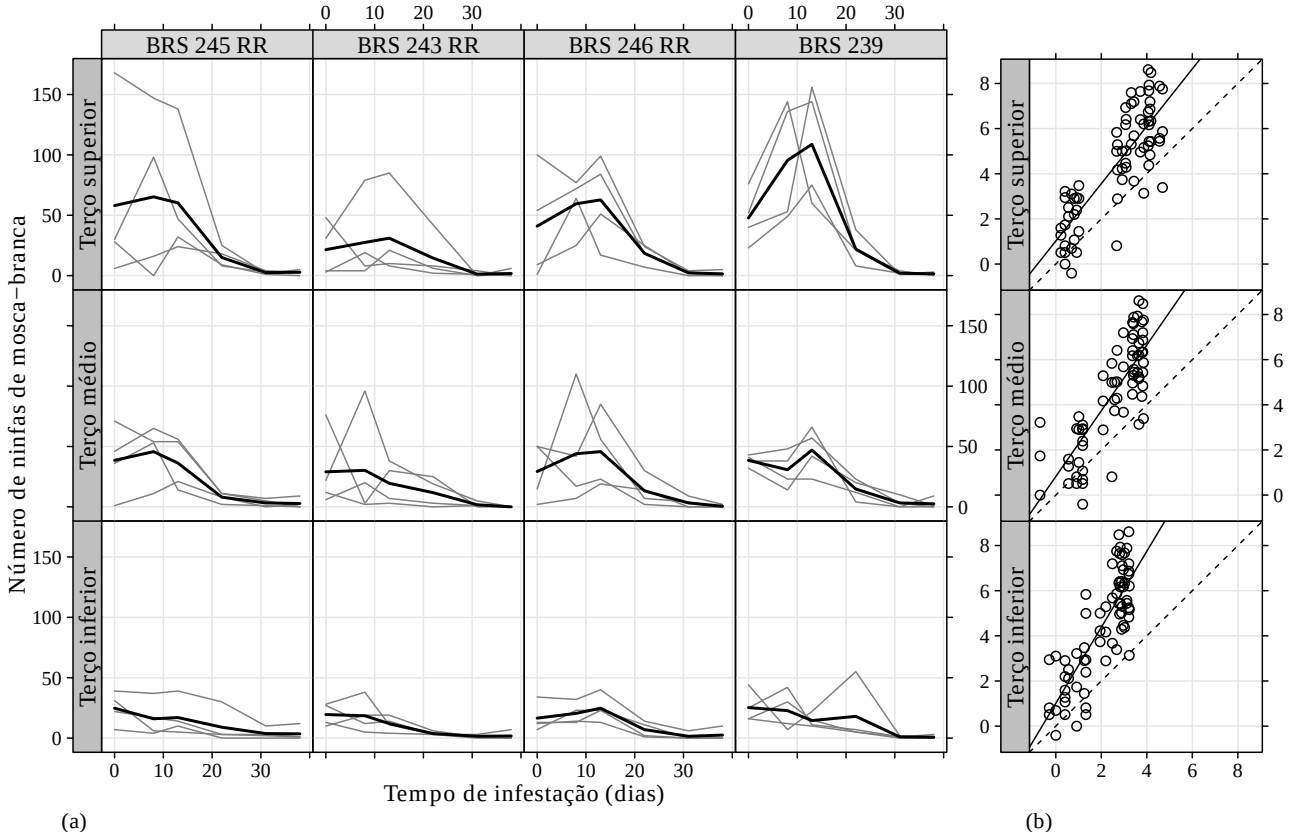


Figura 6 – (a) Perfis das contagens de ninfas observadas para cada bloco em cada uma das quatro cultivares BRS, as linhas em preto representam os perfis médios. (b) Dispersão das médias (eixo x) e variâncias (eixo y) amostrais, na escala logarítmica, calculadas sobre os blocos em cada tempo e cultivar para cada variável de interesse, a linha pontilhada representa a reta média = variância (equidispersão) e a contínua representa um ajuste de mínimos quadrados.

de blocos para avaliar o número de ninfas de mosca-branca em dez diferentes cultivares de soja. As

plantas das diferentes cultivares de soja foram cultivadas em vasos contendo duas plantas. No início do experimento, houve infestação artificial de adultos de mosca-branca na casa de vegetação. Unidades amostrais, contendo dois vasos, foram avaliadas antes da infestação e nos dias 8, 13, 22, 31 e 38, quando se mensurou o número de ninfas da mosca-branca em um folíolo do terço superior, médio e inferior da planta. Nessa pesquisa, inicialmente, pretende-se utilizar somente as cultivares BRS das 10 cultivares experimentadas, devido às contagens nessas cultivares se comportarem de forma similar e ao interesse na comparação das cultivares transgênicas BRS produzidas pela Embrapa. Na Figura 6(a), são apresentados os perfis, das três contagens realizadas, em cada bloco ao longo dos dias de infestação para cada cultivar, e os respectivos perfis médios. O comportamento dos perfis se mostra não linear, com um pico no número de ninfas ocorrendo entre 10 e 20 dias de infestação. Além disso, ressalta-se que a variabilidade das contagens é muito superior nos primeiros dias de infestação. Na Figura 6(b), apresentam-se as médias e variâncias amostrais, em escala logarítmica, para cada variável de interesse. Há uma forte evidência de superdispersão. Para análise desses dados, modelos flexíveis para dados de contagens como Poisson-Tweedie, COM-Poisson, *Gamma-Count* e Poisson-generalizado, com a inclusão de efeitos para modelar a estrutura longitudinal, são alternativas bastante relevantes que ainda não são bem consolidadas na literatura. Em adição, ressalta-se a possibilidade de modelar o parâmetro de dispersão dessas distribuições com covariáveis, uma vez que se observou que a variabilidade tem comportamento tempo-dependente.

4 Metodologia

Nesse capítulo são apresentados alguns modelos probabilísticos flexíveis para dados discretos com suporte no conjunto dos números naturais, como especificar os modelos de regressão e direções para estimação e inferência.

4.1 Distribuição Poisson

A distribuição Poisson é a principal referência para análise de dados na forma de contagens. É uma distribuição pertencente à família exponencial e tem interpretação como um modelo exponencial de dispersão (Jørgensen 1997). A distribuição de Poisson $\text{Po}(\mu)$, com média μ, tem função massa de probabilidade

$$\Pr(Y = y \mid \mu) = \frac{\mu^y \exp(-\mu)}{y!}, \quad y \in \mathbb{N}, \tag{4.1}$$

em que $\mu > 0$. Pela função geradora de cumulantes da família exponencial é fácil mostrar que $\text{E}(Y) = \text{Var(Y)} = \mu$. Como discutido no Capítulo 1, essa é uma particularidade bastante restritiva da distribuição Poisson. A linearidade das razões de probabilidades consecutivas $\Pr(Y = y-1)/\Pr(Y = y) = y/\mu$ e a relação da distribuição Poisson com a distribuição exponencial são características que motivam classes de generalizações do modelo Poisson e serão discutidas nas próximas seções.

Para especificar um modelo de regressão baseado na distribuição Poisson, considere y_i observações de uma variável Poisson e $\boldsymbol{x}_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ip})$ um vetor de covariáveis conhecidas, $i = 1, 2, \ldots, n$. O modelo de regressão Poisson é especificado como

$$Y_i \sim \text{Po}(\mu_i), \quad \text{em que} \quad \mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

em que $\boldsymbol{\beta}$ são parâmetros desconhecidos a serem estimados e $g^{-1}(\cdot)$ uma função de ligação conhecida, por exemplo, logaritmo (função de ligação canônica) e raiz quadrada.

4.2 Distribuição COM-Poisson

A distribuição COM-Poisson é a principal representante da família de distribuições Poisson ponderadas (WPD) (Del Castillo & Pérez-Casany 1998). Uma variável aleatória Y pertence à família WPD se sua função massa de probabilidade puder ser escrita como

$$\Pr(Y = y) = \frac{w(y)\exp(-\lambda)\lambda^y}{y!\text{E}_\lambda[w(Y)]}, \quad y \in \mathbb{N}, \tag{4.2}$$

em que $\text{E}_\lambda(\cdot)$ é o valor médio calculado a partir de uma variável aleatória Poisson de parâmetro λ, chamada de constante de normalização; e $w(y)$ é uma função peso, não negativa e tal que $\text{E}_\lambda[w(Y)]$ seja finita. A função peso $w(y) \equiv w(y, \nu)$, pode depender de um parâmetro adicional de tal forma que sub e superdispersão sejam abrangidas. Obtém-se a distribuição COM-Poisson $\text{CMP}(\lambda, \nu)$ tomando-se $w(y, \nu) = (y!)^{1-\nu}$, para $\nu \geq 0$ (Sellers et al. 2012). Sua função de probabilidade assume a forma

$$\Pr(Y = y) = \frac{\lambda^y \exp(-\lambda)}{(y!)^\nu \text{E}_\lambda[(Y!)^{1-\nu}]} = \frac{\lambda^y}{(y!)^\nu \text{Z}(\lambda, \nu)}, \quad \text{Z}(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}, \tag{4.3}$$

em que ν é dito como o parâmetro de dispersão tal que para $0 < \nu < 1$ e $\nu < 1$ têm-se os casos de superdispersão e subdispersão, respectivamente. Para as distribuições pertencentes à família WPD e, particularmente, para a distribuição COM-Poisson, a razão entre as probabilidades de dois eventos consecutivos é dada por

$$\frac{\Pr(Y = y-1)}{\Pr(Y = y)} = \frac{y}{\lambda}\frac{w(y-1)}{w(y)} \overset{\text{CMP}}{\equiv} \frac{y^\lambda}{\lambda}, \tag{4.4}$$

enquanto que para a distribuição Poisson essa razão é y/λ, correspondente a $w(y)$ constante ou, no caso COM-Poisson, $\nu = 1$. Além da Poisson ($\nu = 1$), a distribuição COM-Poisson também tem como caso particular a distribuição geométrica ($\nu = 0$ e $\lambda < 1$) e como caso limite a distribuição Bernoulli ($\nu \to \infty$, sendo a probabilidade de sucesso $\lambda/(\lambda + 1)$).

Um inconveniente desse modelo é que os momentos média e variância, em geral, não são obtidos em forma fechada. A partir de uma aproximação para $\mathrm{Z}(\lambda, \nu)$, Shmueli et al. (2005) e Sellers & Shmueli (2010) mostram que a esperança e a variância de uma variável $Y \sim \mathrm{CMP}(\lambda, \nu)$ podem ser aproximadas por

$$\mathrm{E}(Y) \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \qquad \text{e} \qquad \mathrm{Var}(Y) \approx \frac{\lambda^{1/\nu}}{\nu}, \tag{4.5}$$

que são particularmente acuradas para $\nu \leq 1$ ou $\lambda > 10^{\nu}$ (Shmueli et al. 2005). Ribeiro Jr et al. (2018) mostraram que a aproximação para a média é acurada, diferentemente da aproximação para variância, que perde acurácia quando $\nu < 1$. Além disso, os autores indicam que a acurácia das aproximações parece não ter relação com as regiões $\nu \leq 1$ ou $\lambda > 10^{\nu}$.

Modelos de regressão baseados na distribuição COM-Poisson foram propostos por Sellers & Shmueli (2010). Para y_i observações independentes do modelo COM-Poisson e $\boldsymbol{x}_i^{\top} = (x_{i1}, x_{i2}, \ldots, x_{ip})$ um vetor de covariáveis conhecidas, $i = 1, 2, \ldots, n$, o modelo de regressão COM-Poisson é definido como

$$Y_i \sim \mathrm{CMP}(\lambda_i, \nu), \quad \text{em que} \quad \lambda_i = g^{-1}(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}).$$

Embora outras funções de ligação sejam possíveis, a função de ligação logarítmica é, frequentemente, adotada.

Essa formulação tem o grande inconveniente de interpretabilidade, uma vez que o modelo de regressão está associado a um parâmetro que não representa a média. Ainda, para a distribuição COM-Poisson, em sua parametrização original, λ e ν são fortemente intra-relacionados na função de verossimilhança (Ribeiro Jr et al. 2018). Sob essas motivações Huang (2017) e Ribeiro Jr et al. (2018) propuseram reparametrizações para a média no modelo COM-Poisson. Nesses casos, o modelo de regressão é especificado como

$$Y_i \sim \mathrm{CMP}_{\mu}(\mu_i, \nu), \quad \text{em que} \quad \mu_i = g^{-1}(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}),$$

sendo $\mathrm{CMP}_{\mu}(\mu_i, \nu)$ a distribuição COM-Poisson reparametrizada para a média. Na abordagem proposta por Huang, o parâmetro μ_i da distribuição $\mathrm{CMP}_{\mu}(\mu_i, \nu)$ é obtido como raiz da equação

$$\sum_{j=0}^{\infty}(j - \mu_i)\frac{\lambda_i^j}{(j!)^{\nu}} = 0,$$

enquanto que Ribeiro Jr et al., obtém μ_i a partir da aproximação para a média,

$$\mu_i = h(\lambda_i, \nu) = \lambda_i^{1/\nu} - \frac{\nu - 1}{2\nu}.$$

Em ambos os casos, μ_i e ν se mostram ortogonais e os parâmetros $\boldsymbol{\beta}$ em um modelo de regressão têm a mesma interpretação do modelo Poisson.

4.3 Distribuição *Gamma-Count*

A distribuição *Gamma-Count* é uma generalização da distribuição Poisson que resulta da relação da Poisson com a distribuição do tempo entre eventos. Seja $\tau_k > 0$, $k \in \mathbb{N}^*$ o tempo entre o $(k-1)$ e o k-ésimo evento e ϑ_n o tempo de chegada do n-ésimo evento. Então $\vartheta_n = \sum_{k-1}^{n}\tau_k$ para $n = 1, 2, \ldots$. Seguindo Winkelmann (1995), seja Y_T o número total de eventos no intervalo $(0, T)$. Para T fixo, Y_T é uma variável aleatória de contagem. Segue das definições de Y_t e ϑ_n que $Y_T < y$ e $\vartheta_y \geq T$ são eventos equivalentes e portanto,

$$\begin{aligned} \Pr(Y_T < y) &= \Pr(\vartheta_y \geq T) = 1 - \mathrm{F}_y(T), \\ \Pr(Y_T = y) &= \Pr(Y_T < y) - \Pr(Y_T < y + 1) = \mathrm{F}_y(T) - \mathrm{F}_{y+1}(T), \end{aligned} \tag{4.6}$$

em que $\mathrm{F}_y(T)$ é a função distribuição acumulada de ϑ_y e T é a amplitude do intervalo de contagem, frequentemente denotado como *offset* nos modelos de regressão.

A Equação (4.6) resulta uma distribuição de contagem a partir, apenas, do conhecimento da distribuição do tempo entre eventos. Por exemplo, a distribuição Poisson deriva da suposição de que τ_k são exponencialmente distribuídos com parâmetro de taxa λ. Assim, tem-se que ϑ_y segue uma distribuição de Erlang de parâmetros y e λ, e a função massa de probabilidade de Y_T é obtida como

$$\Pr(Y_T = y) = \mathrm{F}_y(T) - \mathrm{F}_{y+1}(T) = \sum_{j=0}^{y-1} \frac{\exp(-\lambda T)(\lambda T)^j}{j!} - \sum_{j=0}^{y} \frac{\exp(-\lambda T)(\lambda T)^j}{j!} = \frac{\exp(-\lambda T)(\lambda T)^j}{y!},$$

que, para $T = 1$, é idêntica à Equação (4.1).

Para a distribuição *Gamma-Count* $\mathrm{GCT}(\alpha, \gamma)$ assume-se distribuição gama, com de parâmetros α e γ, para os tempos entre eventos τ_k. Consequentemente, a distribuição de ϑ_y também é gama, porém com parâmetros $y\alpha$ e γ. Sua função massa de probabilidade de uma variável aleatória $Y_t \sim \mathrm{GCT}((\alpha, \gamma))$ é obtida por

$$\Pr(Y_T = y) = \mathrm{F}_y(T) - \mathrm{F}_{n+1}(T) = \int_0^T \frac{\gamma^{y\alpha} t^{y\alpha-1}}{\Gamma(y\alpha)\exp(\gamma t)} dt - \int_0^T \frac{\gamma^{(y+1)\alpha} t^{(y+1)\alpha-1}}{\Gamma[(y+1)\alpha]\exp(\gamma t)} dt, \tag{4.7}$$

que não tem forma fechada, a menos do caso particular $\alpha = 1$ quando a distribuição reduz-se à Poisson. Os momentos da distribuição também não são obtidos em forma fechada. Winkelmann (1995) mostra que para intervalos de observação suficientemente grandes, $T \to \infty$, sustenta-se que

$$Y_T \overset{a}{\sim} \mathcal{N}\left(\frac{\gamma T}{\alpha}, \frac{\gamma T}{\alpha^2}\right),$$

assim a razão média-variância assintótica é $1/\alpha$. Consequentemente, a distribuição *Gamma-Count* é capaz de modelar superdispersão ($0 < \alpha < 1$) e subdispersão ($\alpha > 1$).

Modelos de regressão *Gamma-Count* foram propostos por Winkelmann (1995) e Zeviani et al. (2014). Para y_i observações independentes do modelo $\mathrm{GCT}(\alpha, \gamma_i)$ e $\boldsymbol{x}_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ip})$ um vetor de covariáveis conhecidas, $i = 1, 2, \ldots, n$, o modelo de regressão *Gamma-Count* é definido como

$$Y_i \sim \mathrm{GCT}(\alpha, \gamma_i), \quad \text{em que} \quad \gamma_i = \alpha \left[g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right]^{-1}.$$

A desvantagem desse modelo é que, embora assintoticamente ($T \to \infty$), γ_i/α represente a média das contagens, em geral, modela-se a variável de contagem na escala do tempo entre eventos, perdendo a interpretação dos parâmetros de regressão β.

4.4 Distribuição Poisson generalizada

A distribuição Poisson generalizada é resultante de uma forma limite da distribuição binomial negativa generalizada e pode modelar sub e superdispersão (Zamani & Ismail 2012). Seja Y uma variável aleatória que segue a distribuição Poisson generalizada $\mathrm{GPo}(\lambda, \gamma)$, então sua função massa de probabilidade é dada por

$$\Pr(Y = y) = \begin{cases} \left[\lambda(\lambda + y\gamma)^{y-1}\exp(-\lambda - y\gamma)\right]/y!, & y = 0, 1, 2, \ldots \\ 0 & \text{para } y > m, \text{quando } \gamma < 0, \end{cases} \tag{4.8}$$

em que $\lambda > 0$, $\max(-1, -\lambda/4) \leq \gamma \leq 1$ e m é o maior inteiro positivo tal que $\lambda + m\gamma > 0$ quando γ é negativo (Consul & Famoye 1992). A média e a variância são dadas por $\mathrm{E}(Y) = \lambda(1-\gamma)^{-1}$ e $\mathrm{Var}(Y) = \lambda(1-\gamma)^{-3}$, respectivamente. A distribuição Poisson é um caso particular, quando $\gamma = 0$.

Para modelos de regressão, sugere-se a parametrização para a média

$$\lambda = \frac{\mu}{1+\alpha\mu} \quad \text{e} \quad \gamma = \frac{\alpha\mu}{1+\alpha\mu},$$

que substituindo em (4.8) leva à função massa de probabilidade

$$\Pr(Y=y)=\left(\frac{\mu}{1+\alpha\mu}\right)^{y}\frac{(1+\alpha y)^{y-1}}{y!}\exp\left[-\mu\frac{(1+\alpha y)}{(1+\alpha\mu)}\right], \tag{4.9}$$

denotada por $\text{GPo}_{\mu}(\mu,\alpha)$. Os momentos da distribuição nessa parametrização são

$$\text{E}(Y)=\mu \quad \text{e} \quad \text{Var}(Y)=\mu(1+\mu\alpha)^2, \tag{4.10}$$

que garantem bastante flexibilidade à distribuição, uma vez que a relação média–variância é determinada como uma função cúbica de μ.

Para y_i observações da distribuição Poisson generalizada e $\boldsymbol{x}_i$ um vetor conhecido de covariáveis, o modelo de regressão baseado na distribuição Poisson generalizada é definido por

$$Y_i\sim\text{GPo}_{\mu}(\mu_i,\alpha), \quad \text{em que} \quad \mu_i=g^{-1}(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}),$$

em que $\alpha>\min[-\max(y_i^{-1}),-\max(\mu_i^{-1})]$, quando $\alpha<0$.

Aplicações do modelo de regressão Poisson generalizada são pouco reportadas na literatura. Embora bastante flexível, a grande dificuldade desse modelo reside na complicada restrição do espaço paramétrico, que é difícil de se incorporar, de forma eficiente, no processo de estimação.

4.5 Distribuição Poisson-Tweedie

Distribuições de variáveis na forma de contagens baseadas em especificações hierárquicas da distribuição Poisson são comuns para análise de dados superdispersos, e.g. distribuição binomial negativa. A distribuição Poisson-Tweedie representa um caso geral dos modelos Poisson com especificação hierárquica (Jørgensen 1997, Seção 4.6).

A distribuição Tweedie pertence à classe dos modelos exponenciais de dispersão (Jørgensen 1997). Seja $\text{E}(Y)=\mu$ e $\text{Var}_p(Y)=\phi\mu^p$ então, Y segue distribuição Tweedie $\text{Tw}_p(\mu,\phi)$, em que $\mu\in\Omega_p$, $\phi>0$ e $p\in(-\infty,0]\cup[1,\infty)$. Note que o suporte da distribuição depende de p, que age como um selecionador de distribuições uma vez que as distribuições Gaussiana ($p=0$), Poisson ($p=1$), non-central gamma ($p=3/2$), gamma ($p=2$) e Gaussiana inversa ($p=3$) são casos particulares. A função densidade de probabilidade da distribuição Tweedie não pode ser obtida em forma fechada.

A distribuição Poisson-Tweedie $\text{PTw}_p(\mu,\phi)$ é resultante da especificação hierárquica

$$Y\mid Z\sim\text{Po}(Z) \quad \text{em que} \quad Z\sim\text{Tw}_p(\mu,\phi), \tag{4.11}$$

que não tem forma fechada para função de probabilidade, exceto para casos especiais. A esperança e a variância de uma variável aleatória Poisson-Tweedie são obtidos por

$$\begin{aligned}\text{E}(Y)&=\text{E}[\text{E}(Y|Z)]=\mu\\ \text{Var}(Y)&=\text{Var}[\text{E}(Y|Z)]+\text{E}[\text{Var}(Y|Z)]=\mu+\phi\mu^p.\end{aligned} \tag{4.12}$$

Devido à flexibilidade da distribuição Tweedie, a Poisson-Tweedie também tem importantes casos particulares que incluem as distribuições Hermite ($p=0$), Neymann tipo-A ($p=1$), Pólya-Aeppli ($p=1,5$), binomial negativa ($p=2$) e Poisson–inversa gaussiana ($p=3$) (Bonat et al. 2018).

Pela definição (4.11), modelos Poisson-Tweedie só modelam superdispersão. Bonat et al. (2018) estendem essa distribuição para contemplar subdispersão, adotando apenas a especificação de momentos baseada nas expressões em (4.12) e permitindo a estimação de $\phi<0$. Essa abordagem é análoga aos modelos de quase-verossimilhança (Wedderburn 1974), e, a menos dos casos particulares, não se conhece a distribuição completa de Y, impossibilitando o cálculo de probabilidades, por exemplo.

Para observações $y_1,y_2,\ldots,y_n$ e seus respectivos vetores de covariáveis $\boldsymbol{x}_i$, os modelos de regressão Poisson-Tweedie são definidos por

$$Y_i\sim\text{PTw}_p(\mu_i,\phi), \quad \text{em que} \quad \mu_i=g^{-1}(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}).$$

Em sua versão estendida, a especificação é dada por

$$\mathrm{E}(Y_i) = \mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \quad \text{e} \quad \mathrm{Var}(Y_i) = \mu_i + \phi\mu_i^p.$$

Note que para superdispersão ambas as especificações são equivalentes.

A grande motivação da utilização dessa distribuição é sua função de variância, bastante flexível, em que a estimação de p funciona como um selecionador automático de distribuições.

4.6 Estimação e inferência em modelos de regressão

Para as versões paramétricas dos modelos apresentados anteriormente, o processo de estimação pode ser feito sob o paradigma de verossimilhança, que é bastante intuitivo e fornece estimadores consistentes e assintoticamente não viesados (Pawitan 2001). Para $\boldsymbol{y}$ um vetor $n \times 1$ conhecido de observações de uma mesma classe de distribuições de probabilidade $f(y, \boldsymbol{\theta})$ e $\boldsymbol{X}$ uma matriz $n \times p$ de delineamento ou matriz do modelo. A função de verossimilhança e seu logaritmo são dados por

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}, \boldsymbol{x}_i) \quad \text{e} \quad \ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} \log[f(y_i; \boldsymbol{\theta}_i, \boldsymbol{x}_i)],$$

respectivamente, em que $\boldsymbol{\theta} \in \Theta$ é o conjunto de parâmetros, desconhecido, que determina a distribuição e $\boldsymbol{x}_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ip})$ o vetor de covariáveis da i-ésima observação. É interessante notar que para dados discretos, em que $f(y_i; \boldsymbol{\theta}; \boldsymbol{x}_i) \equiv \Pr(Y_i = y_i; \boldsymbol{\theta}, \boldsymbol{x}_i)$, a função de verossimilhança representa a probabilidade de se observar $\boldsymbol{y}$, dado que $\boldsymbol{\theta}$ é verdadeiro. O estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ é obtido tal que $\mathcal{L}(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) = \max_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})$. Frequentemente, não é possível obter expressões em forma fechada para os estimadores de máxima verossimilhança (Bonat et al. 2017). Entretanto, é usual assumir as estimativas de máxima verossimilhança como soluções para as equações escore

$$\mathcal{U}(\boldsymbol{\theta}) = \left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_2}, \ldots, \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p} \right)^\top = \mathbf{0}, \tag{4.13}$$

que são obtidas por métodos iterativos como Newton-escore, sendo a matriz Hessiana

$$\mathcal{H}(\theta) = \begin{bmatrix} \frac{\partial \ell^2(\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial \ell^2(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial \ell^2(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial \ell^2(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial \ell^2(\boldsymbol{\theta})}{\partial \theta_2^2} & \cdots & \frac{\partial \ell^2(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \ell^2(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_1} & \frac{\partial \ell^2(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_2} & \cdots & \frac{\partial \ell^2(\boldsymbol{\theta})}{\partial \theta_p^2} \end{bmatrix}$$

obtida de forma analítica (e.g. Poisson). Quando $\mathcal{H}(\boldsymbol{\theta})$ não tem forma fechada (e.g. COM-Poisson, *Gamma-Count*, Poisson generalizada), métodos que não necessitam da especificação das derivadas de $\ell(\boldsymbol{\theta})$, como BFGS (Nocedal & Wright 1995) são adotados. Nesses casos, a matriz Hessiana é aproximada, geralmente, por diferenças finitas o que demanda um número maior de avaliações do logaritmo da função de verossimilhança, tornando o processo de estimação mais lento.

A inferência nos modelos paramétricos, ajustados por meio da maximização da função de verossimilhança, baseia-se na distribuição assintótica dos estimadores de máxima verossimilhança

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\theta}, \mathcal{I}(\boldsymbol{\theta})) \quad e \quad g(\hat{\boldsymbol{\theta}}) \overset{a}{\sim} \mathcal{N}(g(\boldsymbol{\theta}), \boldsymbol{G}^\top \mathcal{I}(\boldsymbol{\theta}) \boldsymbol{G}),$$

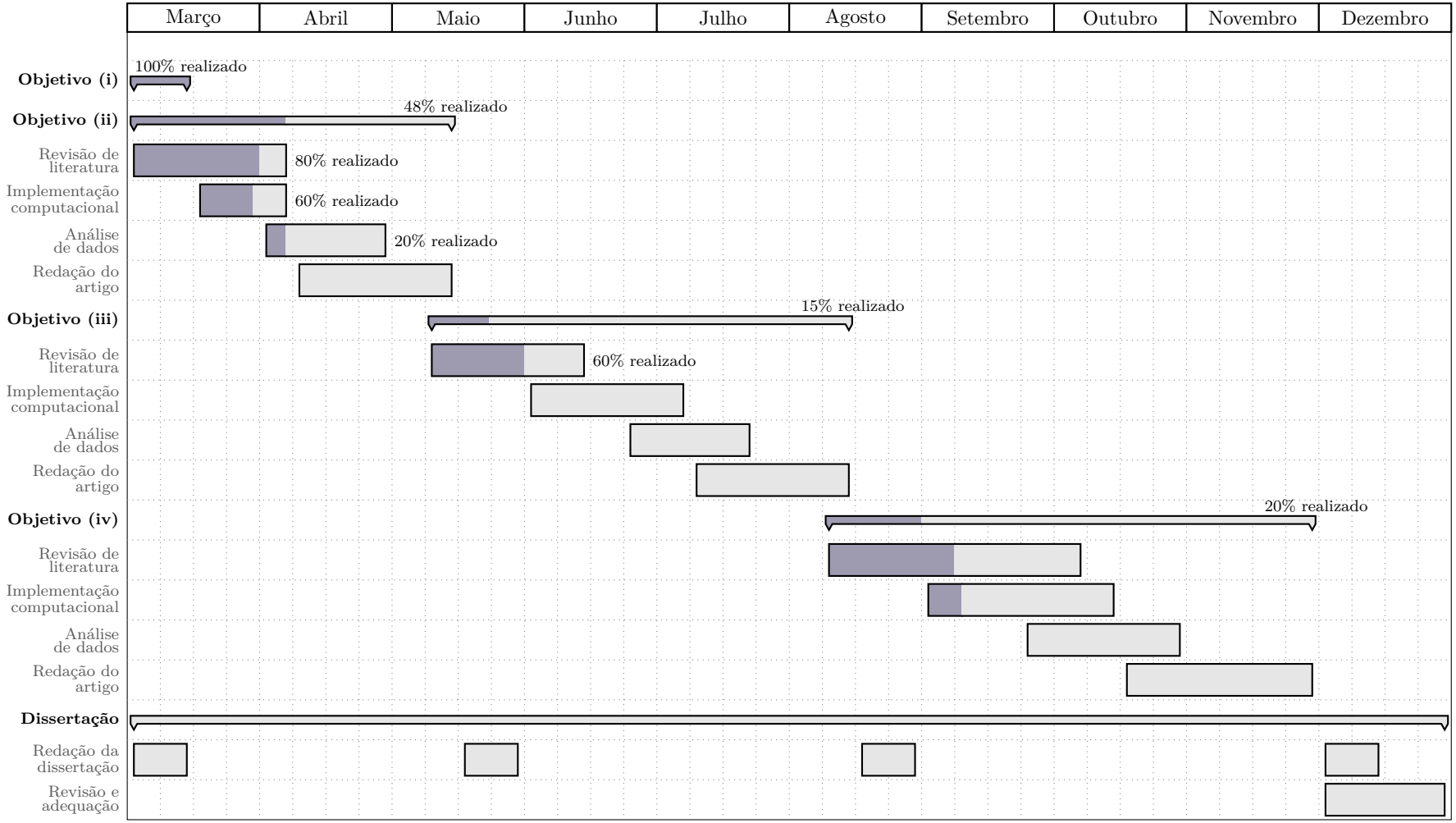
em que $\mathcal{I}(\boldsymbol{\theta}) = -\mathcal{H}^{-1}$ é a matriz de informação observada de Fisher, $g(\cdot)$ é uma função monótona e diferenciável e $\boldsymbol{G}^\top = (\partial g/\partial \beta_1, \ldots, \partial g/\partial \beta_p)^\top$ o vetor de derivadas primeiras de $g(\cdot)$.

Para a distribuição Poisson-Tweedie a estimação pelo método da máxima verossimilhança é computacionalmente exigente, uma vez que as probabilidades são dadas por uma integral intratável, exigindo o uso de métodos numéricos. Considerando a especificação de momentos, Bonat et al. (2018)

propuseram o uso combinado das funções de estimação quase-escore e de Pearson para estimação dos parâmetros de regressão e de dispersão (p, ϕ), respectivamente.

Com essa pesquisa, a análise de dados na forma de contagens pretende ser mais fiel ao processo gerador dos dados, levando a conclusões mais assertivas nos experimentos conduzidos, principalmente, na área agronômica. Com a apresentação e caracterização de distribuições flexíveis para dados dessa natureza, definição dos modelos de regressão, extensões para modelagem da dispersão e inclusão de efeitos aleatórios e discussão dos métodos de estimação, a pesquisa almeja ser uma referência metodologia para a estatística aplicada.

5 Cronograma de Atividades



Piracicaba, 15 de fevereiro de 2018.

Eduardo Elias Ribeiro Junior

De acordo

Clarice Garcia Borges Demétrio

Aprovado pelo programa em _____.

Carlos Tadeu dos Santos Dias
Coordenador do PPG em Estatística e
Experimentação Agronômica

REFERÊNCIAS

Bonat, W. H., Jørgensen, B., Kokonendji, C. C., Hinde, J. & Démetrio, C. G. B. (2018), 'Extended Poisson-Tweedie: properties and regression model for count data', *Statistical Modelling* **18**(1), 24–49.

Bonat, W. H., Zeviani, W. M. & Ribeiro Jr, E. E. (2017), *Regression Models for Count Data: beyond Poisson model*, XV EMR - Brazilian Regression Model School, Goiás, Brazil.

Consul, P. C. & Famoye, F. (1992), 'Generalized Poisson regression model', *Communication in Statistics – Theory and Methods* **21**(1), 89–109.

Del Castillo, J. & Pérez-Casany, M. (1998), 'Weighted Poisson distributions for overdispersion and underdispersion situations', *Annals of the Institute of Statistical Mathematics* **50**(3), 567–585.

Dobson, A. J. & Barnett, A. G. (2008), *An Introduction to Generalized Linear Models*, 3rd edition edn, Chapman & Hall/CRC, Boca Raton.

Faught, E., Wilder, B. J., Ramsay, R. E., Reife, R. A., Kramer, L. D., Pledger, G. W. & Karim, R. M. (1996), 'Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosages', *American Academy of Neurology* **46**(6), 1684–1690.

Hinde, J. & Demétrio, C. G. B. (1998), 'Overdispersion: models and estimation', *Computational Statistics & Data Analysis* **27**(2), 151–170.

Huang, A. (2017), 'Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts', *Statistical Modelling* **17**(6), 1–22.

Jørgensen, B. (1997), *The Theory of Dispersion Models*, Chapman & Hall, London.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edition edn, Chapman & Hall, London.

Molenberghs, G. & Verbeke, G. (2005), *Models for Discrete Longitudinal Data*, Springer, New York.

Molenberghs, G., Verbeke, G. & Demétrio, C. G. B. (2007), 'An extended random-effects approach to modelling repetead, overdispersed count data', *Lifetime Data Analysis* **13**, 513–531.

Molenberghs, G., Verbeke, G. & Demétrio, C. G. B. (2017), 'Hierarchical models with normal and conjugate random effects: a review', *SORT : statistics and operations research transactions* **41**(2), 191–254 (Invited articles).

Molenberghs, G., Verbeke, G., Demétrio, C. G. B. & Vieira, A. M. C. (2010), 'A family of generalized linear models for repeated measures with normal and conjugate random effects', *Statistical Science* **25**(3), 325–347.

Nelder, J. A. & Wedderburn, R. W. M. (1972), 'Generalized Linear Models', *Journal of the Royal Statistical Society. Series A (General)* **135**, 370–384.

Nocedal, J. & Wright, S. J. (1995), *Numerical optimization*, Springer, New York.

Paula, G. A. (2013), *Modelos de Regressão com apoio computacional*, IME–USP, São Paulo.

Pawitan, Y. (2001), *In all likelihood: statistical modelling and inference using likelihood*, Oxford University Press.

Ribeiro Jr, E. E., Zeviani, W. M., Bonat, W. H., Demétrio, C. G. B. & Hinde, J. (2018), 'Reparametrization of COM-Poisson regression models with applications in the analysis of experimental data', *arXiv (Statistics Applications and Statistics Methodology)* .

Sellers, K. F., Borle, S. & Shmueli, G. (2012), 'The COM-Poisson model for count data: a survey of methods and applications', *Applied Stochastic Models in Business and Industry* **28**(2), 104–116.

Sellers, K. F. & Shmueli, G. (2010), 'A flexible regression model for count data', *Annals of Applied Statistics* **4**(2), 943–961.

Serafim, M. E., Ono, F. B., Zeviani, W. M., Novelino, J. O. & Silva, J. V. (2012), 'Umidade do solo e doses de potássio na cultura da soja', *Revista Ciência Agronômica* **43**(2), 222–227.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. & Boatwright, P. (2005), 'A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution', *Journal of the Royal Statistical Society. Series C: Applied Statistics* **54**(1), 127–142.

Silva, A. M., Degrande, P. E., Suekane, R., Fernandes, M. G. & Zeviani, W. M. (2012), 'Impacto de diferentes níveis de desfolha artificial nos estágios fenológicos do algodoeiro', *Revista de Ciências Agrárias* **35**(1), 163–172.

Steutel, F. W. & Thiemann, J. G. F. (1989), 'The gamma process and the Poisson distribution', *(Memorandum COSOR; Vol. 8924). Eindhoven: Technische Universiteit Eindhoven.* .

Suekane, R., Degrande, P., de Lima Junior, I., de Queiroz, M. & Rigoni, E. (2013), 'Danos da mosca-branca *Bemisia tabaci* e distribuição vertical das ninfas em cultivares de soja em casa de vegetação', *Arquivos do Instituto Biológico* **80**(2), 151–158.

Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S*, 4th edition edn, Springer.

Wedderburn, R. W. M. (1974), 'Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method', *Biometrika* **61**(3), 439.

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

Winkelmann, R. & Zimmermann, K. F. (1994), 'Count data models for demographic data', *Mathematical Population Studies* **4**(3), 205–221.

Zamani, H. & Ismail, N. (2012), 'Functional form for the generalized Poisson regression model', *Communication in Statistics – Theory and Methods* **41**, 3666–3675.

Zeviani, W. M., Ribeiro Jr, P. J., Bonat, W. H., Shimakura, S. E. & Muniz, J. A. (2014), 'The Gamma-count distribution in the analysis of experimental underdispersed data', *Journal of Applied Statistics* pp. 1–11.