**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Modelos para dados de contagem não equidispersos com aplicação à
ecologia e em estudos longitudinais**

Douglas Toledo Batista

Tese apresentada para obtenção do título de Doutor em
Ciências. Área de concentração: Estatística e Experi-
mentação Agronômica

**Piracicaba
2020**

Douglas Toledo Batista
Bacharel em Estatística

**Modelos para dados de contagem não equidispersos com aplicação à
ecologia e em estudos longitudinais**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:

Prof. Dr. **IDEMAURO ANTONIO RODRIGUES
DE LARA**

Tese apresentada para obtenção do título de Doutor em
Ciências. Área de concentração: Estatística e Experi-
mentação Agronômica

Piracicaba
2020

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP**

Batista, Douglas Toledo

Modelos para dados de contagem não equidispersos com aplicação à ecologia e em estudos longitudinais / Douglas Toledo Batista. -- versão revisada de acordo com a resolução CoPGr 6018 de 2011. -- Piracicaba, 2020 .

80 p.

Tese (Doutorado) -- USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Macrófitas aquáticas 2. Distribuição de probabilidade 3. Modelos mistos 4. *Half-normal plot* 5. Dados correlacionados 6. Progabide . I. Título.

DEDICATÓRIA

Essa etapa da minha vida pôde ser concluída graças ao apoio e incentivo incondicional da minha mãe, Adileuza Maria Toledo Siqueira, pessoa amável e admirável que sempre esteve ao meu lado nos momentos felizes e difíceis. Nunca me deixou desistir dos sonhos e objetivos da minha vida.

AGRADECIMENTOS

Meus agradecimentos vão a todos os colegas, amigos e familiares que acompanharam essa longa trajetória e contribuíram para a concretização deste trabalho. Em especial:

À minha mãe Adileuza Maria Toledo Siqueira, pelo amor incondicional e incentivo constante em minha vida;

À minha esposa Cristiane Umetsu, pelo amor, dedicação, incentivo e presença constante nesta etapa da minha vida. Agradeço imensamente pela paciência em inúmeros momentos desta trajetória e por aguentar o meu mau humor. Ter você ao meu lado na construção e elaboração desse trabalho foi muito importante, pois sem você tudo teria sido mais angustiante. Obrigado pelos momentos de discussão sobre o tema e contribuições valiosas para toda uma vida;

Ao meu orientador Prof. Dr. Idemauro Antonio Rodrigues de Lara, pela disponibilidade e entrega ao ensinar, pelas valiosas conversas, troca de experiências e competência na condução deste trabalho;

Aos professores e funcionários do Departamento de Matemática e Estatística da ESALQ-USP;

Aos meus amigos de Piracicaba, em especial Daniel Miquelluti, Erasnilson Camilo, Fábio Prata, Fernando Mayer, Júlio Vasconcelos, Otávio Barros, Rick Manguiera, Simone Grego, Talita Tanaka e Valiana Teodoro;

Ao Prof. Dr. David Jose Miquelluti, que em um bate papo sobre o tema nas “botecagens” da vida, conseguiu contribuir de forma relevante para elaboração deste trabalho;

Ao Prof. Dr. Altemir da Silva Braga, que mesmo distante sempre esteve disposto a me auxiliar e sanar as dúvidas que surgiram ao longo do doutorado;

Aos membros da banca examinadora, Dr. Cesar Augusto Taconeli e Dra. Maria Imaculada de Lima Montebello, cujas observações foram de extrema importância para incremento deste trabalho;

À Profa. Dra. Elizabeth Mie Hashimoto, pela prontidão e gentileza em revisar a primeira versão da tese;

À Comissão de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) pelo aporte financeiro com a bolsa de estudos e recursos ao meu programa de Pós-Graduação;

Finalmente, a todos que de alguma forma contribuíram direta ou indiretamente para a realização deste trabalho.

SUMÁRIO

Resumo	7
Abstract	8
1 Introdução	9
Referências	12
2 Modelos para dados de contagem não equidispersos: Uma aplicação em dados ecológicos com subdispersão	17
Resumo	17
2.1 Introdução	17
2.2 Revisão de modelos de dados de contagem	18
2.2.1 Distribuição Poisson	19
2.2.2 Distribuição COM-Poisson	20
2.2.2.1 Regressão COM-Poisson e estimação dos parâmetros	22
2.2.3 Distribuição Poisson generalizada e Poisson generalizada restrita	24
2.2.3.1 Regressão Poisson generalizada restrita e estimação dos parâmetros	24
2.2.4 Distribuição <i>double</i> Poisson	26
2.2.4.1 Regressão <i>double</i> Poisson e estimação dos parâmetros	26
2.2.5 Distribuição contagem gama (“gamma count”)	27
2.2.5.1 Regressão contagem gama e estimação dos parâmetros	28
2.3 Estudo de caso - Efeito do aumento da biomassa de uma macrófita exótica invasora sobre a estrutura da comunidade de macrófitas aquáticas	29
2.4 Métodos	31
2.5 Resultados	33
2.6 Discussão	38
2.7 Considerações finais	39
Referências	39
3 Modelo de regressão <i>double</i> Poisson adaptado para considerar subdispersão e superdispersão conjuntamente em dados de contagem longitudinal	49
Resumo	49
3.1 Introdução	49
3.2 Métodos	50
3.2.1 Distribuição de probabilidade <i>double</i> Poisson	50
3.2.2 Modelo de regressão <i>double</i> Poisson com efeito aleatório	51
3.2.2.1 Inclusão da variável <i>dummy</i> para o parâmetro de dispersão	52
3.2.3 Estimação dos parâmetros	53
3.2.3.1 Predição de b_i via <i>empirical</i> Bayes	55
3.3 Análise dos resíduos	55
3.4 Estudo de simulação	58
3.5 Aplicação	63
3.5.1 Estudo de Motivação: análise de crises epiléticas	63
3.5.2 Formulação do modelo estatístico	67

3.5.3	Resultados e Discussão	68
3.5.4	Considerações finais	74
Referências		74

RESUMO

Modelos para dados de contagem não equidispersos com aplicação à ecologia e em estudos longitudinais

A obtenção de dados em forma de contagem é muito comum em pesquisas das áreas biológicas e o modelo de regressão Poisson é uma primeira alternativa de análise. No entanto, o modelo de Poisson exige independência, taxa constante de ocorrência e, com isso, variância igual a média. A condição de equidispersão limita seu uso em muitas aplicações, dado que essa relação nem sempre ocorre, podendo a variância ser menor (subdispersão) ou maior (superdispersão) do que a média. Como a ocorrência de dados superdispersos é mais comum do que dados subdispersos, houve um maior avanço em metodologias estatísticas na modelagem de dados com superdispersão. Porém, estudos com subdispersão têm surgido com maior regularidade, indicando que mais atenção é necessária para modelos que melhor expliquem essa estrutura de dados. No primeiro capítulo desta tese são abordados modelos alternativos para o problema da subdispersão, para o qual apresenta-se uma aplicação à ecologia. Como alternativas, foram propostos os modelos *double* Poisson, COM-Poisson, contagem gama e Poisson generalizada restrita. Essa classe de modelos é mais flexível para se trabalhar a subdispersão assim como a superdispersão (não equidispersão) presente nos dados, devido a presença de um parâmetro adicional. O fenômeno da não equidispersão também é comumente encontrado em dados de contagem longitudinais, isto é, quando a variável resposta em estudo é obtida repetidamente na mesma unidade amostral ao longo do tempo. Neste caso, modelos lineares generalizados mistos com a inclusão de efeitos aleatórios no preditor linear têm sido utilizados para acomodar a não equidispersão. No segundo capítulo é apresentada uma nova abordagem para análise de dados longitudinais com diferentes graus de dispersão, fundamentada na distribuição *double* Poisson em uma estrutura hierárquica. Para isso, utilizou-se uma variável *dummy* para particionar a sub- e superdispersão presente nos dados. A motivação deste estudo provém de um estudo clínico da literatura científica, que testa a eficiência de um medicamento na condição clínica de pacientes portadores de epilepsia. A estimação dos parâmetros foi realizada via máxima verossimilhança e o parâmetro de dispersão foi avaliado por meio do teste de razão de verossimilhança. A qualidade do ajuste dos modelos aos dados foi avaliada por meio da análise dos resíduos e, em particular, com o auxílio da técnica exploratória do gráfico meio normal de probabilidade.

Palavras-chave: Macrófitas aquáticas; Distribuição de probabilidade; Modelos mistos; *Half-normal plot*; Dados correlacionados; Progabide.

ABSTRACT

Models for non-equidispersed counting data with application to ecology and longitudinal studies

The use of count data is very common in biological areas research and Poisson regression model is the first alternative method to analyze this type of data. However, the Poisson model requires independence, constant rate of occurrence, and thus, variance equal to media. The equidispersion condition limits its use in many applications, since this relationship does not always occur, where the variance may be smaller (underdispersion) or greater (overdispersion) than the average. Because overdispersion data is more common than undispersed data, there has been a major advance in statistical methodologies in overdispersion data modeling. However, examples of underdispersion studies have emerged more regularly, indicating that more attention is needed for models that better explain this data structure. In the first chapter of this series, alternative models for the underdispersion problem are discussed, for which an application in ecology is presented. Alternatively, double Poisson, COM-Poisson, Gamma count and restricted generalized Poisson models were proposed. This class of models is more flexible to deal with both under- and overdispersion in the data, due to the presence of an additional parameter. The non-equidispersion phenomenon is also commonly found in longitudinal count data, *i.e.*, when a variable under study is repeatedly selected over time. In this case, generalized linear mixed models have been used, where additional normal random effects are included in the linear predictor to account for the non-equidispersion. The second chapter presents a new approach for longitudinal data analysis with different degrees of dispersion, based on double Poisson distribution in a hierarchical structure. For this, a *dummy* variable was used to partition the under- and overdispersion present in the data. The motivation of this study comes from a clinical study available in the scientific literature, which tested the efficacy of a drug in the clinical condition of patients with epilepsy. Parameter estimation was performed via maximum likelihood and the dispersion parameter was tested by the likelihood ratio test. The quality of the fit of the models was assessed by residual analysis and, in particular, with the help of the exploratory technique of the half-normal plot.

Keywords: Aquatic macrophytes; Probability distribution; Mixed models; Half-normal plot; Correlated data; Progabide.

1 INTRODUÇÃO

A análise de dados de contagem tem ocupado um lugar importante na estatística aplicada, uma vez que várias áreas do conhecimento apresentam estudos e pesquisas envolvendo este tipo específico de variável resposta. As contagens se referem ao número de eventos específicos que ocorrem em um intervalo de tempo ou espaço (área) e, por definição, consistem apenas em números inteiros positivos (Winkelmann, 2008; Cameron e Trivedi, 2013). O desenvolvimento de modelos de regressão para dados de contagem foi impulsionado pelos modelos lineares generalizados (MLG), sendo o modelo de regressão Poisson um caso especial, descrito pela primeira vez por Nelder e Wedderburn (1972) e detalhado em McCullagh e Nelder (1989). Winkelmann (2008) e Cameron e Trivedi (2013) são referências adicionais para modelos de contagem. O modelo de regressão Poisson exige equidispersão (variância igual a média). Entretanto, na prática, o que geralmente ocorre é a presença da superdispersão (variância maior do que a média) ou subdispersão (variância menor do que a média). O fenômeno da superdispersão em dados de contagem é comum em diversas áreas, tais como ecologia (Jay e Boveng, 2007; Richards, 2008), entomologia (Demétrio et al., 2014), epidemiologia (Lee et al., 2012), psicologia (Gardner et al., 1995), dentre outras. Conforme descrito por Demétrio et al. (2014), diferentes fontes de variação podem ser responsáveis pela presença da superdispersão nos dados, como variabilidade do material experimental, presença de *outliers*, omissão de covariáveis capazes de explicar a heterogeneidade, correlação entre as respostas individuais, amostragem por conglomerados e dados inflacionados de zeros. Em contrapartida, a análise de dados subdispersos ainda representa uma área de estudo em ascensão, uma vez que os mecanismos que levam ao fenômeno da subdispersão ainda não são muito conhecidos.

O problema da superdispersão quando ignorado pode resultar em um modelo que se ajustará pobremente aos dados. Como consequência, tem-se estimativas dos erros padrões dos parâmetros imprecisas, o que compromete as inferências do modelo, como interpretações incorretas e previsões inadequadas (Hinde e Demétrio, 1998). O modelo mais aplicado na literatura para dados de contagem superdispersos é o modelo binomial negativo. No entanto, sua aplicação é específica ao fenômeno da superdispersão (Winkelmann, 2008). O método de estimação da quasi-verossimilhança (Wedderburn, 1974) com suposição Poisson (aqui denominado modelo quasi-Poisson) pode ser aplicado a problemas tanto de subdispersão quanto superdispersão, porém esse método tem a desvantagem de ser não paramétrico. Assim, medidas de interesse científico não podem ser obtidas, tais como medidas de probabilidades para a distribuição das contagens e valores de verossimilhança para a comparação de modelos e submodelos. Jay e Boveng (2007) compararam os modelos binomial negativo e o modelo quasi-Poisson para dados de contagem superdispersos em um estudo referente a contagem de focas portuárias. Apesar de possuírem igual número de parâmetros e ambos serem usados para dados com superdispersão, os autores encontraram diferenças substanciais nas estimativas dos efeitos das covariáveis entre os modelos.

Dado que o modelo de regressão Poisson é restritivo, pesquisadores desenvolveram novas distribuições de probabilidade de modo a flexibilizar a análise de dados com sub- ou superdispersão. Modelos com dois parâmetros para dados de contagens, vêm sendo explorados na literatura, tais como os modelos COM-Poisson (Conway e Maxwell, 1962; Shmueli et al., 2005), *double* Poisson (Efron, 1986), contagem gama (*gamma-count*) (Winkelmann, 1995) e Poisson generalizada restrita (Famoye, 1993) são capazes de modelar a sub- e superdispersão, consideradas neste trabalho no termo único “não equidispersão”. Adicionalmente, o modelo Poisson aparece como um caso particular para os modelos descritos. Como estes modelos são paramétricos, pode-se explorar medidas de probabilidades bem como veros-

semelhança para efeito de comparação entre modelos. Contudo, uma desvantagem dos modelos COM-Poisson, *double* Poisson e contagem gama é que eles não têm média e a variância de forma fechada, sendo apenas obtidas por aproximações.

Como já salientado, existem muitos exemplos da utilização de modelos para dados de contagem superdispersos na literatura científica (Richards, 2008; Jamil et al., 2017; Moral et al., 2018). Em contrapartida, menos comuns são modelos direcionados a dados subdispersos, uma vez que, em geral, esse tipo de informação é descartada, pois acredita-se que sua estrutura é anômala (Sellers e Morris, 2017). Entretanto, nas últimas décadas modelos alternativos para esse problema têm despertado o interesse de pesquisadores. Os modelos contagem gama (Zeviani et al., 2014) e COM-Poisson (Ribeiro Jr et al., 2018), por exemplo, foram utilizados como modelos potenciais para análise de dados agrônômicos subdispersos, apresentando uma maior flexibilidade em relação ao modelo Poisson. Sellers e Morris (2017) compararam a performance dos modelos COM-Poisson, contagem gama, *double* Poisson e Poisson generalizada restrita em dois estudos de caso envolvendo dados subdispersos: um deles considerando o número de focos de greve (em períodos de 4 semanas) na indústria de mineração de carvão no Reino Unido entre 1948-1959 e, o outro, considerando o número de colisões entre locomotivas e veículos automotores selecionados aleatoriamente em cruzamentos na Coreia do Sul. Para o primeiro caso, o modelo *double* Poisson apresentou estimativas mais consistentes, enquanto para o segundo caso, os modelos *double* Poisson, contagem gama e COM-Poisson produziram estimativas concordantes. Ridout e Bebeas (2004) também verificaram a performance de diferentes distribuições, incluindo os modelos *double* Poisson e COM-Poisson, ajustados a dados com subdispersão, referentes ao número de surtos de greve (Sellers e Morris, 2017), número de ovos por ninhada de um pássaro no Reino Unido e o número de espermatozoides que penetraram cada óvulo em um experimento de polispermia. Recentemente, Bonat et al. (2016) propuseram o uso de uma nova classe de modelos baseada em modelos de dispersão fatorial Poisson-Tweedie (Jørgensen e Kokonendji, 2016). Essa classe de modelos também tem como objetivo unificar a análise de dados de contagem, seja para subdispersão, equidispersão ou superdispersão. No entanto, na presença da subdispersão a distribuição de probabilidade não é conhecida.

As classes de modelos para dados de contagem supracitadas foram inicialmente propostas no contexto dos estudos transversais, *i.e.* quando analisa-se uma ou múltiplas variáveis em um momento específico. Por outro lado, quando o objetivo é explorar medidas repetidas no tempo e observadas nos mesmos indivíduos durante um determinado período, trata-se de um planejamento longitudinal. Tais estudos permitem fazer inferências sobre efeitos e progressões temporais, os quais não seriam possíveis por meio dos estudos transversais. Segundo Singer e Andrade (1986), os principais objetivos em um estudo longitudinal são:

- i. Investigar se há diferença entre os tratamentos, fatores ou grupos quanto às distribuições médias de resposta, ou seja, verificar se há efeito de tratamento;
- ii. Averiguar os efeitos de tratamentos (ou grupos) ao padrão de variação da distribuição das respostas ao longo do tempo (ocasião), *i.e.*, verificar se há interação entre o tratamento e o fator tempo;
- iii. Verificar se há efeito do fator tempo comparando as distintas ocasiões em relação às distribuições médias de resposta.

Para dados longitudinais deve-se considerar que, medidas repetidas para o mesmo indivíduo tendem a estar correlacionadas. O uso de técnicas estatísticas que não levam em consideração a cor-

relação intraindividual das medidas pode acarretar em conclusões potencialmente imprecisas (Caruana et al., 2015). Assim, faz-se necessário considerar a correlação intraunidade experimental e a heterogeneidade entre as unidades experimentais, além da eventual possibilidade da existência da não equidispersão. Na prática, estudos longitudinais podem corriqueiramente apresentar unidades amostrais com diferentes níveis de variabilidade, podendo exibir unidades amostrais com subdispersão, equidispersão ou superdispersão.

Na presença de superdispersão, por exemplo, o modelo binomial negativo misto tem sido utilizado como modelo alternativo ao modelo Poisson misto (Zhang et al., 2018; Sung e Lee, 2018). Outros métodos alternativos estão disponíveis, como modelo Poisson generalizado (Toscas e Faddy, 2003), modelo Poisson generalizado INGARCH (Zhu, 2012), modelo regressão Weibull (Luyts et al., 2018), modelo COM-Poisson (Khan et al., 2018), dentre outros. Thall e Vail (1990) propuseram uma família de modelos de covariâncias para dados de contagem longitudinais, segundo os autores estes modelos consideram a superdispersão e a dependência entre as unidades experimentais. Molenberghs et al. (2007) propuseram um modelo linear generalizado misto (MLGM) para dados de contagem longitudinais usando efeitos aleatórios gama e normal, que tem os modelos binomial negativo padrão (efeitos aleatórios gama) e Poisson-normal (efeitos aleatórios normais) como casos especiais. Verifica-se, então, que trabalhos sobre dados de contagem longitudinais superdispersos são mais comuns na literatura científica. No entanto, quando se trata da análise de dados de contagem longitudinais subdispersos, o tema é ainda mais restrito do que em estudos transversais. Grunwald et al. (2011) em um estudo clínico, baseado no modelo hierárquico de Faddy (1997), propuseram um método que generaliza a distribuição Poisson para acomodar os fenômenos da subdispersão ou superdispersão. Em suma, apesar dos esforços em encontrar métodos satisfatórios para a análise de dados de contagem longitudinais, trabalhos analisando a subdispersão e/ou superdispersão presente nos dados (Grunwald et al., 2011; Niu e Ryan Cho, 2018; Tapia et al., 2019) ainda são emergentes na pesquisa científica.

Dessa forma, esta tese tem como objetivo principal explorar o desempenho dos principais modelos paramétricos para dados de contagem não equidispersos, tanto no contexto dos estudos transversais quanto longitudinais. Como objetivo específico, no âmbito do estudo transversal, foi examinado o efeito das covariáveis em relação ao número de grupos funcionais, dados de contagem com a presença de subdispersão, comparando-se a eficácia de quatro modelos distintos. Já no âmbito do estudo longitudinal, objetivou-se utilizar um modelo misto alternativo ao modelo Poisson misto, para explorar a não equidispersão presente nos dados por meio de um parâmetro adicional, a fim de que a variância seja controlada separadamente da média. Como contribuições específicas desta tese apresentam-se uma revisão dos modelos para fenômenos de não equidispersão e a proposta de segmentação dos indivíduos no estudo longitudinal, em subclasses com sub, super e equidispersão. Para tanto, esta tese apresenta-se organizada em dois capítulos.

O primeiro capítulo apresenta e compara a performance dos modelos COM-Poisson, *double Poisson*, contagem gama e Poisson generalizada restrita como métodos alternativos para a análise de dados subdispersos. Como motivação, tem-se um estudo que provém de um trabalho de campo realizado na bacia hidrográfica do rio Itanhaém, que objetivou avaliar a influência do aumento da biomassa de uma espécie exótica de capim aquático *Urochloa arrecta* sobre o número de grupos funcionais da comunidade de macrófitas nativas. A estimação dos parâmetros foi realizada via máxima verossimilhança e o parâmetro de dispersão foi testado por meio do teste de razão de verossimilhança. Para a validação dos

modelos utilizou-se o gráfico meio normal de probabilidade com envelope simulado (*half-normal plot*).

O segundo capítulo propõe a utilização do modelo *double* Poisson (Efron, 1986) com efeito aleatório, para a modelagem de dados de contagem longitudinais não equidispersos. A motivação deste trabalho provém de um conjunto de dados referente ao estudo clínico do uso terapêutico da droga Progabide para controlar o número de ataques epiléticos (Leppik et al., 1985). Os dados provenientes deste estudo foram utilizados na literatura (Thall e Vail, 1990; Breslow e Clayton, 1993; Diggle et al., 2002; Rigby e Stasinopoulos, 2005; Tapia et al., 2019) no âmbito da análise de dados longitudinais. Contudo, os estudos habitualmente levavam em consideração apenas a superdispersão presente neste conjunto de dados, uma vez que a variabilidade analisada era oriunda da variação interindividual, ou seja, entre os indivíduos estudados. No entanto, ao analisar o número de ataques epiléticos intraindividualmente, observa-se que os dados apresentam tanto subdispersão quanto superdispersão. Neste sentido, propõe-se utilizar um modelo de regressão *double* Poisson com intercepto aleatório, de modo a acomodar a subdispersão e superdispersão concomitantemente, adicionando uma variável *dummy* para o parâmetro de dispersão. As estimativas dos parâmetros foram obtidas pelo método da máxima verossimilhança. Também propõe-se a análise de resíduo utilizando o gráfico *half-normal plot*.

Referências

- Bonat, W. H., Jørgensen, B., Kokonendji, C. C., Hinde, J., e Demétrio, C. G. (2016). Extended poisson-tweedie: properties and regression models for count data. *arXiv preprint arXiv:1608.06888*.
- Breslow, N. E. e Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.
- Cameron, A. C. e Trivedi, P. (2013). *Regression analysis of count data*, volume 53. Cambridge University Press.
- Caruana, E. J., Roman, M., Hernández-Sánchez, J., e Solli, P. (2015). Longitudinal studies. *Journal of thoracic disease*, 7(11):E537.
- Conway, R. W. e Maxwell, W. L. (1962). Network dispatching by the shortest-operation discipline. *Operations Research*, 10(1):51–73.
- Demétrio, C. G. B., Hinde, J., e Moral, R. A. (2014). Models for overdispersed data in entomology. In *Ecological modelling applied to entomology*, pages 219–259. Springer.
- Diggle, P. J., Heagerty, P., Liang, K. Y., e Zeger, S. L. (2002). *Analysis of longitudinal data*. Oxford University Press.
- Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81(395):709–721.
- Faddy, M. (1997). Extended poisson process modelling and analysis of count data. *Biometrical Journal*, 39(4):431–440.
- Famoye, F. (1993). Restricted generalized poisson regression model. *Communications in Statistics-Theory and Methods*, 22(5):1335–1354.

- Gardner, W., Mulvey, E. P., e Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed poisson, and negative binomial models. *Psychological bulletin*, 118(3):392.
- Grunwald, G. K., Bruce, S. L., Jiang, L., Strand, M., e Rabinovitch, N. (2011). A statistical model for under-or overdispersed clustered and longitudinal count data. *Biometrical Journal*, 53(4):578–594.
- Hinde, J. e Demétrio, C. G. (1998). Overdispersion: models and estimation. *Computational statistics & data analysis*, 27(2):151–170.
- Jamil, S. A. M., Abdullah, M. A. A., Kek, S. L., Nor, M. E., Mohamed, M., e Ismail, N. (2017). Detecting overdispersion in count data: A zero-inflated poisson regression analysis. In *Journal of Physics: Conference Series*, volume 890, page 012170. IOP Publishing.
- Jay, M. V. H. e Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772.
- Jørgensen, B. e Kokonendji, C. C. (2016). Discrete dispersion models and their tweedie asymptotics. *AStA Advances in Statistical Analysis*, 100(1):43–78.
- Khan, M. N., Jowaheer, V., Sunecher, Y., Bourguignon, M., et al. (2018). Modeling longitudinal inma (1) with com–poisson innovation under non-stationarity: application to medical data. *Computational and Applied Mathematics*, 37(4).
- Lee, J., Han, G., Fulp, W., AR, e others Giuliano (2012). Analysis of overdispersed count data: application to the human papillomavirus infection in men (him) study. *Epidemiology and infection*, 140(6):1087–1094.
- Leppik, I., Dreifuss, F., Bowman, T., Santilli, N., Jacobs, M., Crosby, C., Cloyd, J., Stockman, J., Graves, N., Sutula, T., et al. (1985). A double-blind crossover evaluation of progabide in partial seizures. *Neurology*, 35(4).
- Luyts, M., Molenberghs, G., Verbeke, G., Matthijs, K., Ribeiro Jr, E. E., Demétrio, C. G., e Hinde, J. (2018). A weibull-count approach for handling under-and overdispersed longitudinal/clustered data structures. *Statistical Modelling*, page 1471082X18789992.
- McCullagh, P. e Nelder, J. A. (1989). *Generalized linear models (Monographs on statistics and applied probability 37)*.
- Molenberghs, G., Verbeke, G., e Demétrio, C. G. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime data analysis*, 13(4):513–531.
- Moral, R. A., Hinde, J., Demétrio, C. G. B., Reigada, C., e Godoy, W. A. C. (2018). Models for jointly estimating abundances of two unmarked site-associated species subject to imperfect detection. *Journal of Agricultural, Biological and Environmental Statistics*, 23(1):20–38.
- Nelder, J. e Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.
- Niu, X. e Ryan Cho, H. (2018). Efficient regression modeling for correlated and overdispersed count data. *Communications in Statistics-Theory and Methods*, pages 1–14.

- Ribeiro Jr, E. E., Zeviani, W. M., Bonat, W. H., Demétrio, C. G. B., e Hinde, J. (2018). Reparametrization of com-poisson regression models with applications in the analysis of experimental data. *arXiv preprint arXiv:1801.09795*.
- Richards, S. A. (2008). Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, 45(1):218–227.
- Ridout, M. S. e Besbeas, P. (2004). An empirical model for underdispersed count data. *Statistical Modelling*, 4(1):77–89.
- Rigby, R. A. e Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554.
- Sellers, K. F. e Morris, D. S. (2017). Underdispersion models: Models that are under the radar. *Communications in Statistics-Theory and Methods*, 46(24):12075–12086.
- Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., e 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.
- Singer, J. M. e Andrade, D. F. (1986). Análise de dados longitudinais. *SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA*, 7.
- Sung, Y. e Lee, K. (2018). Negative binomial loglinear mixed models with general random effects covariance matrix. *Communications for Statistical Applications and Methods*, 25(1):61–70.
- Tapia, A., Giampaoli, V., Diaz, M. P., e Leiva, V. (2019). Sensitivity analysis of longitudinal count responses: a local influence approach and application to medical data. *Journal of Applied Statistics*, 46(6):1021–1042.
- Thall, P. F. e Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, pages 657–671.
- Toscas, P. J. e Faddy, M. J. (2003). Likelihood-based analysis of longitudinal count data using a generalized poisson model. *Statistical Modelling*, 3(2):99–108.
- Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61(3):439–447.
- Winkelmann, R. (1995). Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics*, 13(4):467–474.
- Winkelmann, R. (2008). *Econometric analysis of count data*. Springer.
- Zeviani, W. M., Ribeiro Jr, P. J., Bonat, W. H., Shimakura, S. E., e Muniz, J. A. (2014). The gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics*, 41(12):2616–2626.
- Zhang, X., Pei, Y.-F., Zhang, L., Guo, B., Pendegraft, A., Zhuang, W., e Yi, N. (2018). Negative binomial mixed models for analyzing longitudinal microbiome data. *Frontiers in microbiology*, 9:1683.

Zhu, F. (2012). Modeling overdispersed or underdispersed count data with generalized poisson integer-valued garch models. *Journal of Mathematical Analysis and Applications*, 389(1):58–71.

2 MODELOS PARA DADOS DE CONTAGEM NÃO EQUIDISPERSOS: UMA APLICAÇÃO EM DADOS ECOLÓGICOS COM SUBDISPERSÃO

Resumo

Em pesquisas nas áreas biológicas e agrárias frequentemente utilizam-se dados de contagem como variável resposta, que assumem valores inteiros e positivos. O modelo de regressão Poisson aparece como candidato principal para esse tipo de variável resposta, com a pressuposição de equidispersão, isto é, média e variância iguais. No entanto, essa relação nem sempre ocorre, podendo a variância ser maior ou menor do que a média, fenômenos conhecidos como superdispersão e subdispersão, respectivamente. Esses fenômenos, quando desconsiderados, podem acarretar em desvios padrões imprecisos, promovendo um modelo de ajuste pobre aos dados, bem como interpretações e previsões incorretas. Neste sentido, é importante a utilização de análises que possam avaliar de forma satisfatória o comportamento da variável em estudo. Verifica-se uma maior ênfase na literatura para os problemas de superdispersão, quando comparado às análises de dados subdispersos. Neste contexto, esse trabalho apresenta os modelos COM-Poisson, Poisson generalizada restrita, *double* Poisson e contagem gama, como modelos alternativos, mais flexíveis para problemas de contagem associados a superdispersão e/ou subdispersão, uma vez que eles tem um parâmetro adicional que permite a flexibilidade da análise desses dados. Como motivação, tem-se um estudo que provém de um trabalho de campo realizado na bacia hidrográfica do rio Itanhaém, que objetivou avaliar a influência do aumento da biomassa de uma espécie exótica de capim aquático *Urochloa arrecta* sobre o número de grupos funcionais da comunidade de macrófitas nativas. A estimação dos parâmetros foi realizada via máxima verossimilhança e o parâmetro de dispersão foi avaliado por meio do teste de razão de verossimilhança, indicando a presença de subdispersão nos dados. Para a validação dos modelos utilizou-se o gráfico meio normal de probabilidade (*half-normal plot*). Verificou-se que há uma redução no número de grupos funcionais quando há um aumento da quantidade de biomassa. Este resultado possui uma implicação biológica importante, pois o modelo sugere que o crescimento da espécie exótica, expresso em biomassa, inibe o desenvolvimento de outros grupos. Dessa forma, o conhecimento das respostas de uma comunidade biológica em relação ao comportamento de uma espécie exótica invasora é importante no que diz respeito à identificação de medidas de manejo adequadas.

Palavras-chave: Poisson; COM-Poisson; Poisson generalizada restrita; *Double* Poisson; Contagem gama; *Urochloa arrecta*; *Half-normal plot*; Superdispersão.

2.1 Introdução

Em estudos biológicos e agrônômicos é comum o uso de dados de contagem como variável resposta, sejam eles em testes de germinação, análise microbiológica de solos, avaliação entomológica, teste de toxicidade, dentre outros. Esse tipo de dados é frequentemente analisado utilizando a distribuição normal. Nesse sentido, muitos métodos estatísticos recomendam transformar os dados para obter a linearidade, homogeneidade das variâncias e normalidade da variável resposta (O'Hara e Kotze, 2010). No entanto, a análise de dados discretos pode ser feita sem transformação, por meio dos modelos lineares generalizados (MLG). O modelo de regressão Poisson, que é um caso particular de um MLG, é o mais utilizado quando a variável resposta é obtida na forma de contagem. Uma característica essencial

do modelo de regressão Poisson é a equidispersão, isto é, média e variância iguais. No entanto, essa relação nem sempre ocorre, podendo a variância ser maior ou menor do que a média, fenômenos esses conhecidos como superdispersão e subdispersão, respectivamente, chamados simplificados de "não equidispersão". As causas da não equidispersão surgem devido a uma ou mais possibilidades, tais como variabilidade do material experimental, correlação entre as respostas individuais, amostragem por conglomerados ou estrutura agrupada, ausência de variáveis explicativas capaz de explicar a variabilidade nos dados (Kokonendji, 2014; Demétrio et al., 2014).

Atualmente a análise de dados superdispersos tem recebido considerável atenção por parte da comunidade científica, uma vez que esse fenômeno é frequente em estudos biológicos e agrônômicos (Jay e Boveng, 2007; Richards, 2008; Jamil et al., 2017; Moral et al., 2018). Por outro lado, menos comuns são modelos alternativos e avaliações disponíveis para a subdispersão. Ocasionalmente dados subdispersos são descartados, por se acreditar que sua estrutura é uma anomalia (Sellers e Morris, 2017). Devido a singularidade e raridade, o estudo da subdispersão é menos consolidado quando comparado com a superdispersão. No entanto, algumas classes de modelos já existem para essa estrutura de dados. Alguns exemplos disponíveis são *condensed* Poisson (Sellers e Morris, 2017), distribuição contagem gama (Winkelmann, 1995; Zeviani et al., 2014), Poisson generalizada (Famoye et al., 2004), COM-Poisson e *double* Poisson (Ridout e Besbeas, 2004; Wu et al., 2013). Esses modelos são flexíveis para dados de contagem não equidispersos, pois possuem um ou dois parâmetros adicionais que permitem acomodar a não equidispersão.

A motivação deste trabalho baseia-se em uma pesquisa ecológica desenvolvida em uma microbacia costeira, na qual se verificou a relação entre o aumento da biomassa da macrófita aquática exótica invasora *Urochloa arrecta* (Hack. ex. T. Durand and Schinz) Morrone e Zuloaga e o número de grupos funcionais da assembleia de macrófitas nativas. Esta espécie é uma Poaceae nativa da África, conhecida por invadir e infestar comunidades aquáticas em zonas tropicais e subtropicais (Reinert et al., 2007; Pott et al., 2011) em todo o mundo. Esta planta tem alto potencial de regeneração e estudos têm mostrado que ela pode competir com outras espécies e tem efeitos negativos sobre a riqueza de espécies e a diversidade funcional (Michelan et al., 2010). O estudo utilizou como variável resposta a diversidade funcional, que diz respeito ao número (riqueza) de grupos funcionais encontrado em determinado ambiente, isto é, o conjunto de espécies que tem forma de vida similares (Tilman, 2001). A riqueza de grupos funcionais pode ser um parâmetro importante para prever mudanças ambientais, uma vez que eles desempenham papéis semelhantes no funcionamento do ecossistema. Neste sentido, no contexto de invasão por espécies exóticas, a diversidade de grupos funcionais pode ser um dos fatores que pode influenciar a habilidade destas espécies em invadir as comunidades naturais. Elton (1958) foi um dos pioneiros a observar que comunidades de plantas com alta diversidade deveriam ser mais resistentes a invasões por novas espécies.

Desta forma, este trabalho objetivou comparar a performance de quatro modelos paramétricos, COM-Poisson, Poisson generalizado restrito, contagem gama e *double* Poisson, como métodos alternativos para dados de contagem não equidispersos e, particularmente, úteis para modelar dados ecológicos.

2.2 Revisão de modelos de dados de contagem

O modelo de regressão linear clássico é de grande importância para descrever uma série de fenômenos aleatórios, entretanto, se limita à suposição de normalidade para a variável resposta, bem como linearidade, homogeneidade de variâncias e independência dos erros. Quando alguma suposição

não é alcançada, um tipo de transformação pode ser sugerida, sendo a transformação de Box-Cox (Box e Cox, 1964) a mais conhecida.

Com o objetivo de expandir a classe de modelos lineares, Nelder e Wedderburn (1972) propuseram os MLG para os quais não há necessidade de aplicar qualquer tipo de transformação na variável resposta em busca de atingir as pressuposições de normalidade e variância constante. A variável resposta pode ser discreta ou contínua e podem assumir diferentes distribuições de probabilidade pertencentes à família exponencial. Para uma variável resposta contínua simétrica, uma escolha plausível é a distribuição normal, enquanto que para dados contínuos assimétricos pode-se optar pelas distribuições gama ou normal inversa. Para dados de proporção, a distribuição binomial mostra-se aconselhável, enquanto que para dados de contagem pode-se utilizar a distribuição Poisson ou binomial negativa (McCullagh e Nelder, 1989). Nelder e Wedderburn (1972) também propuseram um processo iterativo padrão para a estimação dos parâmetros do MLG e apresentaram a *deviance* como medida para avaliação da qualidade do ajuste do modelo bem como para a seleção de modelos.

Os MLG apresentam três componentes: componente aleatório, sistemático e função de ligação (McCullagh e Nelder, 1989, página 26).

- i Componente aleatório: responsável por identificar a variável resposta e sua distribuição de probabilidade. Sejam $Y_i, i = 1, 2, \dots, n$, n observações independentes da variável resposta e sua distribuição de probabilidade pertencente à família exponencial escrita na forma canônica:

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}, \quad (2.1)$$

em que $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções específicas e $a(\phi) > 0$, e θ_i é denominado de parâmetro natural ou canônico. Usualmente $a(\phi) = \phi/w_i$, no qual w_i são os pesos *a priori* (em geral $w_i = 1$) e ϕ representa um parâmetro de dispersão constante e suporte conhecido. A média e a variância de Y_i são dadas por: $E(Y_i) = b'(\theta_i) = \mu_i$ e $\text{Var}(Y_i) = a(\phi)b''(\theta_i) = a(\phi)V(\mu_i)$. Dessa forma, $b'(\theta_i)$ e $b''(\theta_i)$ são as derivadas de 1ª e 2ª ordens em relação à θ_i , respectivamente. A função $V(\mu_i)$ depende apenas de μ_i , sendo conhecida como função de variância.

- ii Componente sistemático: responsável por especificar as variáveis explicativas $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$, em que $x_{i1} = 1$ para todo i , ao preditor linear η_i dado por:

$$\eta_i = \sum_{j=1}^p x_{ij} \beta_j, \quad j = 1, 2, \dots, p$$

no qual $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$ é um vetor de p parâmetros desconhecidos, ou, sob a forma matricial: $\eta = \mathbf{X}^T \beta$ em que $\mathbf{X}^T$ é uma matriz $n \times p$.

- iii Função de ligação: responsável por relacionar o componente sistemático ao componente aleatório, sendo uma função monótona e diferenciável, ou seja:

$$\eta_i = g(\mu_i) \quad \Rightarrow \quad \mu_i = g^{-1}(\eta_i).$$

2.2.1 Distribuição Poisson

Essa distribuição é de grande importância para a análise de dados em forma de contagens. Sendo assim, considere $Y_i (i = 1, 2, \dots, n)$ uma variável aleatória com distribuição Poisson $Y_i \sim \text{Poisson}(\lambda_i)$, $\lambda_i > 0$, com função de probabilidade de Y_i dado por:

$$P(Y = y_i | \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \dots,$$

Verifica-se que a distribuição Poisson pertence à família exponencial canônica, ou seja:

$$f(y_i; \theta_i, \phi) = \exp \{y_i \ln(\lambda_i) - \lambda_i - \ln(y_i!)\}.$$

Logo, assumindo que $\theta_i = \ln(\lambda_i)$ é o parâmetro canônico e $b(\theta_i) = \exp(\theta_i) = \lambda_i$ é uma função monótona e derivável, verifica-se que $E(Y_i) = b'(\theta_i) = \lambda_i$ e $\text{Var}(Y_i) = a(\phi)b''(\theta_i) = a(\phi)V(\lambda_i)$, e $V(\lambda_i)$ é a função de variância e $a(\phi) = 1$. Pode-se observar que $E(Y_i) = \text{Var}(Y_i)$ (equidispersão). Essa condição da distribuição Poisson faz com que na prática sua aplicação se torne limitada, uma vez que experimentalmente dificilmente se tem média igual a variância. Assim, uma forma simples de identificar a presença da não equidispersão se dá por meio da análise descritiva, ao comparar a média amostral com a variância amostral, ou seja, quando $s_y^2 > \hat{y}$ ou $s_y^2 < \hat{y}$ pode se ter indícios de um problema de superdispersão e subdispersão, respectivamente.

Outra forma de identificar a subdispersão ou superdispersão é estimar o parâmetro de dispersão σ^2 . Quando $\hat{\sigma}^2 = 1$ representa equidispersão, quando $0 < \hat{\sigma}^2 < 1$ demonstra a presença de subdispersão e, por outro lado, quando $\hat{\sigma}^2 > 1$ representa a superdispersão. O parâmetro $\hat{\sigma}^2$ pode ser estimado por:

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

em que X^2 é a estimativa da estatística de Pearson ajustada à regressão Poisson, com n observações e p parâmetros. Na seção 2.4 é apresentado um teste formal para a não equidispersão.

Cameron e Trivedi (2013) salientam que em um modelo de regressão Poisson a média e a variância estão condicionadas às variáveis explicativas:

$$E[Y_i | \mathbf{x}_i] = V[Y_i | \mathbf{x}_i] = \exp(\mathbf{x}_i^T \boldsymbol{\beta}) = \lambda_i. \quad (2.2)$$

Dessa forma, se a variância da amostra for mais do que o dobro da sua média amostral, é provável que os dados permaneçam superdispersos após a inclusão das variáveis explicativas (Cameron e Trivedi, 2013). Isto é particularmente verdadeiro para dados de estudos transversais (“cross-section”), para os quais as variáveis explicativas geralmente explicam menos da metade da variação nos dados. Se a variância da amostra for menor do que a média da amostra, os dados necessariamente são ainda mais subdispersos, mesmo que as variáveis explicativas sejam incluídas (Cameron e Trivedi, 2013). A relação 2.2 é chamada de função média exponencial e na literatura estatística é chamada de modelo log linear, porque o logaritmo da média é condicional e linear nos parâmetros, logo $\ln(E[Y_i | \mathbf{x}_i]) = \mathbf{x}_i^T \boldsymbol{\beta}$. Pode-se observar na equação 2.2 que a variância condicional não é constante e, portanto, a regressão Poisson é intrinsecamente heterocedástica.

Como a proposta de estudo é a subdispersão, apresenta-se sucintamente os principais modelos para tal fenômeno.

2.2.2 Distribuição COM-Poisson

A distribuição COM-Poisson (CP) foi proposta por Conway e Maxwell (1962) e é uma generalização da distribuição Poisson com adição de um parâmetro, para flexibilizar problemas relativos à não

equidispersão. Shmueli et al. (2005) estudaram suas propriedades estatísticas e aplicações. A função de probabilidade CP é definida por:

$$P(Y_i = y_i | \lambda_i, \nu) = \frac{\lambda_i^{y_i}}{(y_i!)^\nu} \frac{1}{Z(\lambda_i, \nu)}, \quad y_i = 0, 1, 2, \dots, \quad i = 1, 2, \dots, n, \quad (2.3)$$

$$Z(\lambda_i, \nu) = \sum_{j=0}^{\infty} \frac{\lambda_i^j}{(j!)^\nu}, \quad \lambda_i > 0, \quad \nu \geq 0$$

em que $Z(\lambda_i, \nu)$ é uma constante de normalização que atende às condições para uma distribuição de probabilidade. A série $\lambda_i^j / (j!)^\nu$ converge para qualquer $\lambda_i > 0$ e $\nu > 0$, uma vez que a razão de dois termos subsequentes da série λ_i / j^ν tende a 0 com $j \rightarrow \infty$. No entanto, se $\nu = 0$ e $\lambda_i \geq 1$, a constante de normalização não converge e a distribuição de probabilidade não é definida. O parâmetro ν é considerado o parâmetro de dispersão e quando $0 < \nu < 1$, tem-se $\text{Var}(Y_i) > E(Y_i)$, $\nu = 1$ a distribuição 2.3 reduz-se à Poisson e para $\nu > 1$, $\text{Var}(Y_i) < E(Y_i)$. É importante salientar que a não equidispersão é uma característica intrínseca dos dados ocasionada por diversas razões. Assim, a distribuição CP por conter um parâmetro adicional, é mais flexível do que a distribuição Poisson para análise de dados de contagem. Além da distribuição Poisson como caso particular, tem-se:

- i) Para $\nu \rightarrow \infty$, $Z(\lambda_i, \nu) \rightarrow 1 + \lambda_i$ a distribuição CP se aproxima da distribuição Bernoulli com $P(Y_i = 1) = \lambda_i(1 + \lambda_i)^{-1}$;
- ii) $\nu = 0$, $0 < \lambda_i < 1$, $\sum_{j=0}^{\infty} \lambda_i^j = 1 / (1 - \lambda_i)$ é uma série geométrica, e portanto, resulta na distribuição de probabilidade geométrica, definida por: $P(Y_i = y_i) = \lambda_i^{y_i} (1 - \lambda_i)$.

A distribuição CP tem a vantagem de permitir uma maior flexibilidade entre as razões de probabilidades sucessivas, quando comparada com a distribuição Poisson, possibilitando uma redução não linear das relações de probabilidades sucessivas, na forma:

$$\frac{P(Y_i = y_i + 1)}{P(Y_i = y_i)} = \frac{y_i^\nu}{\lambda_i},$$

permitindo caudas mais leves ou pesadas (Shmueli et al., 2005). A distribuição CP não tem forma fechada para média e variância, porém podem ser calculadas por meio do método dos momentos. O r -ésimo momento é definido por:

$$E(Y_i^{r+1}) = \begin{cases} \lambda_i [E(Y_i + 1)]^{1-\nu}, & r = 0 \\ \lambda_i \frac{\partial}{\partial \lambda_i} E(Y_i^r) + E(Y_i) E(Y_i^r), & r > 0. \end{cases}$$

Uma outra alternativa para se calcularem a média e a variância é pela definição do valor esperado para variáveis aleatórias discretas, ou seja:

$$E(Y_i) = \sum_{i=1}^{\infty} y_i P(Y_i = y_i), \quad E(Y_i^2) = \sum_{i=1}^{\infty} y_i^2 P(Y_i = y_i) \quad \text{e} \quad \text{Var}(Y_i) = E(Y_i^2) - [E(Y_i)]^2.$$

Minka et al. (2003) propuseram uma aproximação assintótica para $Z(\lambda_i, \nu)$ definida por:

$$Z(\lambda_i, \nu) = \frac{\exp\left(\nu \lambda_i^{\frac{1}{\nu}}\right)}{\lambda_i^{\frac{\nu-1}{2\nu}} (2\pi)^{\frac{\nu-1}{2}} \nu^{\frac{1}{2}}}, \quad (2.4)$$

que é exata quando $\lambda_i > 10^v$. Ao utilizar a aproximação assintótica para $Z(\lambda_i, v)$ é possível obter uma aproximação para primeiro momento. Assim,

$$E(Y_i) = \lambda_i \frac{\partial \log Z(\lambda_i, v)}{\partial \lambda_i} \approx \lambda_i^{1/v} - \frac{v-1}{2v}. \quad (2.5)$$

Guikema e Goffelt (2008), assim como Sellers e Shmueli (2008), apresentam outra forma de obter o primeiro momento e a variância:

$$E(Y_i) = \frac{\partial \log Z(\lambda_i, v)}{\partial \log(\lambda_i)} \approx \lambda_i^{1/v} - \frac{v-1}{2v} \quad \text{e} \quad \text{Var}(Y_i) = \frac{\partial E(Y_i)}{\partial \log(\lambda_i)} \approx \frac{1}{v} \lambda_i^{1/v}.$$

Entretanto, essas aproximações são satisfatórias quando $v \leq 1$ ou $\lambda_i > 10^v$ (Shmueli et al., 2005).

2.2.2.1 Regressão COM-Poisson e estimação dos parâmetros

Sejam Y_i , $i = 1, 2, \dots, n$, n observações independentes da variável resposta proveniente de um estudo de contagem, e $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$, valores das p variáveis explicativas associadas ao i -ésimo indivíduo e $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$ um vetor de p parâmetros desconhecidos. A regressão CP com função de ligação logarítmica é dada por:

$$\eta_i = \log(\lambda_i) = \sum_{j=1}^p x_{ij} \beta_j = \mathbf{x}_i^T \beta,$$

assim, modelando η_i ao invés do próprio λ_i como uma função linear das variáveis explicativas. É importante salientar que é possível utilizar outro tipo de função de ligação, desde que seja uma função monótona e diferenciável, e portanto, $\eta_i = g(\lambda_i) \rightarrow \lambda_i = g^{-1}(\eta_i)$.

Como a distribuição CP não tem forma fechada para os momentos, pode-se usar a aproximação assintótica para $Z(\lambda_i, v)$, dada em 2.4 (Minka et al., 2003). Dessa forma, pode-se reformular a equação 2.3 e incluir os efeitos das covariáveis Jowaheer e Khan (2009), ou seja:

$$P(Y_i = y_i | \mathbf{x}_i) = \frac{\exp(\mathbf{x}_{ij} \beta_j)^{y_i} \exp(\mathbf{x}_{ij} \beta_j)^{\frac{v-1}{2v}} (2\pi)^{\frac{v-1}{2}} v^{0.5}}{(y_i!)^v \exp\left(v \exp\left(\frac{\mathbf{x}_{ij} \beta_j}{v}\right)\right)} \quad (2.6)$$

e para estimar os parâmetros β_j e v utiliza-se o método da máxima verossimilhança. O logaritmo natural da função de verossimilhança do modelo 2.6 é dado por:

$$l(\beta_j, v) = \sum_{i=1}^n y_i (x_{ij} \beta_j) - v \sum_{i=1}^n \log(y_i!) + \frac{v-1}{2v} \sum_{i=1}^n (x_{ij} \beta_j) - v \sum_{i=1}^n \exp\left(\frac{x_{ij} \beta_j}{v}\right) + \frac{n(v-1)}{2} \log(2\pi) + \frac{n}{2} \log(v)$$

e as funções escores são obtidas derivando o logaritmo natural da função de verossimilhança em função de cada um dos respectivos parâmetros. Portanto:

$$U(\beta_j) = \frac{\partial l(\beta_j, v)}{\partial \beta_j} = \sum_{i=1}^n y_i x_{ij} - \sum_{i=1}^n x_{ij} \exp\left(\frac{x_{ij} \beta_j}{v}\right) + 0.5 \left(\frac{v-1}{v}\right) \sum_{i=1}^n x_{ij} \quad (2.7)$$

e

$$U(v) = \frac{\partial l(\beta_j, v)}{\partial v} = \sum_{i=1}^n \left[-\log(y_i!) + \frac{x_{ij} \beta_j}{2v^2} - \exp\left(\frac{x_{ij} \beta_j}{v}\right) + \frac{x_{ij} \beta_j}{v} \exp\left(\frac{x_{ij} \beta_j}{v}\right) \right] + \frac{n}{2} \log(2\pi) + \frac{n}{2v}. \quad (2.8)$$

Dessa forma, o vetor escore de dimensão $(p+1) \times 1$ é definido por:

$$U(\beta_j, v) = \left(\frac{\partial l(\beta_j, v)}{\partial \beta_1}, \frac{\partial l(\beta_j, v)}{\partial \beta_2}, \dots, \frac{\partial l(\beta_j, v)}{\partial \beta_j}, \frac{\partial l(\beta_j, v)}{\partial v} \right)^T.$$

Para obter as estimativas de máxima verossimilhança igualam-se as expressões 2.7 e 2.8 a zero e encontram-se as raízes do sistema de equações em relação a β_j e v . Como este sistema não tem solução analítica fechada, faz-se necessário o uso de procedimentos numéricos, como o método numérico iterativo de Newton-Raphson, para obter as estimativas dos parâmetros. Para implementar este algoritmo, faz-se necessário obter a segunda derivada, ou seja, encontrar a matriz hessiana, mas como a distribuição CP não tem média na forma fechada, utiliza-se a matriz de informação observada. Assim, tem-se:

$$\begin{aligned} \frac{\partial^2 l(\beta_j, v)}{\partial \beta_j^2} &= - \sum_{i=1}^n \left[\frac{x_{ij}^2}{v} \exp(x_{ij} \beta_j) \right], \\ \frac{\partial^2 l(\beta_j, v)}{\partial \beta_j \partial v} &= \sum_{i=1}^n \left[\frac{x_{ij}}{2v^2} + \frac{x_{ij}^2 \beta_j}{v^2} \exp\left(\frac{x_{ij} \beta_j}{v}\right) \right] \end{aligned} \quad (2.9)$$

e

$$\frac{\partial^2 l(\beta_j, v)}{\partial v^2} = -\frac{n}{2v^2} + \sum_{i=1}^n \left[-\frac{x_{ij} \beta_j}{v^3} + \frac{(x_{ij} \beta_j)^2}{v^3} \exp\left(\frac{x_{ij} \beta_j}{v}\right) \right]. \quad (2.10)$$

Para finalizar o algoritmo de estimação dos parâmetros, é necessário resolver de forma iterativa a equação 2.11, em que $\beta_{j(0)}$ e $v_{(0)}$ são os valores iniciais dos parâmetros a serem estimados. Valores iniciais plausíveis para $\beta_{j(0)}$ podem ser extraídos do modelo Poisson e $v_{(0)} = 1$, $I_0(\beta_j, v)$ denota a matriz hessiana observada.

$$\begin{pmatrix} \beta_{1(r+1)} \\ \vdots \\ \beta_{j(r+1)} \\ v_{(r+1)} \end{pmatrix} = \begin{pmatrix} \beta_{1(r)} \\ \vdots \\ \beta_{j(r)} \\ v_{(r)} \end{pmatrix} - \underbrace{\begin{pmatrix} \frac{\partial^2 l(\beta_j, v)}{\partial \beta_1^2} & \frac{\partial^2 l(\beta_j, v)}{\partial \beta_1 \partial \beta_2} & \dots & \frac{\partial^2 l(\beta_j, v)}{\partial \beta_1 \partial v} \\ \frac{\partial^2 l(\beta_j, v)}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 l(\beta_j, v)}{\partial \beta_2^2} & \dots & \frac{\partial^2 l(\beta_j, v)}{\partial \beta_2 \partial v} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\beta_j, v)}{\partial \beta_1 \partial v} & \frac{\partial^2 l(\beta_j, v)}{\partial \beta_2 \partial v} & \dots & \frac{\partial^2 l(\beta_j, v)}{\partial v^2} \end{pmatrix}}_{I_0(\beta_j, v)} \begin{pmatrix} \frac{\partial l(\beta_j, v)}{\partial \beta_1} \\ \vdots \\ \frac{\partial l(\beta_j, v)}{\partial \beta_j} \\ \frac{\partial l(\beta_j, v)}{\partial v} \end{pmatrix} \quad (2.11)$$

Esse procedimento deve ser executado considerando a r -ésima iteração, até que um critério de convergência seja atingido. Um critério plausível de convergência pode ser:

$$\sum_{j=1}^p \left(\frac{\beta_j^r - \beta_j^{(r+1)}}{\beta_j^r} \right) + \frac{v^r - v^{(r+1)}}{v^r} < \xi,$$

assumindo para ξ um valor satisfatoriamente pequeno. Para construir os intervalos de confiança assintóticos e realizar os testes de hipóteses dos parâmetros, faz-se necessário o cálculo da matriz de informação observada (hessiana), no ponto encontrado.

Os termos fora da diagonal são não nulos, ou seja, os parâmetros não são ortogonais. A ortogonalidade é uma propriedade vantajosa, pois simplifica as inferências dos parâmetros. A função escore e a matriz de informação observada (matriz de informação de Fischer) exibem dois resultados importantes que são $E[U(\theta)] = 0$ e $\text{Var}[U(\theta)] = E[I_0(\theta)] = I_E[\theta]$, em que I_E é a matriz de informação esperada e $\theta = (\beta_j, v)$ o vetor de parâmetros. Considerando um problema de estimação regular, se θ é o verdadeiro vetor de parâmetros, assintoticamente tem-se:

$$\hat{\theta} \overset{a}{\sim} NM_d(\theta, I_o(\theta)^{-1}),$$

isto é, a distribuição assintótica dos parâmetros de $\hat{\theta}$ é uma normal multivariada com matriz de variâncias e covariâncias dada pela inversa da matriz de informação observada. Pode-se construir os intervalos de confiança $100(1 - \alpha)\%$ para os parâmetros na forma $\hat{\theta}_i \pm z_{\alpha/2} \sqrt{\text{Var}(\hat{\theta}_{ii})}$, sendo z o quantil da distribuição normal e $\text{Var}(\hat{\theta}_{ii})$ são os elementos da diagonal da matriz $I_o(\hat{\theta})^{-1}$ de variâncias e covariâncias. Os intervalos construídos, desta forma, serão denominados intervalos de Wald.

Os valores preditos são computados com o auxílio da equação 2.5 e é importante ressaltar que a regressão é formulada no parâmetro λ_i , e não na média da variável resposta, como é usual em MLG, ou seja, modela a influência das covariáveis sobre λ_i .

2.2.3 Distribuição Poisson generalizada e Poisson generalizada restrita

A distribuição Poisson generalizada (PG) tem sua função de probabilidade dada por:

$$P(Y_i = y_i | \mu_i, \gamma) = \begin{cases} \frac{\mu_i(\mu_i + \gamma y_i)^{y_i-1} \exp(-(\mu_i + \gamma y_i))}{y_i!} & y_i = 0, 1, 2, \dots \\ 0, & \text{para } y_i > m \text{ quando } \gamma < 0 \end{cases} \quad (2.12)$$

e, zero caso contrário, em que $\mu_i > 0$, $\max(-1, -\mu_i/4) \leq \gamma \leq 1$ e m é o maior inteiro positivo para o qual $\mu_i + m\gamma > 0$ quando γ é negativo. A média e variância são dadas por: $E(Y_i) = \mu_i(1 - \gamma)^{-1}$ e $\text{Var}(Y_i) = \mu_i(1 - \gamma)^{-3}$ (Consul e Famoye, 1992). A PG se reduz à distribuição de Poisson quando $\gamma = 0$. Quando $\gamma > 0$, tem-se $\text{Var}(Y_i) > E(Y_i)$ e para $\gamma < 0$, a $\text{Var}(Y_i) < E(Y_i)$. Quando $\gamma < 0$, o espaço paramétrico é restrito, porém a distribuição de probabilidade é válida apenas para $y_i \leq m$, em que m pode depender do parâmetro γ , e segundo Consul e Famoye (1992), os estimadores de máxima verossimilhança para μ e γ são únicos. Quando $\gamma \geq 0$, o suporte da distribuição é independente dos parâmetros desconhecidos e o estimador de máxima verossimilhança é único (Consul e Famoye, 1992).

Famoye (1993) apresenta a distribuição Poisson generalizada restrita (PGR), a qual pode ser aplicada para dados de contagem seja qual for o grau de dispersão. A definição da restrição da proporcionalidade direta é dada como: $\gamma = \phi \mu_i$, e ao substituir na equação 2.12, obtém-se a PGR. Assim, o parâmetro γ torna-se restritivo e fornece a função de probabilidade dada por:

$$P(Y_i = y_i | \mu_i, \phi) = \begin{cases} \frac{\mu_i^{y_i} (1 + \phi y_i)^{y_i-1} \exp(-\mu_i(1 + \phi y_i))}{y_i!} & y_i = 0, 1, 2, \dots \\ 0, & \text{para } y_i > m \text{ quando } \phi < 0 \end{cases} \quad (2.13)$$

e, zero caso contrário, em que $\max(-\mu_i^{-1}, -1/4) \leq \phi \leq \mu_i^{-1}$ e m como definido em 2.12. Quando $\phi = 0$ a PGR se resume ao modelo Poisson. Quando $\phi \geq 0$ o suporte da PGR não depende do parâmetro desconhecido. E quando $\phi < 0$ a PGR é definida para $y_i = 0, 1, \dots, \text{inteiro}(-1/\phi)$ e para ambos os casos o modelo PGR possui estimativas de máxima verossimilhança únicas para os parâmetros (Famoye, 1993). Um inconveniente do modelo PGR é que quando o parâmetro de dispersão é negativo, o suporte de y é redefinido, de modo que a soma das probabilidades é aproximadamente 1, com erro ínfimo de aproximação menor que 0,5%.

2.2.3.1 Regressão Poisson generalizada restrita e estimação dos parâmetros

Seja $\mathbf{x}_i^T$ o vetor de covariáveis (como na seção 2.2.2.1) e Y_i uma variável aleatória com distribuição PGR. A média μ_i para função de probabilidade 2.13 é dada por: $E(Y_i) = \lambda_i = \mu_i(1 - \phi \mu_i)^{-1}$,

portanto, $\mu_i = \lambda_i(1 + \phi\lambda_i)^{-1}$. Ao substituí-la na equação 2.13 obtém-se a parametrização de média para a distribuição de probabilidade PGR, dada por:

$$P(Y_i = y_i | \lambda_i, \phi) = \left(\frac{\lambda_i}{1 + \phi\lambda_i} \right)^{y_i} \frac{(1 + \phi y_i)^{y_i - 1}}{y_i!} \exp \left(- \frac{\lambda_i(1 + \phi y_i)}{1 + \phi\lambda_i} \right) \quad y_i = 1, 2, \dots \quad (2.14)$$

e, por sua vez, a média e variância são representadas por: $E(Y_i) = \lambda_i$ e $\text{Var}(Y_i) = \lambda_i(1 + \phi\lambda_i)^2$, ϕ é o parâmetro de dispersão (Famoye, 1993). A regressão PGR é dada por:

$$\eta_i = \log(\lambda_i) = \sum_{j=1}^p x_{ij}\beta_j = \mathbf{x}_i^T \boldsymbol{\beta}.$$

O logaritmo natural da função de verossimilhança é dado por:

$$l(\boldsymbol{\beta}_j, \phi) = \sum_{i=1}^n [y_i x_{ij} \beta_j - y_i \log(1 + \phi \exp(x_{ij} \beta_j)) + (y_i - 1) \log(1 + \phi y_i) - \log(y_i!) + \frac{\exp(x_{ij} \beta_j)(1 + \phi y_i)}{1 + \phi \exp(x_{ij} \beta_j)}]$$

e as funções escore são obtidas derivando o logaritmo natural da função de verossimilhança em função de cada um dos parâmetros, portanto:

$$\frac{\partial l(\boldsymbol{\beta}_j, \phi)}{\partial \beta_j} = \sum_{i=1}^n \left[\frac{x_{ij}(y_i - \exp(x_{ij} \beta_j))}{(1 + \phi \exp(x_{ij} \beta_j))^2} \right] \quad (2.15)$$

e

$$\frac{\partial l(\boldsymbol{\beta}_j, \phi)}{\partial \phi} = \sum_{i=1}^n \left[- \frac{y_i \exp(x_{ij} \beta_j)}{1 + \phi \exp(x_{ij} \beta_j)} + \frac{y_i(y_i - 1)}{1 + \phi y_i} - \frac{\exp(x_{ij} \beta_j)(y_i - \exp(x_{ij} \beta_j))}{(1 + \phi \exp(x_{ij} \beta_j))^2} \right]. \quad (2.16)$$

Para obter as estimativas de máxima verossimilhança igualam-se as expressões 2.15 e 2.16 a zero e encontram-se as raízes do sistema de equações em relação a β_j e ϕ . Como este sistema não tem solução analítica faz-se necessário o uso de procedimentos numéricos. As estimativas iniciais para os parâmetros são as estimativas do modelo de regressão Poisson. Quando $\phi \geq 0$, o modelo PGR é dado em 2.14, entretanto, quando $\phi < 0$, y só pode assumir valores $0, 1, \dots$, inteiro $(-\phi^{-1})$.

Para fazer inferências sobre os parâmetros, faz-se necessária a matriz de informação observada (hessiana) no ponto encontrado, como descrito nas seções 2.2.2.1. Assim, as derivadas parciais de segunda ordem são:

$$\begin{aligned} \frac{\partial^2 l(\boldsymbol{\beta}_j, \phi)}{\partial \beta_j \partial \beta_s} = \sum_{i=1}^n & \left[\frac{y_i - \exp(x_{ij} \beta_j)}{\exp(x_{ij} \beta_j)(1 + \phi \exp(x_{ij} \beta_j))^2} (x_{ij} \exp(x_{ij} \beta_j) x_{is}) + \right. \\ & \left. + \frac{2\phi \exp(x_{ij} \beta_j)^2 - y_i(+3\phi \exp(x_{ij} \beta_j))}{\exp(x_{ij} \beta_j)^2 (1 + \phi \exp(x_{ij} \beta_j))^3} (\exp(x_{ij} \beta_j) x_{ij} \exp(x_{ij} \beta_j) x_{is}) \right] \\ j, s = 1, 2, \dots, p \end{aligned}$$

$$\frac{\partial^2 l(\boldsymbol{\beta}_j, \phi)}{\partial \phi^2} = \sum_{i=1}^n \left[\frac{y_i \exp(x_{ij} \beta_j)^2}{(1 + \phi \exp(x_{ij} \beta_j))^2} - \frac{y_i^2(y_i - 1)}{(1 + \phi y_i)^2} + \frac{2\exp(x_{ij} \beta_j)^2(y_i - \exp(x_{ij} \beta_j))}{(1 + \phi \exp(x_{ij} \beta_j))^3} \right]$$

e

$$\frac{\partial^2 l(\boldsymbol{\beta}_j, \phi)}{\partial \phi \partial \beta_j} = \sum_{i=1}^n \left[\frac{-2(y_i - \exp(x_{ij} \beta_j))}{(1 + \phi \exp(x_{ij} \beta_j))^2} x_{ij} \exp(x_{ij} \beta_j) \right].$$

E, dessa forma, a matriz de informação observada é definida por:

$$I_O(\beta_j, \phi) = \begin{bmatrix} -\frac{\partial^2 l(\beta_j, \phi)}{\partial \beta_1^2} & -\frac{\partial^2 l(\beta_j, \phi)}{\partial \beta_1 \partial \beta_2} & \dots & -\frac{\partial^2 l(\beta_j, \phi)}{\partial \beta_1 \partial \phi} \\ -\frac{\partial^2 l(\beta_j, \phi)}{\partial \beta_2 \partial \beta_1} & -\frac{\partial^2 l(\beta_j, \phi)}{\partial \beta_2^2} & \dots & -\frac{\partial^2 l(\beta_j, \phi)}{\partial \beta_2 \partial \phi} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial^2 l(\beta_j, \phi)}{\partial \beta_1 \partial \phi} & -\frac{\partial^2 l(\beta_j, \phi)}{\partial \beta_2 \partial \phi} & \dots & -\frac{\partial^2 l(\beta_j, \phi)}{\partial \phi^2} \end{bmatrix}.$$

2.2.4 Distribuição *double* Poisson

A distribuição *double* Poisson (DP) foi originalmente desenvolvida por Efron (1986) em um contexto da família exponencial dupla, e teve como propósito introduzir um segundo parâmetro na distribuição Poisson para permitir que a variância seja controlada separadamente da média. A DP pode ser usada para modelar dados de contagem sub e superdispersos. A sua distribuição é obtida como uma combinação exponencial de duas distribuições de Poisson $P[(\lambda_i)]$ e $P[(y_i)]$, dada por:

$$f(y_i | \lambda_i, \gamma) = k(\lambda_i, \gamma) \gamma^{\frac{1}{2}} [P(\lambda_i)]^\gamma [P(y_i)]^{1-\gamma},$$

em que γ é o parâmetro de dispersão. Se $\gamma = 1$, tem-se a distribuição Poisson. Se $0 < \gamma < 1$ tem-se $\text{Var}(Y_i) > E(Y_i)$ e se $\gamma > 1$ tem-se $\text{Var}(Y_i) < E(Y_i)$, $K(\lambda_i, \gamma)$ é a constante de normalização e o seu valor exato depende dos parâmetros λ_i e γ . Expandindo a densidade DP tem-se:

$$P(Y_i = y_i | \lambda_i, \gamma) = k(\lambda_i, \gamma) \gamma^{1/2} e^{-\gamma \lambda_i} \left(\frac{e^{-y_i} \lambda_i^{y_i}}{y_i!} \right) \left(\frac{e \lambda_i}{y_i} \right)^{\gamma y_i}, \quad y = 0, 1, 2, \dots, \quad (2.17)$$

em que

$$\frac{1}{k(\lambda_i, \gamma)} \simeq 1 + \frac{1-\gamma}{12\gamma\lambda_i} \left(1 + \frac{1}{\gamma\lambda_i} \right).$$

Efron (1986) mostrou que a constante $k(\lambda_i, \gamma)$ é próxima de 1. Uma desvantagem da distribuição DP é que os resultados para a constante de normalização, média e variância não são exatos. Efron (1986) mostra que:

$$E(Y_i) \approx \lambda_i \quad \text{e} \quad \text{Var}(Y_i) \approx \frac{\lambda_i}{\gamma}.$$

2.2.4.1 Regressão *double* Poisson e estimação dos parâmetros

Assumindo-se Y_i com distribuição de DP e o vetor de covariáveis $\mathbf{x}_i^T$ para cada $i = 1, 2, \dots, n$, o preditor da regressão DP é dado por:

$$\eta_i = \log(\lambda_i) = \sum_{j=1}^p x_{ij} \beta_j = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Substituindo $\lambda_i = \exp(x_{ij} \beta_j)$ na equação 2.17, obtém-se:

$$P(Y_i = y_i | x_i) = \left(1 + \frac{1-\gamma}{12\gamma \exp(x_{ij} \beta_j)} \left(1 + \frac{1}{\gamma \exp(x_{ij} \beta_j)} \right) \right)^{-1} \left(\gamma^{\frac{1}{2}} \exp(-\gamma \exp(x_{ij} \beta_j)) \right) \left(\frac{\exp(-y_i) y_i^{y_i}}{y_i!} \right) \left(\frac{\exp(1) \exp(x_{ij} \beta_j)}{y_i} \right)^{\gamma y_i}.$$

A média e a variância são dadas por $E(Y_i|x_i) \approx \lambda_i$ e $\text{Var}(Y_i|x_i) \approx \lambda_i/\gamma$, respectivamente. O logaritmo natural da função de verossimilhança é dado por:

$$l(\beta_j, \gamma) = \sum_{i=1}^n \left[-\log \left(1 + \frac{1-\gamma}{12\gamma \exp(x_{ij}\beta_j)} \left(1 + \frac{1}{\gamma \exp(x_{ij}\beta_j)} \right) \right) - y_i + \right. \\ \left. + y_i \log(y_i) - \log(y_i!) + \gamma y_i \log \left(\frac{\exp(1) \exp(x_{ij}\beta_j)}{y_i} \right) + \frac{1}{2} \log(\gamma) - \gamma \exp(x_{ij}\beta_j) \right],$$

cujas maximização ocorre de modo análogo ao processo discutido na seção 2.2.2.1.

2.2.5 Distribuição contagem gama (“gamma count”)

Existe uma relação entre a distribuição exponencial e Poisson que é bem discutida na literatura. Considerando Z_i uma variável aleatória que representa o número de ocorrências de um evento, em um determinado intervalo de tempo, isto é $Z_i \sim \text{Poisson}(\lambda_i)$. Se o tamanho do intervalo de interesse variar, pode-se considerar que o número de ocorrências dentro desse intervalo ainda é Poisson. No entanto, com parâmetro ajustado, agora tem-se a variável $Z_{it} \sim \text{Poisson}(\lambda_i t)$. Considerando que se tem uma ocorrência e Y_i é a variável aleatória que indica o tempo até tal fenômeno ocorrer novamente, tem-se para $y_i > 0$ que,

$$F_Y(y_i) = P(Y_i \leq y_i) = 1 - P(Y_i > y_i) = 1 - P(Z_{iy_i} = 0) = 1 - e^{-\lambda_i y_i},$$

portanto, tem-se que Y_i tem distribuição exponencial com parâmetro λ_i (Winkelmann, 2008). Com a proposta de flexibilizar a conexão entre os modelos de contagem, Winkelmann (1995) explorou essa relação com uma distribuição não negativa menos restritiva, pois a distribuição exponencial tem a propriedade da falta de memória e sua função de risco de sobrevivência é constante, ou seja, $Y_i \sim \exp(\lambda_i)$ então $P(Y_i \geq t+s | Y_i \geq s) = P(Y_i \geq t)$ para todo $t, s \geq 0$.

Algumas distribuições mais flexíveis são: lognormal, Weibull e gama, que têm aplicações em análise de sobrevivência. Neste contexto, Winkelmann (1995) propôs o uso da distribuição gama, para flexibilizar a suposição de equidispersão da distribuição Poisson ao custo de um parâmetro adicional. Essa escolha é plausível uma vez que a soma de distribuições gama independentes resulta em uma distribuição gama. Seja $\tau_k > 0$, $k \in \mathbb{N}$ denotando uma sequência de tempos de espera entre os eventos de ordens $k-1$ e k , em que k representa o k -ésimo evento, então o tempo de chegada do n -ésimo evento é dado por:

$$\vartheta_n = \sum_{k=1}^n \tau_k, \quad n = 1, 2, \dots$$

Seja N_T o número total de eventos em um intervalo de 0 a T . Se T é fixo, N_T é uma variável de contagem. Assim, resulta nas seguintes definições de N_T e ϑ_n , tal que

$$N_T < n \quad \leftrightarrow \quad \vartheta_n \geq T,$$

ou seja, o número total de eventos em um intervalo $(0, T)$ é menor do que o tempo de chegada do evento, se e somente se, o tempo de chegada do evento for maior ou igual ao comprimento do intervalo $(0, T)$. Portanto, pode-se estabelecer que

$$P(N_T < n) = P(\vartheta_n \geq T) = 1 - F_n(T),$$

em que $F_n(T)$ é a distribuição acumulada de ϑ_n . Da mesma forma, a

$$P(N_T = n) = F_n(T) - F_{n+1}(T). \quad (2.18)$$

A equação 2.18 fornece a relação fundamental entre a distribuição dos tempos de chegada e a distribuição de contagem. Assim, a distribuição de probabilidade de N_T pode ser obtida explicitamente para todos os tempos de chegada da distribuição ϑ_n (Winkelmann, 1995, 2008). Assumindo que os tempos de espera ϑ_n são independentes e identicamente distribuídos (i.i.d), com distribuição gama $G(\alpha, \beta^*)$, e omitindo o índice k , sua função de densidade pode ser escrita como:

$$f(\tau|\alpha, \beta^*) = \frac{(\beta^*)^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \exp(-\beta^* \tau), \quad \alpha, \beta^* \in \mathbb{R}^+,$$

com $\tau > 0$, média $E(\tau) = \alpha/\beta^*$ e variância $\text{Var}(\tau) = \alpha/(\beta^*)^2$. Como ϑ_n é a soma de n variáveis aleatórias i.i.d tais que, pode-se escrever diretamente a função de densidade de ϑ_n como:

$$f_n(\vartheta|\alpha, \beta^*) = \frac{(\beta^*)^{n\alpha}}{\Gamma(n\alpha)} \vartheta^{n\alpha-1} \exp(-\beta^* \vartheta).$$

E, para obter a distribuição de probabilidade contagem gama, tem-se que avaliar a função de distribuição acumulada:

$$F_n(T) = \int_0^T \frac{(\beta^*)^{n\alpha}}{\Gamma(n\alpha)} \vartheta^{n\alpha-1} \exp(-\beta^* \vartheta) d\vartheta = \frac{1}{\Gamma(n\alpha)} \int_0^{(\beta^*)T} u^{n\alpha-1} \exp(-u) du, \quad (2.19)$$

em que na segunda parte da igualdade aplica-se a mudança de variável para $u = \beta^* \vartheta$, representando a integral da gama incompleta que será denotada como $G(n\alpha, \beta^* T)$. A probabilidade do número de ocorrências é:

$$P(N = n) = G(n\alpha, \beta^* T) - G((n+1)\alpha, \beta^* T), \quad (2.20)$$

isto é, a probabilidade de ocorrência de um evento é dada pela diferença da distribuição acumulada da função gama em dois pontos. Para $\alpha = 1$, $f(\tau)$ reduz-se à distribuição exponencial e a equação 2.20 refere-se à distribuição de Poisson (Winkelmann, 1995, 2008). Quando não fixado o parâmetro α e sim estimado tem-se uma distribuição mais flexível para modelar dados de contagem não equidispersos. Quando $\alpha > 1$ ou $0 < \alpha < 1$ tem-se a presença de sub ou superdispersão, respectivamente. Para o caso de α não inteiro, $G(n\alpha, \beta^* T)$ não tem forma fechada, fazendo assim necessário o uso de avaliações numéricas da integral, podendo ser utilizadas expressões assintóticas (Winkelmann, 1995). Conforme Winkelmann (1995, 2008) o valor esperado e a variância não estão disponíveis em forma fechada, e devem ser computados como:

$$E[N(T)] = \sum_{n=1}^{\infty} n(P(N = n)) \quad \text{e} \quad \text{Var}[N(T)] = \sum_{n=1}^{\infty} [n^2 P(N = n)] - E[N(T)]^2.$$

E para T crescente, tem-se que

$$N(T) \stackrel{a}{\sim} N\left(\frac{\beta^* T}{\alpha}, \frac{\beta^* T}{\alpha^2}\right).$$

2.2.5.1 Regressão contagem gama e estimação dos parâmetros

Segundo Winkelmann (1995), para obter o modelo de regressão contagem gama, a suposição de homogeneidade da população é flexibilizada por meio da formulação de um modelo condicional em que os parâmetros dependem de um vetor de covariáveis individuais $\mathbf{x}_i^T$. Supondo que o intervalo do tempo entre eventos são idênticos para todas as observações, T pode ser definida como unidade sem perda de generalidade, ou seja, $T = 1$. Assumindo que:

$$\frac{\beta^*}{\alpha} = \exp(\mathbf{x}_i^T \boldsymbol{\theta}) \Rightarrow \beta^* = \alpha \exp(\mathbf{x}_i^T \boldsymbol{\theta}),$$

em que $\theta = (\beta_j)$ é o vetor de parâmetros a ser estimado. Esta parametrização produz um modelo de regressão definido por:

$$E(\tau_i|\mathbf{x}_i) = \exp(-\mathbf{x}_i^T \theta) = \frac{\alpha}{\beta^*}.$$

É importante destacar que esse modelo contagem gama é para a média do tempo entre eventos τ_i e não diretamente para as contagem dos eventos Y_i , porque a menos que $\alpha = 1$ não é correto que $E(Y_i|\mathbf{x}_i) = [E(\tau_i|\mathbf{x}_i)]^{-1} \Rightarrow E(Y_i|\mathbf{x}_i) = \beta$ (Winkelmann, 1995). A estimativa do vetor de parâmetros ($\hat{\theta}$) devem ser interpretadas da seguinte forma: $-\theta$ mensura a variação percentual no tempo de espera causado por um aumento de uma unidade em $\mathbf{x}_i$.

Dado uma amostra de observações independentes, as estimativas de $\hat{\alpha}$ e $\hat{\theta}$ podem ser obtidas maximizando o logaritmo natural da função de verossimilhança, dada por:

$$l(y_i|x_i, \alpha, \theta) = \sum_{i=1}^n \ln [G(\alpha y_i, \alpha \exp(\mathbf{x}_i^T \theta)) - G(\alpha y_i + \alpha, \alpha \exp(\mathbf{x}_i^T \theta))]$$

em que y_i são as observações da amostra, $\mathbf{x}_i$ é o vetor de covariáveis, θ é o vetor de parâmetros do modelo de regressão, α é o parâmetro de dispersão e $G(\cdot)$ é a função de densidade acumulada da distribuição contagem gama, dada pela equação 2.19. Como a função é não linear nos parâmetros, a estimação dos parâmetros requer métodos numéricos.

2.3 Estudo de caso - Efeito do aumento da biomassa de uma macrófita exótica invasora sobre a estrutura da comunidade de macrófitas aquáticas

Os dados utilizados neste estudo são provenientes de um trabalho de campo, conduzido por Amorim et al. (2015) realizado na bacia hidrográfica do rio Itanhaém, situado no município de Itanhaém, litoral sul do Estado de São Paulo, Brasil. Os objetos de estudo foram a exótica *Urochloa arrecta*, conhecida como *African signalgrass*, um capim aquático perene com ramos longos flutuantes que podem formar touceiras, geralmente com denso sistema radicular (Figura 2.1) e a assembleia de macrófitas aquáticas nativas presentes nos ambientes da bacia.

Figura 2.1: Banco da espécie exótica *Urochloa arrecta*, com ênfase para seu denso sistema radicular.

A amostragem em campo foi realizada no verão de 2012, seguindo a metodologia adotada por Michelin et al. (2010). Foram selecionados bancos de macrófitas localizados nas zonas litorâneas da bacia e o critério para a escolha dos bancos foi a predominância ou não da espécie exótica *U. arrecta*, medida em biomassa, totalizando 90 bancos. Definiu-se como predominância a ocorrência de *U. arrecta* cobrindo mais de 50% da amostra, por inspeção visual, e confirmado posteriormente em laboratório. As amostras foram obtidas por meio de um quadrado de 1m² (1m x 1m) e em cada quadrado foram identificadas as espécies presentes (Figura 2.2a) e seus respectivos grupos funcionais. Os grupos funcionais representam o conjunto de espécies que possuem forma de vida similares, ou seja, possuem hábitos e papéis semelhantes no ecossistema. Conforme Chambers et al. (2008), as macrófitas podem pertencer aos seguintes grupos funcionais: emergentes (plantas enraizadas com as partes vegetativas emergindo acima da superfície da água), flutuantes livres (plantas que flutuam sobre a superfície da água), flutuantes fixas (plantas enraizadas com folhas e flores flutuantes na superfície da água), submersas fixas (plantas enraizadas e com as partes vegetativas predominantemente submersas), submersas livres (plantas com as partes vegetativas submersas, porém sem estarem enraizadas no substrato) e epífitas (plantas que se desenvolvem sobre outros indivíduos). A biomassa das macrófitas em cada amostra foi quantificada a partir de um quadrado de 0,25m² posicionado no centro de cada quadrado de 1m² (Figura 2.2b). O material vegetal foi levado ao laboratório e seco em estufa a 60 °C, até peso constante. As 90 amostras foram selecionadas de modo a criar um gradiente de biomassa que variasse entre bancos com ausência e dominância da espécie exótica. Assim, o gradiente de biomassa de *U. arrecta* encontrado variou de 0 até 433,88 gPS.m⁻² (grama de peso seco por metro quadrado). Em suma, o atributo da estrutura da comunidade de macrófitas aquáticas avaliado foi a riqueza de grupos funcionais encontrada em cada quadrado amostral. A identidade dos grupos foi desprezada para fins de análise, uma vez que o objetivo do trabalho foi verificar se o aumento da biomassa da espécie exótica era capaz de afetar a riqueza de grupos funcionais, independentemente de qual forma de vida pertencesse.

Figura 2.2: Amostragem das macrófitas aquáticas por meio de quadrado de 1m² (a); e ilustração do esquema utilizado para realizar a identificação das espécies e coleta da biomassa (b), de acordo com estudo desenvolvido por Amorim et al. (2015).

2.4 Métodos

Para estabelecer a notação considere a variável resposta, Y_i como sendo o número de grupos funcionais, em que $i = 1, 2, \dots, 90$. Seja também o vetor de covariáveis $\mathbf{x}_i = (x_{i1}, x_{i2})^T$ e em que $x_{i1} = 1$ para todo i e x_{i2} é quantidade de biomassa em gPS.m⁻² (grama por peso seco por metro quadrado). Inicialmente considerou-se o modelo em que a variável resposta tem distribuição de Poisson com função de ligação logarítmica. Para a situação descrita e considerando os objetivos do estudo, tem-se como estrutura para o preditor linear os modelos:

$$\text{preditor 1: } \eta_i = \lambda_i = \exp(\beta_1);$$

$$\text{preditor 2: } \eta_i = \lambda_i = \exp(\beta_1 + \beta_2 x_{i2}).$$

A estrutura dos preditores lineares permite que a hipótese relevante seja testada pelo teste da razão de verossimilhanças. Para o estudo em questão, o interesse está em verificar se há efeito da quantidade de biomassa em relação ao número de grupos funcionais, ou seja, verificar se $H_0 : \beta_2 = 0$ é verdadeira ou falsa. Assim, se a diferença entre as logaritmo natural da função de verossimilhança do modelo sob pesquisa (preditor 2) e do modelo nulo (preditor 1) for pequena, pode-se concluir que o modelo sob pesquisa é tão informativo quanto o modelo completo. Assim, tem-se:

$$\xi_{RV} = 2 \left(l(\hat{\theta}|y)_{f2} - l(\hat{\theta}|y)_{f1} \right)$$

em que $l(\hat{\theta}|y)_{f1}$ é o logaritmo natural da função de verossimilhança do modelo nulo (preditor 1) encaixado em um modelo maior (preditor 2) com logaritmo natural da função de verossimilhança $l(\hat{\theta}|y)_{f2}$. Sob a hipótese nula, assintoticamente tem-se que $\xi_{RV} \sim \chi^2_{f2-f1}$, no qual $f1$ e $f2$ são os graus de liberdade associados aos modelos. O preditor 1 contém somente o intercepto, ou seja, não leva em consideração o efeito da quantidade de biomassa, já no preditor 2 tem-se o efeito da quantidade biomassa.

Sob a hipótese de não equidispersão, faz-se necessário o uso de modelos que levem em conta a sub ou superdispersão. Sendo assim, as estruturas dos preditores linear 1 e 2 foram consideradas, porém por meio das distribuições de probabilidade: COM-Poisson, Poisson generalizada restrita, *double* Poisson, contagem gama, para a variável resposta. Para se constatar a presença de não equidispersão faz-se necessário um teste de hipóteses para o parâmetro adicional das distribuições. Como descrito nas seções 2.2.2 a 2.2.5, para valores específicos do parâmetro de dispersão (COM-Poisson $v = 1$, Poisson generalizado restrita $\phi = 0$, *double* Poisson $\gamma = 1$, contagem gama $\alpha = 1$) essas distribuições reduzem à Poisson. Por isso, um teste de hipótese formal para esses parâmetros é importante. O critério utilizado é o da razão de verossimilhança. Para o modelo PGR tem-se:

$$\begin{cases} H_0 : \phi = 0 \\ H_1 : \phi \neq 0 \end{cases}.$$

e como estatística,

$$\xi_{RV} = -2 \left(l(\hat{\beta}, \phi = 0) - l(\hat{\beta}, \hat{\phi}) \right),$$

em que $l(\hat{\beta}, \phi = 0)$ é o logaritmo natural da função de verossimilhança do modelo de regressão Poisson e $l(\hat{\beta}, \hat{\phi})$ é o logaritmo natural da função de verossimilhança do modelo Poisson generalizada restrita. Sob a hipótese nula, assintoticamente tem-se que $\xi_{RV} \sim \chi^2_1$. Portanto, é possível verificar se a inclusão do parâmetro de dispersão é significativa ou não para o modelo. Em geral, $\hat{\beta}$ não é igual $\hat{\beta}$ (Famoye,

1993). Para os modelos COM-Poisson, *double* Poisson e contagem gama, o teste é elaborado de forma análoga.

Para discriminar cada modelo produzido por seus diferentes preditores lineares, o critério da informação de Akaike - AIC (*Akaike Criterion Information*) (Akaike, 1974) é calculado. O AIC é uma medida usada para descrever a troca entre precisão e complexidade entre os preditores lineares. O preditor linear com menor AIC é indicado. Assim, define-se o AIC como:

$$\text{AIC} = -2l(\hat{\theta}|y) + 2p,$$

em que $l(y|\hat{\theta})$ é logaritmo natural da função verossimilhança e p é o número de parâmetros do modelo sob pesquisa.

Adicionalmente, para avaliar a qualidade do ajuste dos modelos aos dados foram usados os gráficos meio normal de probabilidades (*half-normal plots* (hnp)). O comportamento esperado dos resíduos para um modelo adequado é aproximadamente uma reta. Para facilitar a análise gráfica Atkinson (1985) propôs a construção de um “envelope” simulado, com o objetivo de diminuir a subjetividade desta análise. Dessa forma, para um modelo adequado espera-se que os resíduos se distribuam satisfatoriamente dentro do “envelope”. O gráfico hnp é obtido plotando-se os valores absolutos ordenados de uma medida de diagnóstico adequada (neste trabalho utilizou-se o resíduo de Pearson generalizado) contra os correspondentes valores esperados das estatísticas de ordem, em valor absoluto, da distribuição meio-normal dada por:

$$\Phi^{-1} \left[\frac{(i + n - \frac{1}{8})}{2n + \frac{1}{2}} \right],$$

em que Φ^{-1} é a função inversa da função distribuição acumulada da distribuição normal padrão, sendo $i = 1, \dots, n$, sendo n a dimensão da amostra. Passos para a construção de um *half-normal plot*:

- i. Ajustar o modelo e calcular o resíduo pertinente representado por r_i , em valor absoluto e colocá-los em ordem crescente;
- ii. Simular 99 amostras para a variável resposta, utilizando o resultado do modelo dado em i;
- iii. Retornar aos modelos ajustados e para cada amostra simulada calcular os novos $r_{j(i)}$, em valores absolutos, $j = 1, \dots, 99$, $i = 1, \dots, n$ e dispor esses valores em ordem crescente;
- iv. Para cada modelo ajustado calcular os percentis (e.g. 2,5%, 50% e 97,5%);
- v. Plotar os valores desses percentis r_i observados contra as estatísticas esperadas da distribuição meio-normal.

Para a construção de intervalos de previsão, utiliza-se a matriz de variâncias e covariâncias condicionada ao parâmetro de dispersão, dada por: $\Sigma_{\beta|\phi} = \Sigma_{\beta\beta} - \Sigma_{\beta\phi} \Sigma_{\phi\phi}^{-1} \Sigma_{\phi\beta}$ (Johnson e Wichern, 2014), em que $\Sigma_{\beta\beta}$ é a matriz de variâncias e covariâncias dos β 's, $\Sigma_{\phi\phi}$ é a variância do parâmetro de dispersão, $\Sigma_{\beta\phi}$ e $\Sigma_{\phi\beta}$ são matrizes de variâncias e covariâncias entre os β 's e o parâmetro de dispersão. Dessa forma, o intervalo de previsão é dado por:

$$\hat{\eta}_i \pm z_{\alpha/2} \sqrt{\mathbf{x}_i^T \Sigma_{\beta|\phi} \mathbf{x}_i}.$$

As análises foram desenvolvidas com o auxílio do *software* R (R Core Team, 2019). As estimativas dos parâmetros foram efetuadas utilizando o pacote *bbmle* (Bolker, 2011), que fornece ferramentas necessárias para estimar os parâmetros dos modelos utilizando o logaritmo natural da função de máxima verossimilhança. Para o gráfico meio normal de probabilidade foi elaborada uma função própria e para simular as amostras das distribuições de probabilidade DP e CG foi usado o pacote *rmutil* (Swihart e Lindsey, 2019). Para as distribuições PGR e CP foi utilizado o pacote *RMKdiscrete* (Kirkpatrick, 2014) e *COMPOissonReg* (Sellers et al., 2011), respectivamente.

2.5 Resultados

Primeiramente uma análise exploratória foi realizada a fim de conhecer o comportamento dos dados. O número de grupos funcionais encontrado foi de 1, 2, 3 e 4 e os seus respectivos percentuais de ocorrência foram de 3,33%, 54,44%, 26,67% e 15,56%. A média e a variância foram 2,54 e 0,63 respectivamente, sendo possível notar que a variância é aproximadamente 4 vezes menor do que a média, havendo indício de subdispersão. A média e a variância são medidas fortemente afetadas por pontos extremos, e em um problema de subdispersão, geralmente, as contagens são aproximadamente homogêneas, como ocorre neste conjunto de dados.

Na Tabela 2.1 encontram-se o logaritmo natural da função de verossimilhança, teste de razão de verossimilhanças para uma sequência de preditores lineares assumindo-se as distribuições Poisson, CP, PGR, DP e CG, AIC e teste de hipótese do parâmetro de dispersão. Para todos os modelos o efeito da biomassa é significativo, ou seja, a quantidade de biomassa contribuiu para explicar o comportamento do número de grupos funcionais. Como indica a análise descritiva, a variância é menor do que a média, havendo indício de subdispersão, o que pode ser confirmado pelas estimativas dos parâmetros de dispersão dos modelos (CP ($\hat{v} = 8,359$), PGR ($\hat{\phi} = -0,196$), DP ($\hat{\gamma} = 6.638$) e CG ($\hat{\alpha} = 7,961$)). Quando considera-se equidispersão (modelo Poisson) o logaritmo natural da função de verossimilhança é maior para ambos os preditores lineares. Desse modo, ao realizar o teste de hipóteses para verificar se a inclusão do parâmetro de dispersão é significativa para os modelos CP, PGR, DP e CG, verificou-se que a hipótese nula foi rejeitada, ou seja, confirmando a existência de subdispersão. Ao observar o p -valor do modelo Poisson que inclui a quantidade de biomassa no modelo, verifica-se que o teste mostrou-se mais conservador, uma vez que este modelo não leva em consideração a subdispersão. Já para os demais modelos foi possível verificar que os p -valores são substancialmente menores.

Tabela 2.1: Medidas de ajuste, comparação entre preditores lineares e p -valores associados ao teste de dispersão para os modelos: Poisson, COM-Poisson (CP), Poisson generalizada restrita (PGR), *double* Poisson (DP) e contagem gama (CG).

Poisson	np	ln	AIC	dif np	2(dif ln)	$P(> \chi^2)$		
preditor 1	1	-136,593	275,186					
preditor 2	2	-133,718	271,436	1	5,750	0,016		
CP	np	ln	AIC	dif np	2(dif ln)	$P(> \chi^2)$	$\hat{v}$	$P(> \chi^2)^a$
preditor 1	2	-97,747	199,494				5,755	<0,001
preditor 2	3	-80,962	167,924	1	33,571	<0,001	8,359	<0,001
PGR	np	ln	AIC	dif np	2(dif ln)	$P(> \chi^2)$	$\hat{\phi}$	$P(> \chi^2)^a$
preditor 1	2	-108,933	221,866				-0,181	<0,001
preditor 2	3	-97,786	201,572	1	22,294	<0,001	-0,196	<0,001
DP	np	ln	AIC	dif np	2(dif ln)	$P(> \chi^2)$	$\hat{\gamma}$	$P(> \chi^2)^a$
preditor 1	2	-106,557	217,114				4,169	<0,001
preditor 2	3	-93,239	192,478	1	26,636	<0,001	6,638	<0,001
CG	np	ln	AIC	dif np	2(dif l)	$P(> \chi^2)$	$\hat{\alpha}$	$P(> \chi^2)^a$
preditor 1	2	-104,936	213,872				5,547	<0,001
preditor 2	3	-90,947	187,894	1	27,978	<0,001	7,961	<0,001

Notas: número de parâmetros (np); logaritmo natural da função de verossimilhança (ln); diferença do número de parâmetros (dif np); diferença do logaritmo natural da função de verossimilhança (dif ln); p -valores associados ao teste de hipótese para o parâmetro de dispersão (^a).

Na Tabela 2.2 são apresentadas as estimativas e erros padrões para os modelos Poisson, CP, PGR, DP e CG. Os resultados apresentaram similaridade para as estimativas dos parâmetros do modelo Poisson, PGR, DP e CG, uma vez que a razão (divisão das estimativas dos parâmetros do modelo Poisson pelas estimativas dos modelos CP, PGR, DP e CG) é próxima de um, porém os erros padrões obtidos pelos modelos PGR, DP e CG são menores quando comparados ao modelo Poisson, com redução, em média, de 2,5 vezes. Isso pode ser explicado pela ausência de um parâmetro para controlar a subdispersão, acarretando em erros padrões imprecisos para o modelo Poisson, o que poderia causar um comprometimento das inferências do modelo. O modelo CP apresentou as maiores diferenças em relação às estimativas dos parâmetros e erros padrões, aproximadamente nove e três vezes maior, respectivamente quando comparada ao modelo Poisson. Entretanto, é importante notar que as estimativas dos modelos CG e CP não são diretamente comparáveis com os demais modelos, uma vez que estes não modelam diretamente uma estrutura de regressão, ou seja, não estão relacionados a uma função da $E(Y_i)$, como é usual na estrutura de um MLG. O modelo CP modela uma estrutura para o parâmetro λ , ou melhor, modela o impacto que as covariáveis causam para λ_i e a média é dada por uma combinação entre λ_i e v . Verificou-se que todos os modelos indicaram uma redução no número de grupos funcionais quando há um aumento na quantidade de biomassa. Em geral, os modelos de dois parâmetros apresentaram resultados semelhantes, em termos de estimativas pontuais, intervalares (Figura 2.4) e testes de hipóteses.

Tabela 2.2: Estimativa dos parâmetros e erros padrões para os modelos Poisson, COM-Poisson (CP), Poisson generalizada restrita (PGR), *double* Poisson (DP) e contagem gama (CG) e razão entre as estimativas e erro padrão dos modelos.

Poisson					
Parâmetro	Estimativa	Erro padrão	P(> z)		
β_1	1,058	0,082	<0,001		
β_2	-0,001	0,001	<0,001		
COM-Poisson				Razão	
Parâmetro	Estimativa	Erro padrão	P(> z)	Estimativa	Erro padrão
β_1	10,021	1,547	<0,001	0,105	0,053
β_2	-0,009	0,002	<0,001	0,134	0,285
v	8,359	1,259	<0,001	-	-
Poisson generalizada restrita				Razão	
Parâmetro	Estimativa	Erro padrão	P(> z)	Estimativa	Erro padrão
β_1	1,089	0,036	<0,001	0,971	2,260
β_2	-0,002	<0,001	<0,001	0,812	1,566
ϕ	-0,196	0,010	<0,001	-	-
<i>double</i> Poisson				Razão	
Parâmetro	Estimativa	Erro padrão	P(> z)	Estimativa	Erro padrão
β_1	1,067	0,032	<0,001	0,991	2,584
β_2	-0,001	<0,001	<0,001	0,971	2,574
γ	6,638	0,991	<0,001	-	-
contagem gama				Razão	
Parâmetro	Estimativa	Erro padrão	P(> z)	Estimativa	Erro padrão
β_1	1,204	0,030	<0,001	0,878	2,752
β_2	-0,001	<0,001	<0,001	1,090	2,827
α	7,961	1,478	<0,001	-	-

Ao observar o gráfico meio normal de probabilidade com envelope simulado para o modelo Poisson verifica-se que praticamente todos os pontos estão fora do envelope simulado, evidenciando que o modelo Poisson não se ajustou de forma satisfatória aos dados (Figura 2.3a). Os modelos CP (Figura 2.3b), DP (Figura 2.3d) e CG (Figura 2.3e) apresentaram ajustes satisfatórios, pois observa-se que os pontos se acomodaram dentro do envelope simulado. O modelo PGR (Figura 2.3c) não apresentou um ajuste adequado, uma vez que a maioria dos pontos estão fora do envelope simulado.

Figura 2.3: Gráfico meio normal de probabilidade com envelope simulado com nível de confiança igual a 95% para os modelos: Poisson (a); COM-Poisson (b); Poisson generalizada restrita (c); *double* Poisson (d) e contagem gama (e).

As curvas de previsão para os modelos Poisson, CP, PGR, DP e CG são apresentadas na Figura 2.4. É possível verificar que os modelos CP, DP e CG apresentaram semelhanças em suas curvas de previsão e bandas de confiança de 95%, enquanto que o modelo PGR apresenta curva de previsão e bandas de confiança distintas dos demais modelos. As bandas de confiança de 95% para o modelo Poisson são as maiores, uma vez que este modelo superestima os erros padrões.

Figura 2.4: Diagrama de dispersão dos valores observados e curvas dos valores previstos com intervalo de 95% confiança.

Na Figura 2.5 encontram-se as medidas de probabilidade estimadas para o número de grupos funcionais (riqueza de grupos funcionais), considerando-se uma biomassa 200 gPS.m^{-2} . Observa-se que os modelos CP, CG e DP estimam maior probabilidade de ocorrência em torno da média do número de grupos funcionais, indicando que estes modelos apresentaram-se mais realistas ao relacionar o número de grupos funcionais à quantidade de biomassa determinada da espécie alvo do estudo. Os modelos Poisson e PGR atribuíram menor probabilidade de ocorrência em torno da média e maior probabilidade de ocorrência para os demais números de grupos funcionais, com exceção quando o número de grupos funcionais é 3 para o modelo Poisson. Ressalta-se que no modelo PGR o suporte de y está redefinido para $y = 0, 1, \dots, 5$, de modo que a soma das probabilidades não é igual a 1, entretanto o erro de aproximação é 0,01%. É interessante notar que os modelos PGR, DP e CG apresentaram resultados semelhantes nas estimativas dos parâmetros e erro padrão (Tabela 2.2). No entanto, as medidas de probabilidade para o modelo PGR não se mostra plausível, ainda que o suporte de y esteja definido em ocorrências realistas,

uma vez que os dados não apresentaram contagens acima de 5 grupos funcionais, sendo essa uma questão intrigante para pesquisas futuras, pois, todas essas distribuições são adequadas para dados de contagem não equidispersos.

Figura 2.5: Probabilidades estimadas considerando-se uma biomassa 200 gPS.m^{-2} para os modelos Poisson, Poisson generalizada restrita, COM-Poisson, *double* Poisson e contagem gama em relação ao número de grupos funcionais das macrófitas aquáticas.

2.6 Discussão

Os resultados apresentados neste estudo sugerem que a biomassa da espécie exótica invasora afeta adversamente o número de grupos funcionais de macrófitas aquáticas, de modo que à medida que a biomassa de *U. arrecta* aumenta, o número de grupos funcionais diminui. Isso possui uma implicação biológica importante, pois os dados sugerem que o crescimento de espécies exóticas, expresso em biomassa, inibe o desenvolvimento de outras espécies nativas. Esses resultados são úteis no que concerne a conservação de ecossistemas aquáticos. Nesse sentido, é importante utilizar análises coerentes que possam avaliar satisfatoriamente o comportamento das variáveis estudadas.

Os modelos alternativos usados para acomodar a subdispersão no estudo ecológico, de modo geral, apresentaram um desempenho mais satisfatório quando comparados ao modelo Poisson. O modelo Poisson apresentou estimativas dos erros padrões maiores, quando comparado aos demais modelos. Assim, suas estimativas dos erros padrões são superestimadas e, portanto, suas estimativas podem não ser suficientemente confiáveis, podendo afetar a inferência do modelo, principalmente em teste de hipóteses

para verificar se há efeito da covariável (ou efeito de tratamento). Além disso, por meio do critério utilizado para verificar a qualidade do ajuste dos modelos (baseado nos resultados do gráfico meio normal de probabilidade), verificou-se que os modelos CP, DP e CG ajustaram-se melhor aos dados. Uma das vantagens dos modelos CP, PGR, DP e CG é que eles são paramétricos. Dessa forma, algumas questões de interesse científico podem ser discutidas, como medidas de probabilidade para distribuição das contagens e medidas de verossimilhança para comparar modelos e submodelos. Estas vantagens não são encontradas, por exemplo, no modelo quasi-Poisson proposto por Wedderburn (1974) e aplicado por Zeviani et al. (2014) em um contexto de subdispersão. Além disso, não pode ser utilizado na classe dos modelos de Poisson-Tweedie proposta por Bonat et al. (2018), porque quando há subdispersão o parâmetro de dispersão é negativo e a distribuição de probabilidade é desconhecida.

Considerando as probabilidades estimadas pelos modelos, assume-se que estas medidas permitem a predição do número de grupos funcionais de acordo com a quantidade de biomassa da espécie exótica. Estas medidas de probabilidade podem ser uma expressão do que possivelmente é encontrado na natureza. Em outras palavras, por meio destas medidas será possível determinar o número potencial de grupos funcionais baseado nas condições da biomassa de *U. arrecta* no habitat aquático em estudo. Neste sentido, estes resultados podem auxiliar especialistas em conservação de ambientes aquáticos continentais na tomada de decisão, no âmbito de manejo e mitigação dos efeitos de espécies invasoras. No entanto, é importante ressaltar que as generalizações devem ser utilizadas com cautela e o ideal seria a possibilidade de testar esses resultados diretamente na comunidade aquática e verificar a eficiência do método *in situ*.

2.7 Considerações finais

Neste trabalho considerou-se quatro modelos para dados de contagem que apresentam um parâmetro de dispersão, flexibilizando assim a modelagem de dados em relação a sub- e superdispersão. Como em ecologia é comum o registro de contagens, recomenda-se que estudos e pesquisas nessa área considerem a possibilidade de não equidispersão. Neste contexto, um diagnóstico correto e a seleção do modelo adequado são imprescindíveis para bons resultados de pesquisa. Portanto, diante do exposto neste trabalho espera-se motivar outros pesquisadores a usar esses modelos perante essa especificidade dos dados, ao invés de tratá-los como dados anômalos, demonstrando suas potencialidades e adequação dessa abordagem em uma variedade de aplicações científicas comuns.

Referências

- Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.
- Amorim, S. R., Umetsu, C. A., Toledo, D., e Camargo, A. F. M. (2015). Effects of a non-native species of poaceae on aquatic macrophyte community composition: a comparison with a native species. *Journal of Aquatic Plant Management*, 53:191–196.
- Atkinson, A. C. (1985). Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis. Technical report.

- Bolker, B. (2011). R development core team. 2014. bbmle: Tools for general maximum likelihood estimation. r package version 1.0. 17. *Computer program*.
- Bonat, W. H., Jørgensen, B., Kokonendji, C. C., Hinde, J., e Demétrio, C. G. B. (2018). Extended poisson–tweedie: properties and regression models for count data. *Statistical Modelling*, 18(1):24–49.
- Box, G. E. e Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26(2):211–252.
- Cameron, A. C. e Trivedi, P. (2013). *Regression analysis of count data*, volume 53. Cambridge University Press.
- Chambers, P., Lacoul, P., Murphy, K., e Thomaz, S. (2008). Global diversity of aquatic macrophytes in freshwater. *Hydrobiologia*, 595(1):9–26.
- Consul, P. C. e Famoye, F. (1992). Generalized poisson regression model. *Communications in Statistics-Theory and Methods*, 21(1):89–109.
- Conway, R. W. e Maxwell, W. L. (1962). Network dispatching by the shortest-operation discipline. *Operations Research*, 10(1):51–73.
- Demétrio, C. G. B., Hinde, J., e Moral, R. A. (2014). Models for overdispersed data in entomology. In *Ecological modelling applied to entomology*, pages 219–259. Springer.
- Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81(395):709–721.
- Elton, C. (1958). Ecology of invasions by animals and plants.
- Famoye, F. (1993). Restricted generalized poisson regression model. *Communications in Statistics-Theory and Methods*, 22(5):1335–1354.
- Famoye, F., Wulu, J. T., e Singh, K. P. (2004). On the generalized poisson regression model with an application to accident data. *Journal of Data Science*, 2(3):287–295.
- Guikema, S. D. e Goffelt, J. P. (2008). A flexible count data regression model for risk analysis. *Risk analysis*, 28(1):213–223.
- Jamil, S. A. M., Abdullah, M. A. A., Kek, S. L., Nor, M. E., Mohamed, M., e Ismail, N. (2017). Detecting overdispersion in count data: A zero-inflated poisson regression analysis. In *Journal of Physics: Conference Series*, volume 890, page 012170. IOP Publishing.
- Jay, M. V. H. e Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772.
- Johnson, R. A. e Wichern, D. W. (2014). *Applied multivariate statistical analysis*, volume 4. Prentice-Hall New Jersey.

- Jowaheer, V. e Khan, N. M. (2009). Estimating regression effects in com-poisson generalized linear model. *International Journal of Computational and Mathematical Sciences*, 1(2):59–63.
- Kirkpatrick, R. M. (2014). *RMKdiscrete: Sundry Discrete Probability Distributions*. R package version 0.1.
- Kokonendji, C. C. (2014). Over-and underdispersion models. *Methods and Applications of Statistics in Clinical Trials*, 2:506–526.
- McCullagh, P. e Nelder, J. A. (1989). *Generalized linear models (Monographs on statistics and applied probability 37)*.
- Michelan, T. S., Thomaz, S., Mormul, R. P., e Carvalho, P. (2010). Effects of an exotic invasive macrophyte (tropical signalgrass) on native plant community composition, species richness and functional diversity. *Freshwater Biology*, 55(6):1315–1326.
- Minka, T. P., Shmeuli, G., Kadane, J. B., Borle, S., e Boatwright, P. (2003). Computing with the com-poisson distribution. *Department of Statistics, Carnegie Mellon University*.
- Moral, R. A., Hinde, J., Demétrio, C. G. B., Reigada, C., e Godoy, W. A. C. (2018). Models for jointly estimating abundances of two unmarked site-associated species subject to imperfect detection. *Journal of Agricultural, Biological and Environmental Statistics*, 23(1):20–38.
- Nelder, J. e Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.
- O'Hara, R. B. e Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, 1(2):118–122.
- Pott, V. J., Pott, A., Lima, L. C. P., Moreira, S. N., e Oliveira, A. K. M. (2011). Aquatic macrophyte diversity of the pantanal wetland and upper basin. *Brazilian Journal of Biology*, 71(1):255–263.
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Reinert, B., Bornschein, M., e Firkowski, C. (2007). Distribution, population size, habitat and conservation of the marsh antwren, stymphalornis acutirostris bornschein, reinert e teixeira, 1995 (thamnophilidae). *Revista Brasileira de Ornitologia*, 15(4):493–519.
- Richards, S. A. (2008). Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, 45(1):218–227.
- Ridout, M. S. e Besbeas, P. (2004). An empirical model for underdispersed count data. *Statistical Modelling*, 4(1):77–89.
- Sellers, K., Lotze, T., e Raim, A. (2011). Compoissonreg: Conway-maxwell poisson (com-poisson) regression. *R package version 0.3*, 2.
- Sellers, K. F. e Morris, D. S. (2017). Underdispersion models: Models that are under the radar. *Communications in Statistics-Theory and Methods*, 46(24):12075–12086.

- Sellers, K. F. e Shmueli, G. (2008). A flexible regression model for count data. *Robert H. Smith School Research Paper No. RHS*, pages 06–060.
- Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., e 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.
- Swihart, B. e Lindsey, J. (2019). *rmutil: Utilities for Nonlinear Regression and Repeated Measurements Models*. R package version 1.1.2.
- Tilman, D. (2001). Functional diversity. *Encyclopedia of biodiversity*, 3(1):109–120.
- Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61(3):439–447.
- Winkelmann, R. (1995). Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics*, 13(4):467–474.
- Winkelmann, R. (2008). *Econometric analysis of count data*. Springer.
- Wu, G., Holan, S. H., e Wikle, C. K. (2013). Hierarchical bayesian spatio-temporal conway–maxwell poisson models with dynamic dispersion. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(3):335–356.
- Zeviani, W. M., Ribeiro Jr, P. J., Bonat, W. H., Shimakura, S. E., e Muniz, J. A. (2014). The gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics*, 41(12):2616–2626.

Apêndice

```
#####
# I used packages
#####
# Installing the package
list.of.packages <- c("corrplot", "rmutil", "scatterplot3d", "bbmle")
new.packages <- list.of.packages[!(list.of.packages %in%
  installed.packages()[,"Package"])]
if(length(new.packages)) install.packages(new.packages)
#####
# entering the data
#####
dados <- read.delim("D:dados.txt")
#####
# Using a glm function for Poisson models
#####
# predictor 1
fit_poisson_1<-glm(Sfuncional~1,family = poisson,data=dados)
```

```

summary(fit_poisson_1)
# Vector of variable and the vector of covariates predictor 1.
V_1 <- with(dados,list(y = Sfuncional, X = model.matrix(fit_poisson_1)))
# predictor 2
fit_poisson_2<-glm(Sfuncional~Biom_Urochloa,family = poisson,data=dados)
summary(fit_poisson_2)
anova(fit_poisson_2,test="Chisq")
# Vector of variable and the vector of covariates predictor 2.
V_2 <- with(dados,list(y = Sfuncional, X = model.matrix(fit_poisson_2)))
#####
# Using the glm function for the quasi-Poisson models
#####
fit_quasi_poisson<-glm(Sfuncional~Biom_Urochloa,family = quasipoisson,data=dados)
summary(fit_quasi_poisson)
#When  $0 < \sigma^2 < 1$  indicates underdispersion,
# $\sigma^2 = 1$  equidispersion and  $\sigma^2 > 1$  overdispersion.
anova(fit_quasi_poisson,test="F")
#####
#The Log-likelihood function of the COM-Poisson regression model
#####
com_poisson=function(params, y, X){
# params: parameter vector;
# params[1]: dispersion parameter  $\nu$ ;
# params[-1]: location parameter  $\lambda$ ;
# y: variable response count;
# X: linear model vector;
v <- params[1]
lambda <- exp(X %*% params[-1])
n<- length(y)
ll<- y*log(lambda)-v*lfactorial(y)+((v-1)/(2*v))*log(lambda)+
((v-1)/(2))*log(2*pi)+0.5*log(v)-
v*(lambda^(1/v))
#Negative of log-likelihood.
-sum(ll)
}
# Uses the Poisson model estimates as initial values of predictor 1.
start_compoisson_1 <- c(v = 1, coef(fit_poisson_1))
parnames(com_poisson) <- names(start_compoisson_1)
# Uses the Poisson model estimates as initial values of predictor 2.
start_compoisson_2 <- c(v = 1, coef(fit_poisson_2))
parnames(com_poisson) <- names(start_compoisson_2)
# uses the Poisson estimates as initial values for the com-poisson model.

```

```

# With  $\nu = 1$  fixed in Poisson's equation use fixed = list (v = 1).
# predictor 1
fit_compoisson_1<- mle2(com_poisson,
start = start_compoisson_1,
data = V_1,
fixed = NULL,
method = "BFGS",
vecpar = TRUE)
summary(fit_compoisson_1);plot(profile(fit_compoisson_1))
logLik(fit_compoisson_1);AIC(fit_compoisson_1)
#p-value: hypothesis for the dispersion parameter predictor 1
diff_1_a<-(-2)*(logLik(fit_poisson_1)-logLik(fit_compoisson_1))
pchisq(q=diff_1_a[1],df=1,lower.tail = F)
# predictor 2
fit_compoisson_2<- mle2(com_poisson,
start = start_compoisson_2,
data = V_2,
fixed = NULL,
method = "BFGS",
vecpar = TRUE)
summary(fit_compoisson_2);plot(profile(fit_compoisson_2))
logLik(fit_compoisson_2);AIC(fit_compoisson_2)
diff_L<-(-2)*(logLik(fit_compoisson_2)-logLik(fit_compoisson_1))
#p-value: hypothesis for the covariate effect
pchisq(q=diff_L[1],df=1,lower.tail = F)
#p-value: hypothesis for the dispersion parameter predictor 2
diff_2_a<-(-2)*(logLik(fit_poisson_2)-logLik(fit_compoisson_2))
pchisq(q=diff_2_a[1],df=1,lower.tail = F)
#####
#The Log-likelihood function of the restricted generalized Poisson model
#####
generalized_restricted_poisson=function(params, y, X){
# params: parameter vector;
# params[1]: dispersion parameter  $\alpha$ ;
# params[-1]: location parameter  $\lambda$ ;
# y: variable response count;
# X: linear model vector;
alpha <- params[1]
lambda <- exp(X %*% params[-1])
z <- 1 + alpha * lambda
w <- 1 + alpha * y
ll <- y * (log(lambda) - log(z)) +

```

```

(y - 1) * log(w) -lambda * (w/z) -lfactorial(y)
# Negative of log-likelihood.
-sum(ll)
}restricted
# Uses the Poisson model estimates as initial values of predictor 2.
start_generalized_restricted_poisson<- c(alpha = 0, coef(fit_poisson_2))
parnames(generalized_restricted_poisson) <- names(start_generalized_restricted_poisson)
# uses the Poisson estimates as initial values for the restricted
  generalized Poisson model.
# With  $\alpha = 0$  fixed in Poisson's equation use fixed = list (alpha = 0).
fit_generalized_restricted_poisson<- mle2(generalized_restricted_poisson,
start = start_generalized_restricted_poisson,
data = V_2,
fixed =NULL,
method = "BFGS",
vecpar = TRUE)
summary(fit_generalized_restricted_poisson)
plot(profile(fit_generalized_restricted_poisson))
#####
#The Log-likelihood function of the Double Poisson model
#####
double_poisson=function(params, y, X){
# params: parameter vector;
# params[1]: dispersion parameter  $\theta$ ;
# params[-1]: location parameter  $\lambda$ ;
# y: variable response count;
# X: linear model vector;
theta <- params[1]
lambda <- exp(X %*% params[-1])
a<- 1+(1/12)*(1-theta)*(1+1/(lambda*theta))/(lambda*theta)
b<- (exp(1)*lambda)/y
ll<- (1/2)*log(theta)-lambda*theta-y*y*log(y)-log(factorial(y))+theta*y*log(b)
-log(a)
# Negativo da log-likelihood.
-sum(ll)
}
# Uses the Poisson model estimates as initial values of predictor 2.
start_double_poisson<- c(theta = 1, coef(fit_poisson_2))
parnames(double_poisson) <- names(start_double_poisson)
# uses the Poisson estimates as initial values for the restricted
  generalized Poisson model.
# With  $\theta = 0$  fixed in Poisson's equation use fixed = list (theta = 1).

```

```

fit_double_poisson<- mle2(double_poisson,
start = start_double_poisson,
data = V_2,
fixed =NULL,
method = "BFGS",
vecpar = TRUE)
summary(fit_double_poisson)
plot(profile(fit_double_poisson))
#####
#The Log-likelihood function of the Gamma-count model
#####
gamma_count <- function(params, y, X){
# params: parameter vector;
# params[1]: dispersion parameter  $\alpha$ ;
# params[-1]: location parameter  $\lambda$ ;
# y: variable response count;
# X: linear model vector;
alpha <- params[1]
eXb <- exp(X%*%params[-1])
alpha.y <- alpha*y
alpha.eXb <- alpha*eXb
a <- pgamma(1, shape=alpha.y, rate=alpha.eXb)
b <- pgamma(1, shape=alpha.y+alpha, rate=alpha.eXb)
# Negativo da log-likelihood.
ll <- -sum(log(a-b))
}
# Uses the Poisson model estimates as initial values of predictor 2.
start_gamma_count<- c(alpha = 1, coef(fit_poisson_2))
parnames(gamma_count) <- names(start_gamma_count)
# uses the Poisson estimates as initial values for the Gamma-count model.
# With  $\alpha = 1$  fixed in Poisson's equation use fixed = list (alpha = 1).
fit_gamma_count<- mle2(gamma_count,
start = start_gamma_count,
data = V_2,
fixed =NULL,
method = "BFGS",
vecpar = TRUE)
summary(fit_gamma_count)
plot(profile(fit_gamma_count))

```


3 MODELO DE REGRESSÃO *DOUBLE POISSON* ADAPTADO PARA CONSIDERAR SUBDISPERSÃO E SUPERDISPERSÃO CONJUNTAMENTE EM DADOS DE CONTAGEM LONGITUDINAL

Resumo

Este trabalho introduz uma extensão dos modelos de efeitos aleatórios fundamentada na distribuição de probabilidade *double Poisson*, como abordagem alternativa para a análise de dados de contagem longitudinais, que apresentam subdispersão e superdispersão concomitantemente. Para tal proposta utilizou-se uma variável *dummy* associada ao parâmetro de dispersão, definida com base na variabilidade intraindividual. Um estudo de simulação também foi desenvolvido para verificar a eficiência do modelo *double Poisson* misto sob esses dois fenômenos conjuntamente. Os resultados do estudo de simulação atestam o bom desempenho da aplicação da metodologia proposta. O método foi aplicado a um conjunto de dados reais provenientes de um estudo clínico, que objetivou verificar se o uso da droga Progabide reduz o número de crises epiléticas dos pacientes ao longo do tempo. Os resultados do modelo *double Poisson* misto considerando a variável *dummy* foram comparados aos modelos Poisson misto e *double Poisson* misto sem a variável *dummy*. A estimação dos parâmetros foi realizada via máxima verossimilhança e a inclusão da variável *dummy* no parâmetro de dispersão foi avaliada por meio do teste de razão de verossimilhança. Para a validação dos modelos, utilizou-se o gráfico *half-normal plot*. Com essa abordagem foi possível flexibilizar a análise de modo a acomodar a não equidispersão e obter inferências para os parâmetros de interesse.

Palavras-chave: Efeitos aleatórios; Máxima verossimilhança; Correlação; *Dummy*.

3.1 Introdução

O modelo de regressão Poisson é uma primeira alternativa para análises de dados de contagem, tanto para estudos transversais quanto para estudos longitudinais. No caso da abordagem longitudinal, deve-se considerar que medidas repetidas obtidas nos mesmo indivíduos naturalmente tendem a ser correlacionadas. O uso de técnicas estatísticas que não levam em consideração a correlação intraindividual pode acarretar em conclusões imprecisas (Caruana et al., 2015).

Os modelos clássicos de regressão para dados de contagem longitudinais pertencem à família dos modelos lineares generalizados mistos (MLGM) e têm sido aplicados em diversos estudos, uma vez que são capazes de acomodar a existência da correlação nas respostas longitudinais por meio dos efeitos aleatórios (Breslow e Clayton, 1993). Na presença de um modelo misto, assume-se que os coeficientes de regressão variam entre indivíduos devido a fatores não medidos e a correlação entre as respostas é resultante dessa heterogeneidade. Há outras classes de modelos para dados longitudinais, como os modelos marginais baseados na equação de estimação generalizadas (EEG) (Zeger e Liang, 1986) e os modelos condicionais estocásticos (Lara et al., 2019). No entanto, eles não estão no escopo desse trabalho.

Dados de contagem longitudinais são frequentemente analisados por meio do modelo Poisson misto. Porém ele apresenta uma relação de média-variância muito restritiva (equidispersão), o que limita a sua aplicação na presença da não equidispersão. Molenberghs et al. (2007) propuseram um MLGM para dados de contagem longitudinais usando efeitos aleatórios gama e normal. Os modelos binomial

negativo gama (efeitos aleatórios gama) e Poisson-normal (efeitos aleatórios normais) são casos especiais dessa família. Na presença de superdispersão, por exemplo, o modelo binomial negativo misto tem sido utilizado como modelo alternativo ao modelo de Poisson misto (Zhang et al., 2018; Sung e Lee, 2018). Não obstante, atualmente alguns pesquisadores têm empregado outros métodos alternativos para considerar a não equidispersão em dados longitudinais, como os modelos Poisson generalizado (Tosca e Faddy, 2003), Poisson generalizado INGARCH (Zhu, 2012), regressão Weibull (Luyts et al., 2018) e COM-Poisson (Khan et al., 2018). Os esforços em encontrar métodos satisfatórios para a análise de dados de contagem longitudinais tendem a estar mais concentrados no problema da superdispersão (Niu e Ryan Cho, 2018; Tapia et al., 2019), já que é uma condição comumente encontrada em diversas áreas do conhecimento. Em contrapartida, trabalhos relacionados à subdispersão ou trabalhos que abordam tanto a subdispersão quanto a superdispersão, são menos frequentes na literatura científica (Grunwald et al., 2011; Zhu, 2012).

A motivação deste trabalho provém de um conjunto de dados da literatura (Leppik et al., 1985), referente ao estudo clínico do uso terapêutico de uma droga para controlar crises epiléticas. Os dados provenientes deste estudo foram utilizados na literatura (Thall e Vail, 1990; Breslow e Clayton, 1993; Diggle et al., 2002; Booth et al., 2003; Rigby e Stasinopoulos, 2005; Tapia et al., 2019) no âmbito da análise de dados longitudinais. Contudo, os estudos supracitados habitualmente levavam em consideração apenas a superdispersão presente neste conjunto de dados, uma vez que a variabilidade analisada era oriunda da variação interindividual, ou seja, entre os indivíduos estudados. No entanto, ao analisar o número de crises epiléticas sob a perspectiva intraindividual, observa-se que os dados apresentam tanto subdispersão quanto superdispersão. Neste contexto, este trabalho apresenta como proposta caracterizar o problema da subdispersão em dados de contagem longitudinais, bem como apresentar uma contribuição para a análise conjunta da subdispersão e superdispersão.

Assim, propõe-se a utilização do modelo *double* Poisson (Efron, 1986) em um contexto da família exponencial dupla, que tem como propósito introduzir um segundo parâmetro na distribuição de Poisson para permitir que a variância seja controlada separadamente da média. Optou-se por estudar a distribuição *double* Poisson, devido a recursos computacionais disponíveis. Dessa forma, explora-se a flexibilidade dos modelos mistos, uma vez que são capazes de considerar a existência da correlação nas respostas longitudinais por meio dos efeitos aleatórios, que por sua vez compõem as estruturas de variância e covariância. Com isso, um modelo de regressão *double* Poisson misto é considerado, de modo a acomodar a subdispersão e superdispersão concomitantemente, por meio da inclusão de uma variável *dummy* ao parâmetro de dispersão.

3.2 Métodos

3.2.1 Distribuição de probabilidade *double* Poisson

A distribuição *double* Poisson (DP) Efron (1986) é obtida como uma combinação exponencial de duas distribuições de Poisson $P[\lambda]$ e $P[y]$ dada por:

$$f(y; \lambda, \gamma) = k(\lambda, \gamma) \gamma^{\frac{1}{2}} [P(\lambda)]^{\gamma} [P(y)]^{1-\gamma}.$$

Expandindo a densidade DP, tem-se:

$$P(Y = y|\lambda, \gamma) = k(\lambda, \gamma) \gamma^{1/2} e^{-\gamma\lambda} \left(\frac{e^{-\gamma} \gamma^y}{y!} \right) \left(\frac{e\lambda}{y} \right)^y, \quad \lambda > 0, \quad y = 0, 1, 2, \dots,$$

em que

$$\frac{1}{k(\lambda, \gamma)} \simeq 1 + \frac{1 - \gamma}{12\gamma\lambda} \left(1 + \frac{1}{\gamma\lambda} \right),$$

sendo γ o parâmetro de dispersão e $K(\lambda, \gamma)$ a constante de normalização, cujo valor exato depende dos parâmetros λ e γ . Efron (1986), de forma numérica, mostrou que a constante $k(\lambda, \gamma)$ é próxima de 1. Assim, na distribuição DP se $\gamma = 1$, tem-se a distribuição Poisson como caso particular e portanto $\text{Var}(Y) = E(Y)$, se $0 < \gamma < 1$ então $\text{Var}(Y) > E(Y)$ e se $\gamma > 1$ então $\text{Var}(Y) < E(Y)$. Dessa forma, o parâmetro adicional γ desempenha um papel importante quando se assume tal distribuição para a variável resposta de contagem uma vez que, se pode modelar a superdispersão e/ou subdispersão. É importante destacar que a não equidispersão é uma característica intrínseca dos dados. A distribuição de probabilidade, por sua vez, contém o parâmetro capaz de modelar tal propriedade dos dados. Efron (1986) mostra que:

$$E(Y) \simeq \lambda \quad \text{e} \quad \text{Var}(Y) \simeq \frac{\lambda}{\gamma}.$$

Com o objetivo de melhorar o desempenho computacional da distribuição DP, Stasinopoulos et al. (2019) apresentam uma nova reparametrização em que $\gamma = (1/\psi)$. Assim:

$$P(Y = y|\lambda, \psi) = k(\lambda, \psi) \left(\frac{1}{\psi} \right)^{1/2} e^{\left(-\frac{\lambda}{\psi}\right)} \left(\frac{e^{-\psi} \psi^y}{y!} \right) \left(\frac{e\lambda}{y} \right)^{\frac{y}{\psi}} \quad (3.1)$$

em que

$$\frac{1}{k(\lambda, \psi)} \simeq 1 + \frac{\psi - 1}{12\lambda} \left(1 + \frac{\psi}{\lambda} \right).$$

Consequentemente, se $\psi = 1$, tem-se a distribuição Poisson como caso particular e portanto $E(Y) = \text{Var}(Y)$, se $0 < \psi < 1$ então $\text{Var}(Y) < E(Y)$ e se $\psi > 1$ então $\text{Var}(Y) > E(Y)$.

3.2.2 Modelo de regressão *double* Poisson com efeito aleatório

Considere um estudo longitudinal com I indivíduos (unidade amostral) e T ocasiões de tempo. Seja a variável aleatória Y_{it} que representa a variável resposta do i -ésimo indivíduo ($i = 1, \dots, I$) e na t -ésima ocasião ($t = 1, \dots, T$). Considere que $Y_{it} \sim \text{DP}(\lambda_{it}, \psi)$ e para relacionar a variável resposta com as covariáveis ou variáveis explicativas, assume-se que todos os indivíduos são independentes e possuem um efeito aleatório específico denotado por b_i e considera-se que os efeitos aleatórios são variáveis aleatórias não observadas (variáveis latentes). Dessa forma, o modelo de regressão para dados correlacionados é caracterizado por uma estrutura hierárquica em que: $b_i \stackrel{iid}{\sim} N(0, \sigma_1^2)$ e $Y_{it}|b_i \stackrel{iid}{\sim} \text{DP}(\lambda_{it}, \psi)$, sendo

$$E(Y_{it}|\mathbf{b}_i) = \log(\lambda_{it}) = \mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{b}_i$$

em que,

$\mathbf{x}_{it} = (x_{it1}, \dots, x_{itp})^T$ é a i -ésima linha da matriz de planejamento conhecida associada aos efeitos fixos; $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ é o vetor de p parâmetros de regressão dos efeitos fixos desconhecidos, comuns a todos os indivíduos;

$\mathbf{z}_{it} = (z_{it}, \dots, z_{it})^T$ é a i -ésima linha da matriz de incidência conhecida dos efeitos aleatórios

$\mathbf{b}_i = (b_1, b_2, \dots, b_i)^T$ é o vetor dos efeitos aleatórios desconhecidos e específicos para cada indivíduo.

Assim, assumindo o modelo DP com efeito aleatório, doravante denominado de *double* Poisson normal (DPN), tem-se como caso especial o modelo Poisson com efeito aleatório, denominado de modelo Poisson normal (PN) quando $\psi = 1$.

3.2.2.1 Inclusão da variável *dummy* para o parâmetro de dispersão

A fim de controlar simultaneamente a sub- e superdispersão presentes nos dados, considera-se a inclusão de uma variável *dummy* associada ao parâmetro de dispersão ψ . Dessa forma, o modelo DPN com a inclusão da variável *dummy* será denominado de *double* Poisson normal *dummy* (DPND). Portanto, define-se:

$$\log(\psi_{it}) = d_{itk}^T \delta_{ik}$$

em que,

d_{itk} é o vetor de dimensão $(k \times 1)$ da variável *dummy* do parâmetro de dispersão;

δ_{ik} é o vetor de dimensão $(k \times 1)$ de incidência do parâmetro de dispersão. Em termos matriciais, o modelo pode ser reescrito como:

$$g_1(\lambda) = \exp(\mathbf{X}\beta + \mathbf{Z}\mathbf{b})$$

$$g_2(\psi) = \exp(\mathbf{D}_k \delta_k)$$

em que, $\mathbf{D}_k$ é uma matriz $(IT \times k)$ de variável *dummy* e $\delta_k = (\delta_1, \delta_2, \delta_3)^T$, indicando o parâmetro de dispersão. Dessa forma, é possível definir as variáveis *dummy* considerando dois cenários:

i) levando em consideração a média e a variância de cada indivíduo:

$$d_{it1} = \begin{cases} 1 & \text{se } \text{Var}(Y_i) < E(Y_i) \\ 0 & \text{caso contrário} \end{cases}, \quad d_{it2} = \begin{cases} 1 & \text{se } \text{Var}(Y_i) = E(Y_i) \\ 0 & \text{caso contrário} \end{cases} \quad \text{e} \quad d_{it3} = \begin{cases} 1 & \text{se } \text{Var}(Y_i) > E(Y_i) \\ 0 & \text{caso contrário} \end{cases}.$$

Assim, nas colunas da matriz $\mathbf{D}_k$ tem-se a indicação dos indivíduos que apresentaram subdispersão, equidispersão e superdispersão, respectivamente. Para exemplificar, supõe-se que $i = 5$ e $t = 2$. Assim, define-se:

$$\text{cenário I: } \mathbf{D}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \text{cenário II: } \mathbf{D}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{e} \quad \text{cenário III: } \mathbf{D}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

No cenário I todos os indivíduos apresentaram subdispersão, ou seja, representa o modelo DPN, enquanto no cenário II os indivíduos 1, 3, e 5 subdispersão, indivíduos 2 e 4 equidispersão e superdispersão,

respectivamente. No cenário III a interpretação é feita de modo análogo. Portanto, para os cenários II e III tem-se aplicação do modelo DPND.

- ii) A segunda opção é ajustar um modelo de regressão Poisson para cada unidade amostral e estimar o parâmetro de dispersão σ^2 . Quando $\hat{\sigma}^2 = 1$, representa equidispersão, quando $0 < \hat{\sigma}^2 < 1$, demonstra a presença de subdispersão e quando $\hat{\sigma}^2 > 1$, representa a superdispersão (McCullagh e Nelder, 1989; Winkelmann, 1995, 2008; Cameron e Trivedi, 2013). O parâmetro σ^2 pode ser estimado por:

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

em que X^2 é a estimativa da estatística de Pearson associada à regressão Poisson, com t observações e p parâmetros. Assim, as d_{ik} são dadas por:

$$d_{it1} = \begin{cases} 1 & \text{se } 0 < \hat{\sigma}^2 < 1 \\ 0 & \text{caso contrário} \end{cases}, \quad d_{it2} = \begin{cases} 1 & \text{se } \hat{\sigma}^2 = 1 \\ 0 & \text{caso contrário} \end{cases} \quad \text{e} \quad d_{it3} = \begin{cases} 1 & \text{se } \hat{\sigma}^2 > 1 \\ 0 & \text{caso contrário} \end{cases}.$$

A vantagem da segunda opção é que a construção da variável *dummy* considera as covariáveis envolvidas no modelo. A desvantagem surge do fato da dependência temporal não estar sendo levada em consideração, além da dimensão intraindivíduo, que de modo geral é pequena. Apesar dessas limitações é uma alternativa para discriminar os indivíduos. É importante enfatizar que a estruturação da variável *dummy* não é imutável e sua construção pode ser definida de vários aspectos distintos.

Ademais, a estrutura da variável *dummy* reflete na suposição de independência condicional e toda dependência entre as medidas na mesma unidade amostral são decorrentes dos efeitos aleatórios.

3.2.3 Estimação dos parâmetros

A estimação dos parâmetros envolvidos nos modelos DPN e DPND pode ser feita utilizando-se o método da máxima verossimilhança. Neste caso, o método consiste em maximizar o logaritmo da função de verossimilhança marginal, obtida pela integração da função de verossimilhança em relação aos efeitos aleatórios b_i . A contribuição para verossimilhança do i -ésimo indivíduo é dado por:

$$f(y_{it}|\beta, \delta, \sigma_1^2) = \int \dots \int \prod_{t=1}^T f(y_{it}|b_i, \beta, \delta) f(b_i|\sigma_1^2) db_i,$$

em que as distribuições foram definidas na seção 3.2. Sob a suposição de independência entre os indivíduos tem-se que a função de verossimilhança marginal é dada pela expressão:

$$\begin{aligned} L(\beta, \delta, \sigma_1^2) &= \prod_{i=1}^I f(y_{it}|\beta, \delta, \sigma_1^2) \\ &= \prod_i \int \dots \int \prod_{t=1}^T f(y_{it}|b_i, \beta, \delta) f(b_i|\sigma_1^2) db_i. \end{aligned} \quad (3.2)$$

A maior dificuldade nesta maximização é a presença das I integrais sobre os efeitos aleatórios. Alguns casos especiais podem ser resolvidos de forma analítica, como o modelo normal e probit-normal (Mollenberghs e Verbeke, 2005). Para os modelos DPN e DPND, a integral 3.2 não possui solução analítica,

fazendo-se necessário o uso de métodos de integração numérica para a sua solução. Uma alternativa é o método de integração numérica de Gauss-Hermite, descrito em Liu e Pierce (1994) e apresentado na literatura específica de dados longitudinais por Lesaffre e Spiessens (2001) e Molenberghs e Verbeke (2005). A quadratura de Gauss-Hermite é aplicada para resolver integrais da forma:

$$\int_{-\infty}^{\infty} f(x)w(x)dx = \int_{-\infty}^{\infty} f(x)\exp(-x^2)dx.$$

E a integral é aproximada por uma soma ponderada é dada por:

$$\int_{-\infty}^{\infty} f(x)w(x)dx \simeq \sum_{s=1}^m w_s \exp(x_s^2)g(x_s),$$

em que m é o número de pontos utilizado na aproximação, ressaltando-se que quanto maior o número de pontos, mais precisa será a aproximação, e w_s está associado ao peso dado por:

$$w_s = \frac{2^{m-1}q!\sqrt{\pi}}{m^2(H_{m-1}(x_s))^2},$$

e $H_m(x_s)$ são as raízes do polinômio de Hermite representadas por:

$$H_m(x) = (-1)^m \exp(-x^2) \frac{d^m}{dx^m} \exp(-x^2).$$

Para a solução aproximada da integral pelo método de Gauss-Hermite é necessário obter os pesos de integração (w_s) e dos pontos x_s . Uma desvantagem do método de Gauss-Hermite é que a qualidade da aproximação da integral é maior se $f(x)w(x) = f(x)\exp(-x^2)$ se comportarem de forma similar a $w(x)$, o que nem sempre ocorre nos problemas gerais (Molenberghs e Verbeke, 2005). Para essas situações a quadratura adaptativa de Gauss-Hermite surge como um procedimento para reescalonar e deslocar os pontos de quadratura e aumentar a precisão. Assim, supõe-se que $f(x)w(x)$ segue uma distribuição normal, na qual a média corresponde a moda de $\hat{x}$ de $\ln(f(x)w(x))$ e a variância é igual a:

$$\left[\frac{d^2}{dx^2} \ln[f(x)w(x)] \Big|_{x=\hat{x}} \right]^{-1}.$$

Dessa forma, os novos pontos de quadratura adaptativos são dados por:

$$x_s^+ = \hat{x} + \left[-\frac{d^2}{dx^2} \ln[f(x)w(x)] \Big|_{x=\hat{x}} \right]^{-0.5} x_s,$$

com seus respectivos pesos:

$$w_s^+ = \left[-\frac{d^2}{dx^2} \ln[f(x)w(x)] \Big|_{x=\hat{x}} \right]^{-0.5} \frac{w(x_s^+)}{w(x_s)} w_s.$$

Portanto, a integral passa a ser aproximada por:

$$\int f(x)w(x)dx \approx \sum_{s=1}^m w_s^+ f(x_s^+).$$

Dessa forma, o método Gauss-Hermite ou adaptativa Gauss-Hermite é utilizado para determinar uma aproximação da integral 3.2. Em geral, quanto maior é a ordem de m pontos de integração, melhor será a aproximação da área. Porém, quanto maior é o número de pontos maior será o esforço computacional para a aproximação. A vantagem da quadratura adaptativa de Gauss-Hermite é que se faz

necessária uma quantidade menor de pontos para uma boa aproximação quando comparado ao método de Gauss-Hermite (Molenberghs e Verbeke, 2005). No entanto, o método da adaptativa de Gauss-Hermite requer o cálculo de $\hat{x}$ para cada indivíduo e essa maximização numérica torna o processo computacional intenso. Outro fator que dificulta o processo de estimação é que o integrando está em função dos parâmetros desconhecidos β , σ_1^2 e δ , os pontos de quadratura, bem como os pesos utilizados na adaptativa de Gauss-Hermite dependem destes parâmetros, e assim precisam ser atualizados a cada passo do processo de estimação iterativo (Molenberghs e Verbeke, 2005). A função `gauss.quad()` do pacote `statmod` (Smyth et al., 2013) do software (R Core Team, 2019) nos proporciona uma ferramenta para o cálculo dos pesos e os pontos de Gauss-Hermite.

3.2.3.1 Predição de b_i via *empirical* Bayes

Para predizer o vetor $\mathbf{b}_i$ pode-se utilizar a abordagem conhecida como *empirical* Bayes ou BLUP (*best linear unbiased predictors*). As predições de $\mathbf{b}_i$ são fundamentais para identificar a variabilidade entre as unidades amostrais e detectar perfis individuais ao longo do tempo. Na prática, tem-se interesse em estimar $E(Y_i|b_i)$, assim é possível avaliar medidas de probabilidades e análise de diagnóstico (Thomas e Have, 1996). O processo para predição consiste em utilizar as estimativas de β_j , σ_1 e δ_k , obtidas pelo método de estimação de máxima verossimilhança e aplicá-las na distribuição marginal a *posteriori* de b_i , dada por:

$$f(\mathbf{b}_i|y_i, \beta, \delta, \sigma_1^2) = \frac{f(y_i|\mathbf{b}_i, \beta, \delta)f(\mathbf{b}_i|\sigma_1^2)}{\int f(y_i|\mathbf{b}_i, \beta, \delta)f(\mathbf{b}_i|\sigma_1^2)d\mathbf{b}_i}.$$

Para maximizar a função de verossimilhança dada em 3.2 e obter as estimativas dos parâmetros, assim como a predição de b_i , utiliza-se o pacote GAMLSS (Rigby e Stasinopoulos, 2005), disponível no software estatístico R (R Core Team, 2019).

3.3 Análise dos resíduos

Avaliar a qualidade do ajuste de um modelo não é uma tarefa simples e a análise dos resíduos é uma etapa fundamental para a seleção do modelo estatístico, que tem por objetivo verificar as pressuposições do modelo adotado e em relação ao conjunto de dados, bem como identificar valores discrepantes e/ou influentes. As técnicas de diagnóstico baseiam-se na análise de resíduos e podem ser formais (como testes) ou informais (como gráficos). Essas técnicas são de grande importância para a validação de um modelo, pois se as pressuposições forem violadas, as previsões e inferências do modelo podem ser enganosas. Nobre e Motta Singer (2007) estudaram os resíduos marginais e condicionais para os modelos mistos e recomendam o uso dos resíduos condicionais para avaliar as suposições de normalidade e homoscedasticidade de variâncias. Os autores também sugerem o uso das predições de $\mathbf{b}_i$ para identificar a presença de unidades amostrais discrepantes (*outliers*), bem como avaliar a normalidade dos efeitos aleatórios. Portanto, os resíduos marginais e condicionais são dados por:

- i) Resíduo marginal: $\hat{\xi}_m = \mathbf{y} - \mathbf{X}\hat{\beta}$ que fazem a predição dos erros marginais;
- ii) Resíduo condicional: $\hat{\xi}_c = \mathbf{y} - \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{\mathbf{b}}$ que fazem a predição dos erros condicionais.

No entanto, neste trabalho, utilizam-se os resíduos quantílicos aleatorizados (Dunn e Smyth, 1996), que são utilizados para análise de diagnóstico no modelo GAMLSS (Stasinopoulos et al., 2017).

A principal vantagem dos resíduos quantílicos aleatorizados é que independente da distribuição de probabilidade para a variável resposta, os resíduos sempre terão uma distribuição normal padrão quando o modelo for corretamente especificado (Dunn e Smyth, 1996). Os resíduos são calculados de duas formas distintas conforme a natureza discreta ou contínua da variável resposta. Como o foco deste trabalho diz respeito a dados de contagem, apresenta-se o cálculo para o caso discreto. Assim, se y_{it} é uma observação de uma variável aleatória discreta, logo $F(y_{it}|\theta)$ é uma função de distribuição acumulada. A técnica de transformação de probabilidade integral consiste em definir u como um valor aleatório de uma distribuição de probabilidade uniforme nos intervalos:

$$[u_1, u_2] = [F(y_{it} - 1|\theta), F(y_{it}|\theta)] \quad \text{e} \quad [\hat{u}_1, \hat{u}_2] = [F(y_{it} - 1|\hat{\theta}), F(y_{it}|\hat{\theta})].$$

Para facilitar a compreensão do processo de transformação de probabilidade integral discreta, considere o gráfico da Figura 3.1, desenvolvido por Stasinopoulos et al. (2017, p. 421). Para uma função de probabilidade discreta (Figura 3.1a), o valor de y_{it} é transformado em um intervalo $(u_1, u_2]$ por meio da função de distribuição acumulada (Figura 3.1b). Portanto, u é selecionado de uma distribuição uniforme com intervalo $(u_1, u_2]$ e finalmente u é transformado no resíduo (Figura 3.1c), denominado de z - score, definido por:

$$\hat{r} = \Phi^{-1}(\hat{u}),$$

em que Φ^{-1} é a função inversa da função distribuição acumulada da distribuição normal padrão. Se o modelo é correto, então os resíduos têm distribuição normal padrão e, similarmente, os resíduos se aproximam de uma distribuição normal padrão (Stasinopoulos et al., 2017).

Figura 3.1: Gráfico do processo de transformação de probabilidade integral discreta. (Extraído de (Stasinopoulos et al., 2017, p. 421))

Por se tratar de um processo aleatório, é prudente estudar não apenas um conjunto de resíduos, mas sim de uma sequência de várias aleatorizações antes de se concluir sobre a adequação do modelo (Stasinopoulos et al., 2017).

Como técnicas para avaliar o ajuste, atualmente existem diversas ferramentas disponíveis na literatura para este fim, com destaque para o gráfico do tipo quantis-quantis (QQ), em que tem-se os resíduos ordenados (eixo y) e os quantis teóricos obtidos da distribuição normal padrão (eixo x). Quando a distribuição dos resíduos é gaussiana, espera-se que os resíduos estejam dispostos em uma reta com inclinação de 45° . No entanto, nem sempre é fácil avaliar se os pontos estão alinhados ou se as irregularidades são causadas devido à aleatoriedade. Com o objetivo de simplificar a interpretação do gráfico, Atkinson (1985) propôs a construção de bandas de confiança, denominadas de envelopes simulados. Essas bandas de confiança são obtidas por meio de simulações computacionais e o resultado esperado para um modelo bem ajustado é que os valores dos resíduos ordenados se distribuam aleatoriamente no interior do envelope simulado. Assim, torna-se possível identificar a presença de observações discrepantes (*outliers*) e verificar se a distribuição de probabilidade foi corretamente especificada. Demétrio et al. (2014) utilizam o gráfico meio normal de probabilidade (*half-normal plots* (hnp)), em que os resíduos são considerados em valores absolutos, para avaliar a qualidade do ajuste dos modelos para dados de contagem com superdispersão. Moral et al. (2017) apresentam a implementação da função hnp no software R Core Team (2019) que fornece funções para gerar gráficos meio normal de probabilidade com um envelope simulado para uma série de modelos lineares generalizados e extensões. O gráfico hnp para o modelo PN pode ser obtido por meio deste pacote.

Dado que para os modelos DPN e DPND o pacote ainda não disponibiliza a elaboração dos gráficos, foi desenvolvida uma função específica para este fim. Para sua implementação fez-se necessário plotar os valores absolutos ordenados de uma medida de diagnóstico adequada (neste trabalho utilizou-se os resíduos quantílicos aleatorizados) contra os correspondentes valores esperados das estatísticas de ordem, em valor absoluto, da distribuição meio-normal dada por:

$$\Phi^{-1} \left[\frac{(i + n - \frac{1}{8})}{2n + \frac{1}{2}} \right],$$

em que Φ^{-1} é a função inversa da função distribuição acumulada da distribuição normal padrão, no qual $i = 1, \dots, n$, sendo $n = IT$ a dimensão da amostra. Abaixo são descritos os passos para a construção de um *half-normal plot* segundo Atkinson (1985) e Moral et al. (2017):

- i. Ajustar o modelo;
- ii. Calcular o resíduo pertinente representado por r_i , em valor absoluto e colocá-los em ordem crescente;
- iii. Simular 99 (ou mais) amostras para a variável resposta, utilizando as mesmas covariáveis e estimativas dos parâmetros do modelo i ;
- iv. Ajustar o mesmo modelo para cada variável resposta obtida em iii e calcular os novos resíduos $r_{j(i)}$, em que $j = 1, \dots, 99$, $i = 1, \dots, n$ e classificá-los em valores absolutos e dispor esses valores em ordem crescente;

- v. Para cada resíduo obtido em iv, calcular os percentis (e.g. 2,5%, 50% e 97,5%) desejados dos valores de diagnóstico simulado para cada valor da ordem estatística desejada e utilizá-los para gerar o envelope simulado;
- vi. Plotar os valores do resultado obtido em ii e v contra as estatísticas esperadas da distribuição meio-normal.

Se o modelo em estudo estiver correto, os valores observados devem estar distribuídos aleatoriamente no interior do envelope simulado.

A análise de resíduos por meio do gráfico hnp é uma ferramenta importante para a seleção do modelo estatístico adequado, dada a capacidade de verificar se, com base em simulações computacionais extraídas de informações *a priori* do modelo ajustado, as estimativas dos parâmetros do modelo representam efetivamente o comportamento real do conjunto de dados. É importante salientar que outras técnicas de diagnóstico podem ser utilizadas para garantir a seleção do modelo estatístico mais adequado. Para os modelos mistos, a questão sobre qual tipo de resíduo deve ser utilizada para avaliar a qualidade do ajuste do modelo selecionado é uma área de pesquisa ainda em aberto (Nobre e Motta Singer, 2007).

3.4 Estudo de simulação

Para avaliar a flexibilidade do modelo *double* Poisson com intercepto aleatório bem como as propriedades do estimador de máxima verossimilhança sob a condição de dados de contagem longitudinais não equidispersos, realizou-se um estudo de simulação.

Consideram-se neste trabalho 45 cenários, dados pelas combinações da condição de dispersão: (subdisperso, superdisperso e sub- e superdispersos), dos tamanhos de amostra ($I = 40, 60, 80, 100$ e 500) e o número de medidas repetidas no tempo de $T = 4, 5, 6$. Para cada cenário foram simulados 1000 conjuntos de dados. Para os cenários de subdispersão e superdispersão, os parâmetros de dispersão foram fixados em $\delta = 0, 3$ e 3 , respectivamente. Estes mesmos valores foram utilizados para o modelo DPND, ou seja, considerando as condições de sub- e superdispersão conjuntamente. Dessa forma, tem-se que $g_2(\psi_{it}) = d_{it1}\delta_1 + d_{it2}\delta_2$, em que:

$\delta_1 = 0, 3$ o parâmetro de dispersão, representando os indivíduos com subdispersão;

$\delta_2 = 3$ o parâmetro de dispersão, representando os indivíduos com superdispersão;

$d_{it1} = 1$ para $i = 1, 2, \dots, (I/2)$ e $d_{it2} = 0$ para $i = I/2 + 1, \dots, I$;

$d_{it1} = 0$ para $i = 1, 2, \dots, (I/2)$ e $d_{it2} = 1$ para $i = I/2 + 1, \dots, I$.

Para a estrutura de modelo de regressão com efeito aleatório, especifica-se o vetor de média condicionado aos efeitos aleatórios, tal como: $E(Y_i|b_i) = \exp[(\beta_0 + b_i) + \beta_1 x_{it}]$, em que x_{it} é dado por:

$$x_{it} = \begin{cases} -0,5 & (i = 1, \dots, I/5) \\ 0 & (i = (I/5) + 1, \dots, 4I/5) \\ 0,5 & (i = (4I/5) + 1, \dots, I). \end{cases}$$

Esta estrutura de covariável foi aplicada por Khan e Jowaheer (2013b) e adaptada para o estudo em questão. Os coeficientes de regressão foram fixados em $\beta_0 = \beta_1 = 1$. Assume-se também $b_i \sim N(0, \sigma_1^2)$

e $\sigma_1^2 = 0,81$. Para simular os dados de contagem para os cenários propostos por meio de um processo aleatório, elaborou-se uma função que encontra-se disponível no apêndice. Optou-se por apresentar os resultados para $T = 5$, visto que para $T = 4$ e $T = 6$, os valores foram similares, indicando que a dimensão do perfil individual (número de medidas repetidas no tempo), particularmente neste estudo, não representou uma fonte de variação nas estimativas dos parâmetros. Na Tabela 3.1 encontram-se os resultados das simulações para os 15 cenários, com base nas 1000 simulações. Nota-se que para os cenários de subdispersão e sub- e superdispersão conjuntamente, os valores médios das estimativas dos parâmetros β_0 , β_1 , σ_1^2 , δ_1 e δ_2 são próximos dos verdadeiros valores dos parâmetros considerados. Também verifica-se que há uma redução no coeficiente de variação à medida que o número de indivíduos aumenta, isto é, a variabilidade em torno do valor estimado dos parâmetros é menor.

Tabela 3.1: Estatísticas descritivas das estimativas dos parâmetros para o modelo *double* Poisson normal, ajustado a dados simulados, considerando os cenários de subdispersão ($\delta = 0,3$), superdispersão ($\delta = 3$) e sub- e superdispersão conjuntamente ($\delta_1 = 0,3$ e $\delta_2 = 3$) e $T = 5$ medidas repetidas no tempo de acordo com o número de indivíduos (I) e valores populacionais fixados em $\beta_0 = \beta_1 = 1$ e $\sigma_1^2 = 0,81$.

δ	I	Descritiva	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\delta}$	$\hat{\sigma}_1^2$	
0,3	40	Média	1,0158	1,0030	0,2457	0,8095	
		Desvio padrão	0,1531	0,4738	0,0307	0,2032	
		Coefficiente de variação	15,07%	47,24%	12,49%	25,10%	
		Viés médio	1,58%	0,30%	-18,12%	-0,06%	
	60	Média	1,0173	1,0071	0,2459	0,8173	
		Desvio padrão	0,1227	0,3802	0,0255	0,1582	
		Coefficiente de variação	12,06%	37,75%	10,39%	19,36%	
		Vies médio	1,73%	0,71%	-18,05%	0,91%	
	80	Média	1,0137	0,9972	0,2470	0,8187	
		Desvio padrão	0,1050	0,3096	0,0206	0,1375	
		Coefficiente de variação	10,36%	31,05%	8,35%	16,79%	
		Viés médio	1,37%	-0,28%	-17,67%	1,07%	
	100	Média	1,0188	0,9796	0,2459	0,8260	
		Desvio padrão	0,0911	0,2888	0,0194	0,1269	
		Coefficiente de variação	8,94%	29,48%	7,90%	15,37%	
		Viés médio	1,88%	-2,04%	-18,05%	1,98%	
	500	Média	1,009	1,001	0,2474	0,8475	
		Desvio padrão	0,0399	0,1289	0,0090	0,0576	
		Coefficiente de variação	3,96%	12,87%	3,65%	6,80%	
		Viés médio	0,96%	0,19%	-17,53	4,64%	
δ	I	Descritiva	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\delta}$	$\hat{\sigma}_1^2$	
3	40	Média	1,1255	0,8967	2,2530	0,7067	
		Desvio padrão	0,1494	0,4685	0,2754	0,2054	
		Coefficiente de variação	13,27%	52,25%	12,22%	29,07%	
		Viés médio	12,55%	-10,33%	-24,90%	-12,75%	
	60	Média	1,1275	0,9000	2,2639	0,7339	
		Desvio padrão	0,1253	0,3775	0,2407	0,1749	
		Coefficiente de variação	11,12%	41,94%	10,63%	23,83%	
		Viés médio	12,77%	-10,12%	-24,54%	-9,35%	
	80	Média	1,1217	0,8989	2,2460	0,7398	
		Desvio padrão	0,1047	0,3459	0,1932	0,1491	
		Coefficiente de variação	9,33%	38,48%	8,60%	20,16%	
		Viés médio	12,17%	-10,11%	-25,13%	-8,67%	
	100	Média	1,1214	0,9091	2,2360	0,7436	
		Desvio padrão	0,0997	0,3002	0,1790	0,1321	
		Coefficiente de variação	8,89%	33,02%	8,00%	17,76%	
		Viés médio	12,14%	-9,09%	-25,47%	-8,20%	
	500	Média	1,1267	0,8982	2,2354	0,7524	
		Desvio padrão	0,0436	0,1301	0,0790	0,0613	
		Coefficiente de variação	3,87%	14,48%	3,53%	8,15%	
		Viés médio	12,67%	-10,18%	-25,49%	-7,11%	
δ_k	I	Descritiva	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\sigma}_1^2$
$\delta_1 = 0,3$ e $\delta_2 = 3$	40	Média	1,0250	1,0377	0,3022	2,7489	0,8274
		Desvio padrão	0,1566	0,4528	0,0455	0,4729	0,2218
		Coefficiente de variação	15,28%	43,64%	15,05%	17,20%	26,81%
		Viés médio	2,50%	3,77%	0,72%	-8,37%	2,14%
	60	Média	1,0321	1,0029	0,3003	2,7082	0,8298
		Desvio padrão	0,1272	0,3283	0,0378	0,3554	0,1743
		Coefficiente de variação	12,32%	32,74%	12,59%	13,12%	21,00%
		Viés médio	3,21%	0,29%	0,12%	-9,73%	2,45%
	80	Média	1,0321	1,0024	0,3029	2,6951	0,8384
		Desvio padrão	0,1086	0,3328	0,0327	0,3270	0,1550
		Coefficiente de variação	10,52%	33,20%	10,79%	12,13%	18,49%
		Viés médio	3,21%	0,24%	0,98%	-10,16%	3,51%
	100	Média	1,0362	1,0091	0,3032	2,6887	0,8408
		Desvio padrão	0,0915	0,3031	0,0286	0,2738	0,1279
		Coefficiente de variação	8,83%	30,03%	9,44%	10,18%	15,21%
		Viés médio	3,62%	0,91%	1,06%	-10,38%	3,80%
	500	Média	1,0328	1,0055	0,3031	2,6877	0,8556
		Desvio padrão	0,0425	0,1275	0,0133	0,1205	0,0597
		Coefficiente de variação	4,12%	12,68%	4,38%	4,48%	6,97%
		Viés médio	3,28%	0,55%	1,05%	-10,41%	5,62%

Por meio das Figuras 3.2a e 3.2c observa-se que, de modo geral, o viés médio em percentual ($B(\hat{\theta}) = (E(\hat{\theta}) - \theta)/\theta$, em que θ denotam os valores verdadeiros dos parâmetros β_0 , β_1 , σ_1^2 , δ , δ_1 e δ_2 e $E(\hat{\theta})$ representa a média das estimativas dos parâmetros ajustado aos dados simulados) é próximo de zero. Entretanto, para os parâmetros δ e δ_2 os vieses médios foram de -18% e -10% para os respectivos cenários da subdispersão e sub- e superdispersão conjuntamente, indicando que os parâmetros de dispersão foram subestimados. Por outro lado, para o cenário da superdispersão, o viés médio em percentual é maior quando comparado aos outros cenários (Figura 3.2b), conforme é esperado quando se tem um nível de dispersão maior.

Figura 3.2: Gráficos dos vieses médios em percentual das estimativas dos parâmetros do modelo *double* Poisson normal para dados simulados, considerando os cenários de subdispersão ($\delta = 0, 3$), superdispersão ($\delta = 3$) e sub- e superdispersão conjuntamente ($\delta_1 = 0, 3$ e $\delta_2 = 3$) e $T = 5$ medidas repetidas no tempo. Valores populacionais fixados em $\beta_0 = \beta_1 = 1$ e $\sigma_1^2 = 0,81$.

Os vícios das estimativas para cada um dos cenários simulados são apresentados na Figura 3.3. De modo geral, para todos os cenários verifica-se que, a medida que o número de indivíduos aumenta o coeficiente variação diminui (Tabela 3.1), isto é, a diferença entre o valor estimado e o verdadeiro valor do parâmetro é reduzida (Figura 3.3).

Figura 3.3: *Box plot* para os vieses das simulações do modelo *double* Poisson normal, considerando os cenários de subdispersão ($\delta = 0,3$; a - e), superdispersão ($\delta = 3$; f - j) e sub- e superdispersão conjuntamente ($\delta_1 = 0,3$ e $\delta_2 = 3$; k - o), com o número de indivíduos variando entre 40 e 500 em cada cenário, e seus coeficientes de regressão fixados em $\beta_0 = \beta_1 = 1$, $\sigma_1^2 = 0,81$ e $T = 5$ medidas repetidas no tempo.

Ao comparar os resultados da simulação dos cenários subdispersos e superdispersos com o cenário sub- e superdispersos conjuntamente, verifica-se que a metodologia proposta possibilitou estimativas dos parâmetros, em média, próximas da verdadeira, dado que os vieses médios em percentual são próximos de zero, indicando que as estimativas geradas se aproximam dos verdadeiros valores. Neste contexto o uso da variável *dummy* é uma alternativa em particionar a variabilidade presente nos indivíduos com diferentes níveis de dispersão, contribuindo para a melhor interpretação dos dados.

3.5 Aplicação

3.5.1 Estudo de Motivação: análise de crises epiléticas

Para ilustrar o método proposto neste trabalho, utiliza-se como motivação dados produzidos por um estudo clínico realizado por Leppik et al. (1985), que se tornou conhecido na literatura após a análise realizada por Thall e Vail (1990). Breslow e Clayton (1993), Diggle et al. (2002), Rigby e Stasinopoulos (2005), Booth et al. (2003), Tapia et al. (2019) também analisaram os dados desse estudo destacando o fenômeno da superdispersão. O estudo foi realizado com pacientes portadores de epilepsia, cujo objetivo foi verificar se o uso da droga Progabide reduz ou não a frequência de crises epiléticas dos pacientes. O ensaio clínico foi realizado com 59 pacientes, dos quais 31 receberam a droga e 28 receberam o tratamento controle (placebo). Todos os pacientes receberam conjuntamente a quimioterapia padrão. O ensaio teve duração de 4 períodos, cada período com duração de duas semanas. A cada consulta clínica realizada o número de crises epiléticas nas duas semanas anteriores à consulta era relatada. Maiores detalhes, assim como o conjunto de dados podem ser encontrados no pacote HSAUR (Brian e Torsten, 2017) do software (R Core Team, 2019).

Primeiramente foi realizada uma análise exploratória com a finalidade de conhecer o comportamento dos dados. Os perfis individuais das crises epiléticas dos pacientes durante o ensaio clínico são mostrados nas Figuras 3.4a, na qual se considera a estratificação por tratamento e Figura 3.4b, isoladamente para cada um dos pacientes. Apesar dos pacientes 25 e 49 terem apresentado as maiores variações no número de crises ao longo do estudo quando comparados aos demais, de modo geral, não se verifica um padrão claro nos perfis que seja capaz de diferenciar o tratamento placebo do Progabide. Em outras palavras, não é possível observar se há uma tendência de aumento ou diminuição das crises epiléticas ao longo do tempo em ambos os tratamentos. Porém ao observar a Figura 3.4b nota-se um possível efeito aleatório nos perfis dos pacientes, uma vez que os dados observados podem ser afetados por comportamentos ou estilos de vida individuais e independentes um do outro.

(a)

(b)

Figura 3.4: Gráficos de perfis individuais do número de crises epilépticas para os pacientes dos tratamentos placebo e Progabide (a); curvas isoladas para cada um dos 59 pacientes do ensaio clínico (b) ao longo dos quatro períodos amostrados, para os dados do ensaio clínico realizado por Leppik et al. (1985).

O número médio de crises epiléticas e as variâncias por período amostrado, são apresentados na Tabela 3.2. Verifica-se que para cada ocasião a variância é maior do que a média. Esse fato foi considerado pelos autores Thall e Vail (1990), Breslow e Clayton (1993), Diggle et al. (2002), Rigby e Stasinopoulos (2005), Booth et al. (2003) e Tapia et al. (2019) para tratar o problema como superdispersão.

Tabela 3.2: Medidas descritivas (média e variância) do número de crises epiléticas ao longo dos quatro períodos amostrados para os tratamentos placebo e Progabide, para os dados do ensaio clínico realizado por Leppik et al. (1985).

Tratamento	Estatística Descritiva	Visitas			
		1	2	3	4
placebo	Média	9,36	8,29	8,79	7,96
	Variância	102,76	66,66	215,29	58,18
Progabide	Média	8,58	8,42	8,13	6,71
	Variância	332,72	140,65	193,05	126,88

A média geral de crises epiléticas foi de 8,60 e 7,96 para os tratamentos placebo e Progabide, respectivamente, apontando uma redução de 7,44% no número de crises epiléticas após o uso do medicamento Progabide em comparação ao tratamento placebo. As variâncias para os tratamentos placebo e Progabide foram de 108,00 e 194,03, respectivamente. Nota-se que a variância foi superior à média em torno de 12,56 vezes para o tratamento placebo e 24,37 para o tratamento Progabide, havendo indício de superdispersão. A ocorrência deste fenômeno se justifica pelo fato de ter sido levado em consideração apenas a variação interindividual, ou seja, a variação entre os indivíduos.

Os pacientes 5, 8, 10, 11, 15, 16, 25, 49, 53 e 56 apresentaram maior grau de dispersão, pois seus intervalos interquartílicos (*i.e.* a diferença entre o terceiro quartil e o primeiro quartil) se destacam dos demais pacientes, conforme observa-se na Figura 3.5. Em contrapartida, os pacientes 1, 2, 9, 13, 19, 21, 22, 23, 26, 27, 28, 34, 40, 46, 48, 50, 51, 55, 57, 58 e 59 exibiram menor grau de dispersão.

Figura 3.5: *Box plot* do número de crises epiléticas para os pacientes dos tratamentos placebo e Progabide, para os dados do ensaio clínico realizado por Leppik et al. (1985).

Ao considerar a média e a variância intraindividualmente, isto é, a resposta dentro do mesmo indivíduo ao longo do tempo, observa-se que os tratamentos placebo e Progabide apresentaram pacientes com subdispersão, equidispersão e superdispersão. Pacientes que apresentaram subdispersão, equidispersão e superdispersão representaram 45,76%, 1,69% e 52,55% da amostra, respectivamente. Na Tabela 3.3 são apresentadas as medidas descritivas dos pacientes, com destaque para aqueles que apresentaram subdispersão.

Tabela 3.3: Medidas descritivas (média e variância) do número de crises epiléticas dos pacientes submetidos ao tratamento placebo (pacientes 1 a 28) e tratamento Progabide (pacientes 29 a 59), para os dados do ensaio clínico realizado por Leppik et al. (1985). Números com destaque em negrito indicam os pacientes que apresentaram subdispersão.

Paciente	1	2	3	4	5	6	7	8	9	10
Média	3,50	3,50	2,75	3,25	13,75	5,50	3,00	23,75	5,50	8,25
Variância	1,00	1,00	4,92	2,25	46,25	7,00	6,67	138,92	0,33	42,92
Paciente	11	12	13	14	15	16	17	18	19	20
Média	16,50	7,50	4,00	10,50	14,75	4,00	1,50	30,75	3,75	4,00
Variância	83,67	11,67	2,67	9,67	47,58	27,33	3,00	17,58	2,25	10,00
Paciente	21	22	23	24	25	26	27	28	29	30
Média	3,50	3,50	3,25	7,50	35,75	1,50	2,50	13,25	10,50	7,00
Variância	0,33	0,33	1,58	17,00	729,58	0,33	1,67	1,58	7,00	4,67
Paciente	31	32	33	34	35	36	37	38	39	40
Média	1,75	3,25	4,75	2,75	18,50	5,00	2,50	6,00	7,25	1,00
Variância	4,25	4,25	4,92	1,58	7,00	2,00	3,67	4,00	52,92	0,67
Paciente	41	42	43	44	45	46	47	48	49	50
Média	1,50	3,00	16,25	6,50	9,75	1,75	8,00	0,75	75,50	3,25
Variância	3,67	4,67	36,92	9,67	38,25	0,92	2,67	0,92	327,00	0,92
Paciente	51	52	53	54	55	56	57	58	59	
Média	6,50	2,50	17,50	3,25	3,75	12,75	1,50	0,00	2,50	
Variância	1,67	3,67	57,67	6,25	0,92	101,58	1,67	0,00	1,67	

3.5.2 Formulação do modelo estatístico

Para estabelecer a notação, considere a variável resposta Y_{it} como sendo o número de crises epiléticas, em que $i = 1, \dots, 59$ e $t = 1, \dots, 4$. Propõe-se uma sequência de 4 preditores lineares com a finalidade de testar as hipóteses do ensaio clínico, ou seja, verificar se o medicamento Progabide auxilia no tratamento da epilepsia ao longo do tempo, com a inclusão da variável *dummy* para o parâmetro de dispersão ψ da distribuição DPN, conforme descrito na seção 3.2.2.1, levando em consideração a média e a variância de cada indivíduo (Tabela 3.3). Como o paciente 58 não manifestou crises epiléticas durante o período do estudo, resultando em média e variância iguais a zero (Tabela 3.3), somou-se uma constante arbitrariamente pequena (0,00001) a fim de estimar o parâmetro de dispersão. Dessa forma, o paciente 58 será considerado como um caso de subdispersão. Portanto:

$$g_2(\psi_{it}) = \exp(d_{it1}\delta_1 + d_{it2}\delta_2),$$

em que:

δ_1 é o parâmetro de dispersão para os pacientes com subdispersão;

δ_2 é o parâmetro de dispersão para os pacientes com superdispersão;

$d_{it1} = 1$ e $d_{it2} = 0$ quando o paciente apresenta subdispersão;

$d_{it1} = 0$ e $d_{it2} = 1$ quando o paciente apresenta superdispersão.

Os preditores lineares são dados por:

$$\begin{aligned}
\text{Preditor 1} - g_1(\lambda_{it}) &= \exp[\beta_0 + b_i]; \\
\text{Preditor 2} - g_1(\lambda_{it}) &= \exp[(\beta_0 + b_i) + \beta_1 t_{it}]; \\
\text{Preditor 3} - g_1(\lambda_{it}) &= \exp[(\beta_0 + b_i) + \beta_1 t_{it} + \beta_2 \text{Trat}_i]; \\
\text{Preditor 4} - g_1(\lambda_{it}) &= \exp[(\beta_0 + b_i) + \beta_1 t_{it} + \beta_2 \text{Trat}_i + \beta_3 \text{Trat}_i t_{it}].
\end{aligned}$$

Em que:

β_1 é o efeito longitudinal;

β_2 é o efeito de tratamento;

β_3 é o efeito da interação entre o tratamento e o efeito longitudinal;

t_{it} indica a visita do indivíduo (paciente) ao médico, considerada neste estudo como uma variável quantitativa;

$\text{Trat}_i = 0$ se o i -ésimo paciente recebeu o tratamento placebo ou 1 se recebeu o tratamento Progabide ;

b_i é o efeito aleatório do indivíduo (intercepto aleatório), para o qual supõe-se $b_i \sim N(0, \sigma_1^2)$.

Para cada um dos preditores lineares, consideram-se às distribuições PN, DPN e DPND para a variável aleatória Y_{it} , número de crises epiléticas. As estruturas dos preditores lineares foram construídas de forma aninhada para facilitar a elaboração de testes de hipóteses para os parâmetros, baseados na teoria da verossimilhança, ou seja, verificar se $H_{01} : \beta_1 = 0$, $H_{02} : \beta_2 = 0$ e $H_{03} : \beta_3 = 0$ são verdadeiras ou falsas. O preditor 1 contém apenas o intercepto, o preditor 2 inclui o efeito longitudinal, o preditor 3 corresponde ao efeito longitudinal mais o efeito de tratamento, e por último, o preditor 4, que corresponde ao preditor 3 somado ao efeito de interação entre o tratamento e o tempo. As estimativas dos parâmetros foram obtidas por meio do pacote GAMLSS (Rigby e Stasinopoulos, 2005) do software R Core Team (2019).

3.5.3 Resultados e Discussão

O logaritmo natural da função de verossimilhança o teste de razão de verossimilhança para uma sequência de preditores lineares para as distribuições PN, DPN e DPND, AIC e teste de hipótese para o parâmetro de dispersão são apresentados na Tabela 3.4. Os modelos PN e DPN são apresentados apenas para efeito de comparação dos resultados, visto que o objetivo do estudo é analisar a subdispersão e superdispersão conjuntamente.

Para todos os modelos analisados, o efeito de tratamento e interação de tempo e tratamento não foram significativos ao nível de significância de 5%, isto é, o medicamento Progabide não contribuiu para a redução do número de crises epiléticas dos pacientes. O efeito longitudinal (tempo) foi somente significativo para os modelos PN e DPN. Como verificado na análise descritiva, a variância global é maior que a média global, havendo indício de superdispersão, o que pode ser validado por meio de um teste de hipótese para o parâmetro de dispersão, uma vez que os modelos PN e DPN são modelos

encaixados. Dessa forma, observa-se que a hipótese nula foi rejeitada ($H_0 : \delta = 0$), ou seja, a inclusão do parâmetro de dispersão contribuiu para explicar o comportamento da variável resposta, confirmando a existência da superdispersão. É importante notar que o modelo DPND apresentou o menor valor do logaritmo natural da função de verossimilhança, assim como o menor AIC. Como os modelos DPN e DPND são modelos encaixados, é possível realizar um teste de hipótese para verificar se a inclusão da variável *dummy* possibilita a melhora do ajuste do modelo. Verificou-se então que a hipótese nula foi rejeitada ($H_0 : \delta_1 + \delta_2 = 0$), indicando que a inclusão da variável *dummy* no parâmetro de dispersão contribuiu de forma significativa para a melhora do ajuste do modelo.

Tabela 3.4: Medidas de ajuste, comparação entre os preditores lineares e teste de dispersão para os modelos Poisson normal (PN), *double* Poisson normal (DPN) e *double* Poisson normal *dummy* (DPND), ajustados aos dados do ensaio clínico realizado por Leppik et al. (1985).

PN	np	ln	AIC	dif np	2(dif ln)	$P(> \chi^2)^a$			
Preditor 1	2	-592,9266	1189,8532						
Preditor 2	3	-588,3383	1182,6766	1	9,1766	0,0025			
Preditor 3	4	-588,0971	1184,1942	1	0,4824	0,4873			
Preditor 4	5	-587,7522	1185,5044	1	0,6898	0,4062			
DPN	np	ln	AIC	dif np	2(dif ln)	$P(> \chi^2)^a$	$\hat{\delta}$	$P(> \chi^2)^b$	
Preditor 1	3	-571,6507	1149,3014				0,5201	<0,0001	
Preditor 2	4	-569,0292	1146,0584	1	5,2430	0,0220	0,5003	<0,0001	
Preditor 3	5	-568,9637	1147,9274	1	0,1310	0,7174	0,5065	<0,0001	
Preditor 4	6	-568,7225	1149,4450	1	0,4824	0,4873	0,4972	<0,0001	
DPND	np	ln	AIC	dif np	2(dif ln)	$P(> \chi^2)^a$	$\hat{\delta}_1$	$\hat{\delta}_2$	$P(> \chi^2)^c$
Preditor 1	4	-513,9259	1035,8518				-1,0346	1,1025	<0,0001
Preditor 2	5	-512,2824	1034,5648	1	3,2870	0,0698	-1,0420	1,0805	<0,0001
Preditor 3	6	-512,0497	1036,0994	1	0,4654	0,4951	-1,0433	1,0788	<0,0001
Preditor 4	7	-511,5918	1037,1836	1	0,9158	0,3386	-1,0499	1,0776	<0,0001

Notas: número de parâmetros (np); logaritmo natural da função de verossimilhança (ln); diferença do número de parâmetros (dif np); diferença do logaritmo natural da função de verossimilhança (dif ln); p -valores associados ao teste de hipótese para o efeito das covariáveis (^a); p -valores associados ao teste de hipótese para o parâmetro de dispersão (^b); p -valores associados ao teste de hipótese para o parâmetro de dispersão considerando a variável *dummy* (^c).

As estimativas e os erros padrões para os modelos PN, DPN e DPND são apresentadas na Tabela 3.5, considerando o preditor 4 apenas para efeito de comparação das estimativas, uma vez que os fatores envolvidos na análise clínica não foram significativos. Ao comparar os resultados do modelo PN com o modelo DPN, verifica-se uma similaridade nas estimativas dos parâmetros, de efeitos fixos e a estimativa do parâmetro σ_1^2 apresentou um aumento de 19%. Devido à não equidispersão presente nos dados, as diferenças nos erros padrões são mais acentuadas, com o modelo DPN apresentando um aumento médio de 22%. Para os modelos PN e DPND o destaque fica para o parâmetro β_1 com uma redução de aproximadamente 152%. Essa diferença notória se justifica pelo fato do modelo PN apresentar o efeito longitudinal como uma variável significativa para o modelo (Tabela 3.4). Quando comparamos os modelos DPN e DPND, o mesmo ocorre para o parâmetro β_1 . Já para os erros padrões, devido à inclusão da variável *dummy*, as diferenças são mais acentuadas, com o modelo DPND apresentando uma redução média de 43%, exceto para o parâmetro de dispersão.

Tabela 3.5: Estimativa dos parâmetros e erros padrões para os modelos Poisson normal (PN), *double* Poisson normal (DPN) e *double* Poisson normal *dummy* (DPND), ajustados aos dados do ensaio clínico realizado por Leppik et al. (1985).

Poisson normal (PN)							
Parâmetro	Estimativa	Erro padrão	P (ltl)				
β_0	1,9320	0,0776	<0,0001				
β_1	-0,0428	0,0288	0,1400				
β_2	-0,1765	0,1084	0,1050				
β_3	-0,0315	0,0406	0,4390				
$\sigma_{\beta_0}^2$	0,7702	—	—				
<i>double</i> Poisson normal (DPN)				Taxa percentual entre PN e DPN			
Parâmetro	Estimativa	Erro padrão	P (ltl)	Estimativa	Erro padrão		
β_0	1,9159	0,0998	<0,0001	-0,84%	22,26%		
β_1	-0,0429	0,0371	0,2500	0,09%	22,29%		
β_2	-0,1741	0,1393	0,2130	-1,34%	22,18%		
β_3	-0,0318	0,0522	0,5430	1,04%	22,21%		
$\sigma_{\beta_0}^2$	0,9519	—	—	19,08%	—		
δ	0,4972	0,0927	<0,0001	—	—		
<i>double</i> Poisson normal <i>dummy</i> (DPND)				Taxa percentual entre PN e DPND		Taxa percentual entre DPN e DPND	
Parâmetro	Estimativa	Erro padrão	P (ltl)	Estimativa	Erro padrão	Estimativa	Erro padrão
β_0	1,8425	0,0689	<0,0001	-4,86%	-12,63%	-3,98%	-44,89%
β_1	-0,0170	0,0253	0,5033	-152,12%	-13,90%	-152,36%	-46,56%
β_2	-0,1984	0,0997	0,0482	11,05%	-8,68%	12,23%	-39,65%
β_3	-0,0350	0,0370	0,3452	10,09%	-9,71%	9,14%	-41,02%
$\sigma_{\beta_0}^2$	0,9410	—	—	18,15%	—	-1,16%	—
δ_{i1}	-1,0499	0,1327	<0,0001	—	—	—	—
δ_{i2}	1,0776	0,1387	<0,0001	—	—	53,86%	33,17% —

Outra diferença entre resultados dos modelos encontra-se nos parâmetros de dispersão. O modelo DPN possui apenas um parâmetro de dispersão que controla e estima a superdispersão presente nos dados, já o modelo DPND, por sua vez, apresenta dois parâmetros de dispersão que controlam a subdispersão e a superdispersão conjuntamente. Neste sentido, quando se comparam os modelos DPN e DPND no que tange o parâmetro de dispersão para a superdispersão, observa-se que há um aumento de aproximadamente 54% na estimativa. Em outras palavras, como o modelo DPN possui uma menor flexibilidade em relação ao parâmetro de dispersão, pode eventualmente subestimar a variabilidade real do dados, uma vez que esta variável em estudo, particularmente, apresenta tanto sub quanto superdispersão. Em suma, o modelo DPND apresenta-se como uma ferramenta útil para dados que apresentam diferentes graus de dispersão entre os perfis. Quando ignora-se essa mistura de dispersão, o modelo DPN é capaz de trabalhar somente com o fenômeno da superdispersão, podendo causar um comprometimento das inferências.

Para detectar a presença de observações discrepantes, bem como a violação da suposição de independência dos erros e homogeneidade de variância, utilizou-se os resíduos quantílicos, como descrito na Sessão 3.3. As análises de resíduos para os modelos PN, DPN e DPND são apresentadas na Figura 3.6. Os pontos dos modelos PN (Figura 3.6a) e DPN (Figura 3.6c) encontram-se majoritariamente fora do envelope simulado, indicando que estes modelos não se ajustaram de forma satisfatória em função da não equidispersão presente nos dados. Por meio das Figuras 3.6b e 3.6d é possível identificar a presença de algumas observações discrepantes. Para o modelo DPND verifica-se que o modelo apresentou um ajuste satisfatório, uma vez que praticamente todos os pontos estão dentro do envelope simulado (Figuras 3.6e).

(a)

(b)

(c)

(d)

(e)

(f)

Figura 3.6: Gráfico meio normal de probabilidade com envelope simulado (nível de confiança = 95%) e gráfico de resíduos versus ordem das observações para os modelos Poisson normal (a, b); *double* Poisson normal (c, d); *double* Poisson normal *dummy* (e, f) referente ao ensaio clínico realizado por Leppik et al. (1985).

As probabilidades estimadas para o número médio de crises epiléticas são apresentadas na Figura 3.7. Em comparação com o modelo DPND, é possível verificar que os modelos PN e DPN subestimaram a probabilidade de ocorrência de crises epiléticas em torno da média para todos os pacientes que apresentaram subdispersão (a) e superestimou a probabilidade de ocorrência para os pacientes com superdispersão (b). Este resultado pode estar associado ao fato dos modelos PN e DPN não acomodarem adequadamente a não equidispersão presente nos dados. Também é importante notar que a probabilidade de ocorrência ao longo do tempo é praticamente constante, o que evidencia que não há efeito de tratamento ao longo do tempo.

Figura 3.7: Gráficos de probabilidades estimadas considerando o número médio de crises epiléticas por paciente durante os quatro períodos amostrados para os modelos Poisson normal (PN), *double* Poisson normal (DPN) e *double* Poisson normal *dummy* (DPND). As letras a e b nos gráficos indicam pacientes que apresentaram subdispersão e superdispersão, respectivamente, no número médio de crises epiléticas ao longo do ensaio clínico realizado por Leppik et al. (1985).

Os modelos PN, DPN e DPND apresentaram médias ajustadas próximas (Figura 3.8) e, de modo geral, para todos os modelos verifica-se que há uma redução ao longo do tempo no número médio de crises epiléticas. No entanto, essa redução na maioria dos casos é inferior a uma unidade. A principal diferença entre os modelos encontra-se no parâmetro de dispersão que influencia nas medidas de probabilidade do modelo. Dessa forma, é importante destacar a relevância de um modelo com ajuste satisfatório, pois as medidas de probabilidades dependem deste parâmetro. Medidas de probabilidades adequadas podem ajudar os pesquisadores a estimar a probabilidade de ocorrência de crises epiléticas por paciente e medicamento utilizado, o que contribui para a tomada de decisão, podendo ser igualmente interessantes para outros ensaios clínicos.

Figura 3.8: Gráfico de comparação dos valores médios ajustados de crises epiléticas por paciente durante os quatro períodos amostrados para os modelos Poisson normal (PN), *double* Poisson normal (DPN) e *double* Poisson normal dummy (DPND). As letras a e b nos gráficos indicam pacientes que apresentaram subdispersão e superdispersão, respectivamente, no número médio de crises epiléticas ao longo do ensaio clínico realizado por Leppik et al. (1985).

3.5.4 Considerações finais

A metodologia proposta e utilizada neste trabalho mostrou ser eficiente em analisar um conjunto de dados que apresenta subdispersão e superdispersão concomitantemente, e esta situação particular pode ocorrer nas mais variadas áreas do conhecimento. Aplicar um método que leva em consideração a análise conjunta destes fenômenos propicia melhores resultados quando comparados a outros métodos convencionais. Diferente da metodologia tradicional, que comumente foca na identificação de observações influentes e sua respectiva interferência no ajuste do modelo (e por vezes são removidos da análise), esta abordagem permite uma nova perspectiva ao ajustar a variabilidade dos dados particionadamente. O modelo *double* Poisson normal *dummy* mostrou maior versatilidade ao modelar a sub- e superdispersão, representando uma vantagem em relação aos modelos Poisson normal e *double* Poisson normal. Com base na aplicação do ensaio clínico a análise de resíduos reforça essa conclusão, em virtude do ajuste satisfatório revelado por meio do gráfico meio normal de probabilidade. Dessa forma, os resultados indicam que o método empregado tem um bom desempenho e pode ser uma alternativa potencial aos modelos convencionais disponíveis na literatura.

Referências

- Atkinson, A. C. (1985). Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis. Technical report.
- Booth, J. G., Casella, G., Friedl, H., e Hobert, J. P. (2003). Negative binomial loglinear mixed models. *Statistical Modelling*, 3(3):179–191.
- Breslow, N. E. e Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.
- Brian, S. E. e Torsten, H. (2017). *HSAUR: A Handbook of Statistical Analyses Using R (1st Edition)*. R package version 1.3-9.
- Cameron, A. C. e Trivedi, P. (2013). *Regression analysis of count data*, volume 53. Cambridge University Press.
- Caruana, E. J., Roman, M., Hernández-Sánchez, J., e Solli, P. (2015). Longitudinal studies. *Journal of thoracic disease*, 7(11):E537.
- Demétrio, C. G. B., Hinde, J., e Moral, R. A. (2014). Models for overdispersed data in entomology. In *Ecological modelling applied to entomology*, pages 219–259. Springer.
- Diggle, P. J., Heagerty, P., Liang, K. Y., e Zeger, S. L. (2002). *Analysis of longitudinal data*. Oxford University Press.
- Dunn, P. K. e Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244.
- Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81(395):709–721.

- Grunwald, G. K., Bruce, S. L., Jiang, L., Strand, M., e Rabinovitch, N. (2011). A statistical model for under-or overdispersed clustered and longitudinal count data. *Biometrical Journal*, 53(4):578–594.
- Khan, M. N., Jowaheer, V., Sunecher, Y., Bourguignon, M., et al. (2018). Modeling longitudinal inma (1) with com–poisson innovation under non-stationarity: application to medical data. *Computational and Applied Mathematics*, 37(4).
- Khan, N. M. e Jowaheer, V. (2013b). Comparing joint ggl estimation and gmm adaptive estimation in com-poisson longitudinal regression model. *Communications in Statistics-Simulation and Computation*, 42(4):755–770.
- Lara, I. A. R. d., Hinde, J., e Taconeli, C. A. (2019). Global and local tests to assess stationarity of markov transition models. *Communications in Statistics-Simulation and Computation*, 48(4):1019–1039.
- Leppik, I., Dreifuss, F., Bowman, T., Santilli, N., Jacobs, M., Crosby, C., Cloyd, J., Stockman, J., Graves, N., Sutula, T., et al. (1985). A double-blind crossover evaluation of progabide in partial seizures. *Neurology*, 35(4).
- Lesaffre, E. e Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):325–335.
- Liu, Q. e Pierce, D. A. (1994). A note on gaussÜhermite quadrature. *Biometrika*, 81(3):624–629.
- Luyts, M., Molenberghs, G., Verbeke, G., Matthijs, K., Ribeiro Jr, E. E., Demétrio, C. G., e Hinde, J. (2018). A weibull-count approach for handling under-and overdispersed longitudinal/clustered data structures. *Statistical Modelling*, page 1471082X18789992.
- McCullagh, P. e Nelder, J. A. (1989). *Generalized linear models (Monographs on statistics and applied probability 37)*.
- Molenberghs, G. e Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag New York.
- Molenberghs, G., Verbeke, G., e Demétrio, C. G. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime data analysis*, 13(4):513–531.
- Moral, R. A., Hinde, J., e Demétrio, C. G. (2017). Half-normal plots and overdispersed models in r: The hnp package. *J. Stat. Softw*, 81(10).
- Niu, X. e Ryan Cho, H. (2018). Efficient regression modeling for correlated and overdispersed count data. *Communications in Statistics-Theory and Methods*, pages 1–14.
- Nobre, S. J. e Motta Singer, J. (2007). Residual analysis for linear mixed models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 49(6):863–875.
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

- Rigby, R. A. e Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554.
- Smyth, G., Hu, Y., Dunn, P., Phipson, B., e Chen, Y. (2013). statmod: Statistical modeling. *R package version*, 1(4):17.
- Stasinopoulos, M., Rigby, B., Akantziliotou, C., Heller, G., Ospina, R., e Stasinopoulos, M. M. (2019). Package *gamlss.dist*.
- Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., e De Bastiani, F. (2017). *Flexible regression and smoothing: using GAMLSS in R*. Chapman and Hall/CRC.
- Sung, Y. e Lee, K. (2018). Negative binomial loglinear mixed models with general random effects covariance matrix. *Communications for Statistical Applications and Methods*, 25(1):61–70.
- Tapia, A., Giampaoli, V., Diaz, M. P., e Leiva, V. (2019). Sensitivity analysis of longitudinal count responses: a local influence approach and application to medical data. *Journal of Applied Statistics*, 46(6):1021–1042.
- Thall, P. F. e Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, pages 657–671.
- Thomas, R. e Have, T. (1996). A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics*, pages 473–491.
- Toscas, P. J. e Faddy, M. J. (2003). Likelihood-based analysis of longitudinal count data using a generalized poisson model. *Statistical Modelling*, 3(2):99–108.
- Winkelmann, R. (1995). Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics*, 13(4):467–474.
- Winkelmann, R. (2008). *Econometric analysis of count data*. Springer.
- Zeger, S. L. e Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130.
- Zhang, X., Pei, Y.-F., Zhang, L., Guo, B., Pendegraft, A., Zhuang, W., e Yi, N. (2018). Negative binomial mixed models for analyzing longitudinal microbiome data. *Frontiers in microbiology*, 9:1683.
- Zhu, F. (2012). Modeling overdispersed or underdispersed count data with generalized poisson integer-valued garch models. *Journal of Mathematical Analysis and Applications*, 389(1):58–71.

Apêndice

```
# pacotes
require(gamlss)
require(Rfast)
```

```
#####
```

```

#simulação
#####
sim.double.poisson <- function(
f.fixo, f.aleat, beta.fixo, sigma_beta0, subject,delta_k){
X <- model.matrix(f.fixo, subject)
Z <- model.matrix(f.aleat, subject)
V <- model.matrix(~-1+var)
n.bloco <- ncol(Z)
n.rep <- nrow(Z)/n.bloco
bi <- rNO(n.bloco,0,sigma_beta0)
XZ <- cbind(X,Z)
beta <- c(beta.fixo,bi)
preditor <- XZ%%beta
lambda <- exp(preditor)
dummy <- V%%delta_k
y <- rDP0(n=length(lambda), mu=lambda,dummy)
return(cbind(y=y))
}

#####
# definindo a estrutura
i<-40 # numero de individuos
t<-5 # numero de repeticoes no tempo
ID<-as.factor(rep(1:i,each=t))
periodo<-rep(1:t,i)
var<-as.factor(rep(1:2,each=i*t/2))
N<-1000 # numero de amostras geradas
#####
#####
# gerando as covariáveis
x1 <- seq(1:(i*t))
rep(1:i,each=t)
covariavel <- (matrix(0, ncol = 1, nrow = i*t))
for(a in 1:length(x1)){
cova<-if(x1[a] <= (i*t)/5){
print(-0.5)
} else
if(x1[a] >= (i*t/5)+1 & x1[a] <= (4*i*t)/5 ) {
print(0)
} else
if(x1[a] >= (4*i*t/5)+1 & x1[a] <= (i*t) ){
print(0.5)
}
}

```

```

covariavel[a] <- as.matrix(cova)
}
#####
resp <- (matrix(0, ncol = N, nrow = i*t))
for(j in 1:N){
simu <- sim.double.poisson(f.fixo=~1+(covariavel),
f.aleat=~1 + ID,
beta.fixo = c(b0=1,b1=1),
sigma_beta0=0.9,
delta_k<-c(delta_1=0.3,delta_2=3),
subject=ID)
resp[,j] <- as.matrix(simu)
}
#####
dados<-data.frame(resp,ID,covariavel,periodo)
#####
# media e variancia
media <- (matrix(0, ncol = N, nrow =i*t ))
for(j in 1:N){
medi <- rep(tapply(dados[,j]+0.0000001,dados$ID,mean),each=t)
media[,j] <- cbind(medi)
}

variancia <- (matrix(0, ncol = N, nrow =i*t ))
for(j in 1:N){
var <- rep(tapply(dados[,j]+0.000001,dados$ID,var),each=t)
variancia[,j] <- cbind(var)
}

media_var<-list(media,variancia)
for (j in 1:N) {
a<-cbind(media_var[[1]][,j],media_var[[2]][,j])
#print(a)
}
#####
dados_simu<-(lapply(1:2, function(x) matrix(0, nrow=i*t, ncol=6)))
options(max.print=999999)
for(j in 1:N){
simu<-cbind(dados[,j],covariavel,
subi=ifelse(variancia[,j] < media[,j],1,0),
super=ifelse(variancia[,j] > media[,j],1,0),
ID)

```

```

dados_simu[[j]]<-data.frame(simu)
}

coef <- (matrix(0, ncol = 5, nrow =N ))
for(j in 1:N){
fit_dpnd<-gamlss(dados_simu[[j]][,1]~dados_simu[[j]][,2]+
random(as.factor(dados_simu[[j]][,5])),
sigma.formula = ~(dados_simu[[j]][,3]+
dados_simu[[j]][,4]-1),
data= dados_simu[[j]],
family = DPO,
trace=FALSE,
method = mixed(60, 60),
method.pb="ML" # método de estimação
)
coef[j,]<-cbind(fit_dpnd$mu.coefficients[1],
fit_dpnd$mu.coefficients[2],
exp(fit_dpnd$sigma.coefficients[1]),
exp(fit_dpnd$sigma.coefficients[2]),
as.numeric(getSmo(fit_dpnd)[4]))
}

colMeans(coef);sqrt(colVars(coef))
####
#análise conjunto de dados
####
rm(list=ls(all=TRUE))
data("epilepsy", package = "HSAUR")

var= rep(tapply(epilepsy$seizure.rate+0.00001,epilepsy$subject,var),each=4)
media=rep(tapply(epilepsy$seizure.rate+0.00001,epilepsy$subject,mean),each=4)
dados<-data.frame(epilepsy,media,var,
seq=1:236,
subi=ifelse(var < media,1,0),
equi=ifelse(var == media,1,0),
super=ifelse(var > media,1,0))

fit_dpnd<-gamlss(seizure.rate~as.numeric(period)*treatment+
random(subject),
sigma.formula = ~(subi+super-1),
data= dados,
family = DPO,

```



```

trace=FALSE, # nao mostra as interacoes
method = mixed(50, 50)
)
summary(fit_dpnd)
(sigma2_b0<-getSmo(fit_dpnd)[4])

# segunda opção dummy

fit<-with(epilepsy,
by(epilepsy, epilepsy[, "subject"],
function(x) summary(glm(seizure.rate ~ 1, family = quasipoisson, data = x))))
fit_dispersion<-with(epilepsy,
by(epilepsy, subject,
function(x) summary(glm(seizure.rate ~ 1, family = quasipoisson, data = x))$dispersion))

phi_hat<-rep(round(as.vector(fit_dispersion[1:59]),4),each=4)

dados2<-data.frame(epilepsy, phi_hat,
seq=1:236,
subi=ifelse(phi_hat < 1,1,0),
equi=ifelse(phi_hat == 1,1,0),
super=ifelse(phi_hat > 1,1,0))

fit_dpnd2<-gamlss(seizure.rate~as.numeric(period)*treatment+
random(subject),
sigma.formula = ~(subi+super-1),
data= dados,
family = DP0,
trace=FALSE, # nao mostra as interacoes
method = mixed(50, 50)
)
summary(fit_dpnd2)
(sigma2_b0<-getSmo(fit_dpnd2)[4])

```