Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Modelagem estatística e desenvolvimento de ferramentas computacionais aplicados à produção *in vitro* de embriões bovinos e hormesis em plantas

Deoclecio Jardim Amorim

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

Piracicaba
2023

Deoclecio Jardim Amorim
Bacharel em Agronomia

Modelagem estatística e desenvolvimento de ferramentas computacionais aplicados à produção *in vitro* de embriões bovinos e hormesis em plantas

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientadora:
Profa. Dra. **CLARICE GARCIA BORGES DEMÉTRIO**

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

Piracicaba
2023

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

Amorim, Deoclecio Jardim

Modelagem estatística e desenvolvimento de ferramentas computacionais aplicados à produção *in vitro* de embriões bovinos e hormesis em plantas/ Deoclecio Jardim Amorim. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2023 .

132 p.

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

1. Modelagem estatística 2. Modelos lineares generalizados mistos 3. Superdispersão 4. Produção de embriões bovinos 5. Hormesis 6. Modelos não lineares . I. Título.

DEDICATÓRIA

A Deus, por cumprir suas promessas em minha vida.
À minha querida esposa, que sempre me apoiou e lutou comigo.
Aos meus amados pais e irmãos.

AGRADECIMENTOS

Agradeço primeiramente ao Senhor nosso Deus todo Poderoso.

À coordenação de aperfeiçoamento de pessoal de nível superior (CAPES), pelo auxílio financeiro concedido nessa trajetória.

A minha orientadora, Profa. Dra. Clarice Garcia Borges Demétrio que sempre acreditou em nosso trabalho, por ser paciente e passar um pouco do seu conhecimento. Sou muito grato por ter vivenciado esses momentos com a senhora.

A Mestre Daniela Garcia Borges Demétrio, pela paciência e por toda contribuição no desenvolvimento deste trabalho.

Ao Prof. Dr. Afrânio Márcio Corrêa Vieira, pela solicitude e por toda contribuição no desenvolvimento deste trabalho.

A todos os professores do curso de Pós-graduação em Estatística e Experimentação Agronômica, pela oportunidade concedida, ensinamentos e apoio que contribuíram para minha formação.

Às secretárias do Departamento de Ciências Exatas, Solange de Assis Paes Sabadin e Luciane Brajão, e aos técnicos de informática, Eduardo Bonilha e Jorge Alexandre Wiendl, pela ajuda e boa vontade.

A minha amada esposa Jania Claudia Camilo dos Santos por todo companheirismo, amor e carinho.

Aos meus amados pais João dos Santos Amorim e Maria Lima Jardim Amorim por participarem da minha vida fornecendo o constante apoio na minha caminhada.

Aos meus irmãos Deuciane Jardim Amorim da Silva, Deucleiton Jardim Amorim, Marilane Jardim Amorim, Marisa Jardim Amorim, Antono José Jardim Amorim e Maria Regina Jardim Amorim.

Aos amigos Vitor dos Santos, Camila Fernanda Boaro Spernega e Lucas dos Santos pela grande ajuda fornecida, descontração e amizade.

A todos os meus amigos do Programa de Pós-graduação e Experimentação Agronômica, pelos momentos de estudos, café, atenção e amizade.

Finalmente, a todos que me ajudaram de forma direta e indireta para o desenvolvimento deste trabalho.

EPÍGRAFE

"Para que, como está escrito: Aquele que se gloria, glorie-se no Senhor".

1 Coríntios 1: 31

SUMÁRIO

RESUMO

Modelagem estatística e desenvolvimento de ferramentas computacionais aplicados à produção *in vitro* de embriões bovinos e hormesis em plantas

Dados provenientes de estudos em ciências agrárias podem apresentar características diferentes, com destaque para estudos de produção *in vitro* de embriões (PIVE) e hormesis em plantas. Na PIVE, os dados, geralmente, não apresentam distribuição normal e são medidos em contextos longitudinais ou hierárquicos. Exemplos comuns incluem dados na forma de contagens, proporções e binários. Por outro lado, na maioria dos estudos de hormesis, os dados são correlacionados, presentes em características morfológicas, fisiológicas e bioquímicas. Nesse sentido, o objetivo deste trabalho é desenvolver modelos e ferramentas computacionais que auxiliem na análise desses tipos de dados, tornando-a mais acessível a pesquisadores de áreas aplicadas. Em relação aos dados da PIVE, inicialmente, abordaram-se os aspectos relacionados a superdispersão e correlação em dados longitudinais, testes de hipóteses sobre os componentes de variância e seleção de modelos com foco em dados na forma de contagens e na proposição de um tutorial utilizando o software `R`. Os modelos lineares generalizados mistos (MLGMs) e os modelos combinados (MC) ajustaram-se satisfatoriamente, captando a variabilidade extra e a correlação entre as medidas longitudinais. O MC tem como característica o uso de dois conjuntos de efeitos aleatórios para capturar a superdispersão e a correlação, sendo mais flexíveis que os tradicionais MLGMs. Apesar da flexibilidade de modelagem oferecida, os MCs ainda não possuem uma ferramenta computacional padronizada no software de código-aberto `R`. Em função disso, foi desenvolvido o pacote **combTMB** para o software `R`. As funcionalidades do pacote são ilustradas com duas aplicações em dados na forma de contagens e de proporções. Finalmente, foi proposta uma modelagem não linear multivariada para estudar hormesis em plantas, sendo que as informações concernentes às correlações das variáveis respostas são consideradas com uma matriz de variâncias e covariâncias obtida, a partir dos resíduos univariados. Uma vantagem dessa estratégia de modelagem é a obtenção de estimativas mais precisas dos parâmetros e uma maior compreensão das relações bifásicas de dose-resposta, pois considera na estrutura de modelagem as inter-relações existentes entre as diversas características mensuradas no sistema vegetal.

Palavras-chave: Modelagem estatística, Modelos lineares generalizados mistos, Superdispersão, Produção de embriões bovinos, Hormesis, Modelos não lineares.

ABSTRACT

Statistical modeling and development of computational tools applied to *in vitro* production of bovine embryos and hormesis in plants

Data from studies in agricultural sciences may present different features, especially in studies of *in vitro* embryo production (IVP) and hormesis in plants. In IVP, data are usually non-Gaussian and are measured in longitudinal or hierarchical contexts. Common examples include count, proportion, and binary data. On the other hand, in most hormesis studies, data are correlated and present in morphological, physiological, and biochemical characteristics. In this sense, the objective of this work is to develop models and computational tools that assist in the analysis of these types of data, making it more accessible to researchers in applied fields. Regarding IVP data, initially, aspects related to overdispersion and correlation in longitudinal data, hypothesis tests for variance components, and model selection focused on count data were addressed, proposing a tutorial using the software `R`. Generalized linear mixed models (GLMMs) and combined models (CM) were satisfactorily fitted, capturing the extra variability and correlation between longitudinal measures. The CM is characterized by the use of two sets of random effects to capture overdispersion and correlation, being more flexible than traditional GLMMs. Despite the modeling flexibility offered, CMs still do not have a standardized computational tool in the open-source software `R`. For this purpose, it was developped the package **combTMB** for `R` software. The package's functionalities are illustrated with two applications on count and proportion data. Finally, a multivariate nonlinear modeling has been proposed to study hormesis in plants, where information concerning to the correlations of response variables is considered with a variance-covariance matrix obtained from univariate residuals. An advantage of this modeling strategy is obtaining more precise estimates of parameters and a better understanding of dose-response biphasic relationships, considering the interrelationships between the various characteristics measured in the plant system.

Keywords: Statistical modeling, Generalized linear mixed model, Overdispersion, Bovine embryo production, Hormesis, Nonlinear models.

1 INTRODUÇÃO

As técnicas de modelagem estatística são amplamente empregadas em muitas áreas de estudo, incluindo as ciências agrárias, que englobam subáreas como agronomia, zootecnia e medicina veterinária. Em cada uma dessas subáreas, os estudos realizados podem apresentar características distintas, tais como dados discretos e contínuos, experimentos univariados e/ou multivariados, dados longitudinais e/ou hierárquicos, dentre outros tipos de dados e estruturas. Essa variedade de características, aliada ao grande volume de dados gerados anualmente, torna essa área muito interessante para a proposição e apresentação de diferentes abordagens estatísticas com o intuito de capturar as particularidades presentes nos dados.

Neste trabalho, as técnicas de modelagem estatística foram empregadas em dois problemas particulares na área das ciências agrárias. O primeiro refere-se à produção *in vitro* de embriões (PIVE), uma biotecnologia poderosa para o melhoramento genético do gado leiteiro, que viabiliza a multiplicação de fêmeas de alto mérito genético (Demétrio et al., 2020; Seneda et al., 2020). Em 2021, mesmo diante da pandemia de Covid-19, que ainda era um grande problema de saúde global, a indústria de transferência de embriões permaneceu em alta em todo o mundo, registrando um aumento no número total de embriões em quase todos os países para todas as espécies representativas (bovinos, equinos, ovinos e caprinos). Em bovinos, a produção total de embriões (tanto *in vivo* quanto *in vitro*) aumentou em cerca de 7,0% em relação a 2020, com mais de 1,5 milhão de embriões registrados. Em todo o mundo, os embriões produzidos *in vitro* representaram 79,7% de todos os embriões bovinos transferíveis em 2021 (Viana, 2022).

Apesar do grande sucesso da PIVE proporcionado pelo avanço das técnicas laboratoriais e genéticas, os programas de coleta de óvulos (OPU) ainda enfrentam alguns desafios que comprometem a eficácia plena da técnica. Por exemplo, a PIVE pode ser afetada pelo estresse térmico (Souza-Cácares et al., 2019), pelo status reprodutivo das doadoras (Wang et al., 2020), pelo intervalo entre as OPU (Pontes et al., 2011), pelos protocolos hormonais (Demétrio et al., 2020) e outros fatores. Além disso, os dados obtidos na PIVE podem apresentar diferentes características, o que torna a análise estatística e a tomada de decisão em programas em larga escala mais desafiadoras.

Em geral, os dados obtidos na PIVE são dados não normais mensurados em contextos longitudinais ou hierárquicos, em que os animais ou "clusters" são medidos repetidamente. Portanto, são dados estruturados com diferentes graus de complexidade. Exemplos comuns incluem dados na forma de contagens, como o números totais de oócitos (Garcia et al., 2020), dados binários, como a ocorrência ou não de prenhez (Demétrio et al., 2020), e dados na forma de proporções, como a taxa de prenhez (Chebel et al., 2008). Esses tipos de dados podem ser analisados usando os modelos lineares generalizados (MLGs) (Nelder e Wedderburn, 1972), que oferecem uma variedade de opções, com distribuições

que pertencem à família exponencial.

Ao utilizar a classe dos MLGs, por exemplo, o modelo Poisson para dados de contagens e o modelo binomial para dados de proporções, é necessário ter cuidado com a superdispersão, fenômeno comum em dados biológicos em que a variabilidade dos dados é maior do que a esperada pelo modelo probabilístico adotado (Hinde e Demétrio, 1998; Demétrio et al., 2014). Ajustar modelos inadequados a dados superdispersos pode resultar em erros-padrão subestimados e interpretação incorreta da significância dos parâmetros. Hinde e Demétrio (1998) fornecem uma visão geral das abordagens para lidar com esse fenômeno, com ênfase em dados na forma de contagens e proporções.

Outra alternativa de modelagem para os dados da PIVE é a utilização de modelos lineares generalizados mistos (MLGM), desenvolvidos por Breslow e Clayton (1993). Essa classe de modelo estende os MLGs para o contexto das medidas repetidas usando efeitos aleatórios normalmente distribuídos, permitindo a modelagem das correlações entre as unidades amostrais. Além disso, os MLGMs podem capturar a superdispersão até certo ponto (Molenberghs et al., 2007; Iddi e Molenberghs, 2012), tornando-se uma ferramenta popular para dados discretos. Vale ressaltar que grande parte da popularidade dos MLGMs é atribuída à disponibilidade de ferramentas computacionais, como o pacote `R` **lme4** (Bates et al., 2015).

As estratégias de modelagem mencionadas acomodam a superdispersão ou a correlação. No entanto, existem situações em que ambos os fenômenos ocorrem simultaneamente (Molenberghs et al., 2012). Para tais casos, os modelos combinados (MC) desenvolvidos por Molenberghs et al. (2007) e Molenberghs et al. (2010) podem ser uma alternativa viável. No entanto, a necessidade de um software padrão, como o pacote **lme4**, torna seu uso difícil para a maioria dos pesquisadores de áreas aplicadas.

O segundo problema abordado neste trabalho está relacionado à exposição de plantas a baixas doses de produtos químicos ou outros estressores que podem estimular o crescimento em várias espécies vegetais. Por exemplo, a aplicação de baixas doses do herbicida glifosato (N-(fosfonometil)glicina) pode estimular o crescimento e/ou a produção em várias culturas agrícolas, incluindo a cana-de-açúcar (*Saccharum officinarum* L.) (Pincelli-Souza et al., 2020), o feijão (*Phaseolus vulgaris* L.) (Bortolheiro e Silva, 2021) e o cártamo (*Carthamus tinctorius* L.) (Santos et al., 2021). Na literatura, esse fenômeno estimulador é conhecido como hormesis, sendo um tipo especial de relação dose-resposta caracterizado por uma estimulação em baixas doses e uma inibição em altas doses, com ampla ocorrência em todos os sistemas biológicos (Calabrese e Blain, 2009).

A hormesis em plantas foi relatada pela primeira vez em 1907 por Jensen (1907), e desde então, os relatos de sua ocorrência têm se acumulado, especialmente aqueles causados por herbicidas (Belz e Duke, 2014). Como resultado, o fenômeno da hormesis tem sido reconhecido como relevante em diversas áreas de estudo (Cedergreen et al., 2005; Calabrese e Blain, 2009; Belz e Duke, 2022). Todavia, reconhecer os casos em que a

hormesis realmente ocorre e estabelecer adequadamente o seu significado ainda é um dos principais desafios enfrentados pelos pesquisadores.

De acordo com Belz e Duke (2022), a forma mais eficiente e confiável para testar a ocorrência da resposta hormética é a geração e modelagem estatística de uma curva completa de dose-resposta. Para isso, a técnica de modelagem estatística mais utilizada é a de modelos de regressão não linear univariados, com destaque para os modelos desenvolvidos por Brain e Cousens (1989) e Cedergreen et al. (2005). Ambos os modelos são amplamente utilizados na área de biologia vegetal e têm contribuído significativamente para aumentar a credibilidade do fenômeno da hormesis nesta área de pesquisa (Belz e Piepho, 2012).

Desde a introdução do modelo proposto por Brain e Cousens (1989), o fenômeno da hormesis tem sido discutido apenas sob o contexto univariado. Entretanto, a maioria dos estudos conduzidos para investigar a hormesis em plantas mede um conjunto de características morfológicas, fisiológicas e bioquímicas, geralmente, correlacionadas (Belz, 2018; Pincelli-Souza et al., 2020; Santos et al., 2022). Dessa forma, ao avaliar o efeito de uma variável por vez, sem considerar a interdependência entre elas, pode resultar em estimativas dos parâmetros menos precisas e comprometer as conclusões. Além disso, não é possível inferir, a partir de uma perspectiva global ou multivariada, sobre a existência da hormesis, ou seja, quando todas as características são consideradas, uma vez que não é possível realizar um teste estatístico formal para confrontar as hipóteses de existência ou não existência da hormesis. Nesse contexto, uma abordagem não linear multivariada pode ser desenvolvida e usada para este tipo de pesquisa.

Diante do exposto, nota-se que há uma demanda crescente por novas teorias, modelos de análise de dados e ferramentas computacionais para o estudo de dados relacionados à produção *in vitro* de embriões bovinos e para a investigação da hormesis em plantas. Dessa forma, os objetivos deste trabalho são:

1. Revisar e aplicar os modelos lineares generalizados (MLGs) e suas extensões a problemas de produção de embriões, buscando solucionar problemas relacionados à superdispersão e estruturas de correlação.

2. Desenvolver uma ferramenta computacional padronizada no software de código-aberto `R` (R Core Team, 2022) que forneça a classe dos MCs no sistema `R` para a computação estatística.

3. Apresentar uma proposta para modelar a hormesis em plantas no contexto multivariado.

Este trabalho está estruturado da seguinte forma: Capítulo 2, apresenta uma revisão dos MLGs e extensões e discute-se como ajustar e selecionar o melhor modelo para dados na forma de contagens correlacionados e superdispersos na PIVE, detalhando

passo a passo usando o software `R` (R Core Team, 2022). Este trabalho é um tutorial para pesquisadores de áreas aplicadas, por exemplo, áreas de zootecnia e medicina veterinária. No Capítulo 3, foi desenvolvido o pacote `R` **combTMB**, inspirado nas limitações das ferramentas computacionais destacadas no Capítulo 2, principalmente, na ausência de um pacote `R` que incorpore a classe dos MLG, MLGM e MCs em uma única ferramenta computacional. Esse pacote possui como características mais atraentes a combinação de velocidade e uma interface flexível e familiar aos usuários de `glm()` e **lme4**. Seu uso é ilustrado com a análise de dados na forma de contagens e de proporções. No Capítulo 4, propõe-se a modelagem não linear multivariada da hormesis, cuja principal vantagem em relação às abordagens univariadas é o fato de considerar as inter-relações existentes entre as diversas características mensuradas no sistema vegetal, levando a estimativas mais precisas dos parâmetros. Finalmente, no Capítulo 5, apresentam-se as considerações finais e algumas sugestões para trabalhos futuros.

Referências

Bates, D., Mächler, M., Bolker, B., e Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1).

Belz, R. G. (2018). Herbicide hormesis can act as a driver of resistance evolution in weeds–psii-target site resistance in *Chenopodium album* L. as a case study. *Pest Management Science*, 74(12):2874–2883.

Belz, R. G. e Duke, S. O. (2014). Herbicides and plant hormesis. *Pest Management Science*, 70(5):698–707.

Belz, R. G. e Duke, S. O. (2022). Modelling biphasic hormetic dose responses to predict sub-noael effects using plant biology as an example. *Current Opinion in Toxicology*.

Belz, R. G. e Piepho, H.-P. (2012). Modeling effective dosages in hormetic dose-response studies. *PloS ONE*, 7(3):e33432.

Bortolheiro, F. P. A. P. e Silva, M. A. (2021). Low doses of glyphosate can affect the nutrient composition of common beans depending on the sowing season. *Science of the Total Environment*, 794:148733.

Brain, P. e Cousens, R. (1989). An equation to describe dose responses where there is stimulation of growth at low doses. *Weed Research*, 29(2):93–96.

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.

Calabrese, E. J. e Blain, R. B. (2009). Hormesis and plant biology. *Environmental Pollution*, 157(1):42–48.

Cedergreen, N., Ritz, C., e Streibig, J. C. (2005). Improved empirical models describing hormesis. *Environmental Toxicology and Chemistry: An International Journal*, 24(12):3166–3172.

Chebel, R. C., Demétrio, D. G. B., e Metzger, J. (2008). Factors affecting success of embryo collection and transfer in large dairy herds. *Theriogenology*, 69(1):98–106.

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.

Demétrio, D. G. B., Benedetti, E., Demétrio, C. G. B., Fonseca, J., Oliveira, M., Magalhaes, A., e Santos, R. M. d. (2020). How can we improve embryo production and pregnancy outcomes of holstein embryos produced in vitro? (12 years of practical results at a california dairy farm). *Animal Reproduction*, 17(3).

Garcia, S. M., Morotti, F., Cavalieri, F. L. B., Lunardelli, P. A., de Oliveira Santos, A., Membrive, C. M. B., Castilho, C., Puelker, R. Z., Silva, J. O. F., Zangirolamo, A. F., e Seneda, M. M. (2020). Synchronization of stage of follicle development before OPU improves embryo production in cows with large antral follicle counts. *Animal Reproduction Science*, 221:106601.

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

Iddi, S. e Molenberghs, G. (2012). A combined overdispersed and marginalized multilevel model. *Computational Statistics & Data Analysis*, 56(6):1944–1951.

Jensen, G. H. (1907). Toxic limits and stimulation effects of some salts and poisons on wheat. *Botanical Gazette*, 43(1):11 – 44.

Molenberghs, G., Verbeke, G., e Demétrio, C. G. B. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, 13(4):513–531.

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

Molenberghs, G., Verbeke, G., Iddi, S., e Demétrio, C. G. B. (2012). A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. *Journal of Multivariate Analysis*, 111:94–109.

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

Pincelli-Souza, R. P., Bortolheiro, F. P. A. P., Carbonari, C. A., Velini, E. D., e Silva, M. A. (2020). Hormetic effect of glyphosate persists during the entire growth period and increases sugarcane yield. *Pest Management Science*, 76(7):2388–2394.

Pontes, J., Sterza, F. M., Basso, A., Ferreira, C., Sanches, B., Rubin, K., e Seneda, M. (2011). Ovum pick up, in vitro embryo production, and pregnancy rates from a large-scale commercial program using nelore cattle (*Bos indicus*) donors. *Theriogenology*, 75(9):1640–1646.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Santos, J. C. C., Silva, D. M. R., Amorim, D. J., Rosa, V. R., Santos, A. L. F., Veline, E. D., Carbonari, C. A., e Silva, M. A. (2022). Glyphosate hormesis attenuates water deficit stress in safflower (*Carthamus tinctorius* L.) by modulating physiological and biochemical mediators. *Science of The Total Environment*, 810:152204.

Santos, J. C. C., Silva, D. M. R., Amorim, D. J., Sab, M. P. V., e Silva, M. A. (2021). Glyphosate hormesis mitigates the effect of water deficit in safflower (*Carthamus tinctorius* L.). *Pest Management Science*, 77(4):2029–2044.

Seneda, M. M., Zangirolamo, A. F., Bergamo, L. Z., e Morotti, F. (2020). Follicular wave synchronization prior to ovum pick-up. *Theriogenology*, 150:180–185.

Souza-Cácares, M. B., Fialho, A. L. L., Silva, W. A. L., Cardoso, C. J. T., Pöhland, R., Martins, M. I. M., e Melo-Sterza, F. A. (2019). Oocyte quality and heat shock proteins in oocytes from bovine breeds adapted to the tropics under different conditions of environmental thermal stress. *Theriogenology*, 130:103–110.

Viana, J. (2022). 2021 statistics of embryo production and transfer in domestic farm animals. *Embryo Technology Newsletter*, 40(4):24–38.

Wang, J., Li, J., Wang, F., Xiao, J., Wang, Y., Yang, H., Li, S., e Cao, Z. (2020). Heat stress on calves and heifers: a review. *Journal of Animal Science and Biotechnology*, 11(1).

2 MODELAGEM ESTATÍSTICA DE DADOS DE CONTAGEM DE OÓCITOS SUPERDISPERSOS, USANDO O SOFTWARE R

Resumo

Na última década, a produção de embriões *in vitro* (PIVE) foi alavancada, principalmente, pela seleção genômica, permitindo a utilização de novilhas muito jovens com alto mérito genético. No entanto, a combinação de fatores genéticos e ambientais tonaram a PIVE de gado holandês desafiadora. Vários fatores precisam ser coordenados para obter um bezerro oriundo da PIVE. Por isso, a modelagem estatística correta tem importância para a análise e a interpretação dos efeitos desses fatores e posterior tomada de decisão. Os dados oriundos desses estudos, geralmente, são dados de contagem superdispersos e correlacionados. Neste trabalho, os aspectos relacionados a superdispersão, correlação em dados longitudinais, testes de hipóteses sobre os componentes de variância, seleção de modelos e implicações práticas são abordados, para análise do número de oócitos totais obtidos de fêmeas da raça bovina holandesa em um estudo observacional conduzido na RuAnn Genetics Laboratory (Riverdale, Califórnia, EUA), em 2020. Apresentam-se os códigos desenvolvidos no software `R` para ajuste dos modelos lineares generalizados, modelos lineares generalizados mistos e modelos combinados. Discute-se como selecionar o modelo que melhor se ajusta aos dados e como identificar os fatores que influenciam na obtenção de oócitos. Os modelos lineares generalizados mistos e modelos combinados ajustaram-se satisfatoriamente, captando a variabilidade extra e a correlação entre as medidas longitudinais.

Palavras-chave: Produção *in vitro* de embriões, Modelos lineares generalizados, Modelos mistos, Dados de contagem.

2.1 Introdução

A biotecnologia reprodutiva que mais se expandiu nos últimos anos foi a produção de embriões *in vitro* (PIVE), contribuindo, efetivamente, para o melhoramento genético do gado leiteiro, com aplicações em escala comercial e possibilitando a propagação dos animais geneticamente superiores (Cavalieri et al., 2018; Sirard, 2018; Demétrio et al., 2020). O sucesso dessa biotecnologia tem grande contribuição da análise genômica que revolucionou a criação de gado leiteiro e aumentou a pressão de seleção, encurtando o intervalo de gerações (Sirard, 2018). Apesar dos constantes avanços, os programas de coleta de óvulos (OPU) continuam a evidenciar alguns desafios que comprometem a plena eficácia da técnica.

Os programas de PIVE podem ser afetados por vários fatores, incluindo o número de oócitos totais coletados, que é um dos mais importantes (Garcia et al., 2020). Outros

fatores que estão relacionados com a eficiência da PIVE são: estresse térmico (Chebel et al., 2008; Souza-Cácares et al., 2019), status reprodutivo (Demétrio et al., 2020; Wang et al., 2020), intervalo das OPUs (Pontes et al., 2011) e protocolos hormonais (Demétrio et al., 2020; Ongaratto et al., 2020). Dessa forma, laboratórios e empresas vêm buscando estratégias que sirvam para aprimorar e otimizar métodos de uso da PIVE em programas de grande escala. Somando-se a esses esforços, metodologias estatísticas adequadas devem ser empregadas para analisar corretamente os dados.

Os conjuntos de dados utilizados nas análises dos resultados da PIVE, muitas vezes, são oriundos de estudos observacionais e não se encaixam no escopo dos métodos estatísticos tradicionais que são usados para analisar dados provenientes da distribuição normal. Nesses estudos, são comuns dados na forma de contagens que assumem apenas valores inteiros não negativos (0, 1, 2, $\cdots$), como, por exemplo, número de oócitos totais (Demétrio et al., 2020; Garcia et al., 2020) e número de embriões (Ongaratto et al., 2020). São, também, comuns os dados na forma de proporções (Stroup, 2015), por exemplo, taxa de clivagem (Demétrio et al., 2020), taxa de produção de blastocisto (Cavalieri et al., 2018) e taxas de prenhez (Chebel et al., 2008).

Em geral, dados na forma de contagens ou de proporções não atendem às pressuposições dos modelos lineares clássicos. Uma das possíveis formas usadas para contornar esse problema é pela transformação de dados tal as pressuposições de normalidade e homogeneidade de variâncias sejam satisfeitas (Bolker et al., 2009). No entanto, nem sempre é eficiente e, além disso, é ignorada a natureza da variável resposta conforme destacado por Bolker et al. (2009). Devido ao avanço dos métodos computacionais nos últimos 30 anos (Stroup, 2015; Kosma et al., 2019), esses tipos de dados podem ser analisados usando a classe dos modelos lineares generalizados (MLGs) desenvolvidos por Nelder e Wedderburn (1972) e de modelos lineares generalizados mistos (MLGMs) (Bolker et al., 2009; Stroup, 2015).

O modelo padrão para análise de dados na forma de contagens é o de Poisson, um caso especial dos MLGs. Sob as pressuposições básicas independência e taxa constante de ocorrência dos eventos o modelo Poisson tem variância igual à média. Se uma (ou ambas) dessas suposições falha(m) a variância observada será maior (menor) (Hoef e Boveng, 2007; Bolker et al., 2009; Demétrio et al., 2014; Kosma et al., 2019) do que a média resultando no que é conhecido como superdispersão (subdispersão) (Hinde e Demétrio, 1998), o que é comum para conjuntos de dados biológicos. Na presença de superdispersão, os modelos quase-Poisson e binomial negativo podem ser utilizados (Hoef e Boveng, 2007; Demétrio et al., 2014).

Outra abordagem para tratar o problema da superdispersão é a utilização dos MLGMs (Molenberghs et al., 2007; Fatoretto et al., 2018), pois a inclusão de efeitos aleatórios no preditor linear permite acomodar a extra-variabilidade causada por correlação intraclasse. Segundo Bolker et al. (2009), os MLGMs são a melhor ferramenta para ana-

lisar dados não normais que envolvem efeitos aleatórios, todavia a inclusão de efeitos aleatórios no preditor linear pode lidar com a variabilidade e a superdispersão até certo ponto. Em situações práticas, os MLGMs podem não capturar toda a superdispersão presente, comprometendo as conclusões (Molenberghs et al., 2007; Iddi e Molenberghs, 2012). Diante dessas limitações, Molenberghs et al. (2007), e Molenberghs et al. (2010) desenvolveram os modelos combinados (MC), que permitem acomodar a superdispersão e a correlação em dados não normais medidos longitudinalmente por meio da introdução de efeitos aleatórios, tanto no preditor linear, quanto na média da distribuição.

Ignorar a presença de superdispersão e de correlação pode ter influência sobre os erros-padrão das estimativas dos parâmetros e sobre as estatísticas de testes levando à seleção de modelos excessivamente complexos e, consequentemente, causando erros na avaliação dos resultados da análise (Molenberghs et al., 2007; Fatoretto et al., 2018; Kosma et al., 2019).

Diante do grande número de softwares e métodos disponíveis para analisar dados de contagem, o desafio consiste em escolher o melhor método, quais efeitos fixos ou aleatórios devem ser considerados no modelo e como interpretar os resultados, principalmente, no contexto dos estudos observacionais. Portanto, os objetivos deste trabalho são: (i) discutir como selecionar modelos para analisar adequadamente os dados de contagem com superdispersão no contexto dos estudos observacionais e (ii) avaliar os fatores ligados à quantidade oócitos obtidos por OPU. Pretende-se fornecer diretrizes metodológicas para profissionais ligados às ciências agrárias trazendo ao mesmo tempo, discussões do ponto de vista estatístico e prático. Ilustram-se, ainda, as aplicações de MLGs, MLGMs e MC, usando o software `R` (R Core Team, 2022).

O restante deste artigo está organizado como se segue. Na Seção 2.2 apresenta-se um estudo de caso que motivou esse trabalho. Modelos usados para a análise de dados na forma de contagens e técnicas de seleção de modelos e diagnósticos são descritos na Seção 2.3. Na Seção 2.4, é apresentada e discutida a análise dos dados do estudo de caso, além dos códigos em `R`. Algumas considerações gerais e conclusões são apresentadas na Seção 2.5.

2.2 Estudo de caso

Para identificar os fatores que afetam a obtenção de oócitos totais (OT) provenientes de fêmeas da raça bovina holandesa, utilizou-se um conjunto de dados obtido de um estudo observacional de propriedade da Fazenda RuAnn and Maddox Dairy, em Riverdale, Califórnia (USA), (N36°28'26.945", W119°55'47.65"). A composição do banco de dados foi obtida a partir de sessões de aspiração folicular guiada por ultrassom (OPU), realizadas entre 06 de janeiro e 28 de dezembro de 2020.

Os dados referem-se a 1148 de sessões de OPU, realizadas em 318 fêmeas bovinas

doadoras de oócitos de três status: novilhas, vacas em lactação e vacas secas. A seleção dos animais foi realizada considerando os altos valores zootécnicos do plantel das fêmeas com a intenção de multiplicar a genética desejada. Todos os animais foram, devidamente, registrados na Holstein Association USA.

As sessões de OPU não tiveram uma rotina específica de frequência, tendo sido realizadas, em média 99 sessões por mês. Para o estudo, avaliaram-se os seguintes fatores: doadora ($n = 318$), status da doadora (H - novilhas, M - em lactação e D - secas), período do ano (P1 e P2), o intervalo entre as OPU realizadas na mesma doadora (2wks - duas, 3wks - três e na - mais do que três semanas) e número de injeções de hormônio folículo-estimulante - FSH (0, 1 e 5), quanto a seus efeitos sobre a obtenção de oócitos. A variável resposta considerada foi o número de oócitos totais (OT).

Para verificar o efeito do período do ano, consideraram-se as sessões de OPU realizadas no P1 (período de temperatura mais elevada que compreende os meses de junho a outubro, 464 aspirações) e no período P2 (período de temperatura menos elevada que compreende os sete meses restantes, 684 aspirações) (Figura 2.1).

Para as doadoras que receberam FSH, utilizou-se o hormônio de liberação de gonadotrofina - GnRH (Fertagyl®, Merck, 129 g, IM) para induzir a formação de um corpo lúteo e tentar aumentar as concentrações de progesterona, e para sincronizar a emergência da onda folicular. O protocolo de uma única injeção de FSH consistiu na aplicação intramuscular (IM) de 100 mg de FSH para novilhas e 140 mg de FSH para vacas (Folltropin®, Vetoquinol) 36 h após GnRH. O FSH consistiu na injeção de 2,5 mL de 400 mg de FSH diluído em 10 mL de ácido hialurônico (HA) à 0,5%. A OPU foi realizada entre 48 a 50 h após o FSH. Para cinco injeções de FSH, o protocolo consistiu em 100 mg de FSH para novilhas e 140 mg de FSH para vacas dividido em 5 injeções IM iguais (intervalos de 10-14 h), 36 h após GnRH. O FSH consistiu em injeções de 5×1 mL e $5 \times 1{,}4$ mL para novilhas e vacas, respectivamente, de 400 mg de FSH diluídos em 20 mL de solução salina. A OPU foi realizada entre 18 a 20 h para novilhas e 24 a 30 h para vacas após a última injeção de FSH.

2.3 Abordagem estatística

Como já descrito, os resultados de interesse são dados de contagens como: os número de oócitos totais. Para esse tipo de variável resposta, o modelo Poisson é o padrão. No entanto, essa distribuição apresenta uma relação média e variância muito restritiva. Com o intuito de permitir maior flexibilidade na relação de média e variância, extensões foram desenvolvidas, como as discutidas nas seções que se seguem. O modelo Poisson padrão para contagens, assim como o modelo binomial para dados de proporção são exemplos típicos de modelo linear generalizado (Nelder e Wedderburn, 1972).

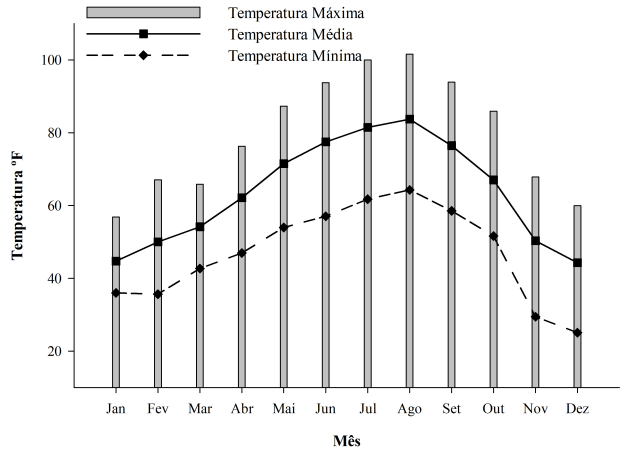

Figura 2.1. Dados climáticos de 2020 - Riverdale, Califórnia (USA).

Introdução aos modelos lineares generalizados

Os modelos lineares generalizados (MLGs) foram definidos por Nelder e Wedderburn (1972) como uma estrutura geral para lidar com uma gama de modelos estatísticos utilizando dados normais ou não, apresentando-se como uma teoria unificadora da modelagem estatística. Um MLG envolve três componentes:

(i) um componente aleatório, representado pela variável aleatória Y_i, ou seja, a variável resposta, com distribuição pertencente à família exponencial, com função densidade de probabilidade (fdp)

$$f(y_i;\theta_i,\phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i,\phi)\} \qquad i = 1,\ldots,n, \tag{2.1}$$

com funções conhecidas $b(\cdot)$ e $c(\cdot)$; θ_i e ϕ são denominados parâmetro canônico e parâmetro de dispersão, respectivamente (Molenberghs et al., 2007). A média e a variância podem ser obtidas por meio das expressões: $\mathrm{E}(Y_i) = \mu_i = b'(\theta_i)$ e $\mathrm{Var}(Y_i) = \phi b''(\theta_i) = V(\mu_i)$, sendo $V(\cdot)$ a função de variância (Demétrio et al., 2014);

(ii) um preditor linear η que é uma combinação linear de variáveis explanatórias

$$\eta_i = \boldsymbol{x_i}^T\boldsymbol{\beta},$$

em que $\boldsymbol{\beta}$ é o vetor $(p \times 1)$ de parâmetros a serem estimados e $\boldsymbol{x_i} = [\boldsymbol{x}_{11},\ldots,\boldsymbol{x}_{1p}]^T$ é a i-ésima coluna da matriz de delineamento do modelo $(n \times p)$; e

(iii) uma função de ligação $g[\mathrm{E}(Y_i)] = \eta_i$, que relaciona a média da distribuição ao preditor linear.

Portanto, para ajustar um MLG é necessário escolher uma distribuição para a variável resposta, uma matriz que representa o preditor linear do modelo e uma função de ligação. Casos particulares dos MLGs são os modelos lineares clássicos, considerando distribuição normal e função de ligação identidade, modelos logísticos para dados de proporções, modelos Poisson para dados de contagens dentre outros.

A estimativa do vetor $\boldsymbol{\beta}$ de parâmetros é obtida pelo método de máxima verossimilhança, usando um procedimento iterativo ponderado baseado no algoritmo do escore de Fisher (Nelder e Wedderburn, 1972) e na convergência

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z},$$

em que $\boldsymbol{X}$ é a matriz de do modelo, $\boldsymbol{W}$ é uma matriz diagonal com elementos $W_i = \frac{1}{V(\mu_i)[g'(\mu_i)]^2}$, $g'(\mu_i) = \frac{dg(\mu_i)}{d\mu_i}$ e $\boldsymbol{z}$ é a variável resposta ajustada com $z_i = \eta_i + g'(\mu_i)(y_i - \mu_i)$.

As estimativas de $\boldsymbol{\beta}$ minimizam a função "deviance" dada por

$$\begin{aligned} S_p &= 2(\hat{\ell}_n - \hat{\ell}_p) = \phi^{-1}D_p \\ &= 2\phi^{-1}\sum_{i=1}^{n}[y_i(\tilde{\theta}_i - \hat{\theta}_i) + b(\hat{\theta}_i) - b(\tilde{\theta}_i)], \end{aligned}$$

em que S_p é a "scaled deviance", D_p é a "deviance"; $\hat{\ell}_n$ e $\hat{\ell}_p$ são os máximos do logaritmo da função de verossimilhança para os modelos saturado (modelo que tem n parâmetros, um por observação) e aquele sob estudo (com p parâmetros), respectivamente e $\hat{\theta}_i$ e $\tilde{\theta}_i$ são as estimativas de máxima verossimilhança, para os modelos saturado e aquele sob estudo. Mede a distância entre os valores observados e os ajustados em unidades de logaritmo da função de verossimilhança. Quanto melhor for o ajuste do modelo, menor será o valor da "deviance".

Uma medida alternativa para verificação do ajuste de um modelo é a estatística X^2 de Pearson, dada por

$$X^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\widehat{\mathrm{Var}}(Y_i)}.$$

Modelo Poisson

Sejam $Y_i, i = 1, \ldots, n$, variáveis aleatórias independentes que representam contagens. É razoável assumir, inicialmente, que Y_i tem distribuição Poisson com média μ_i e função de probabilidade dada por

$$P(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots, \quad \mu > 0,$$

que é um caso particular dos MLGs, com parâmetro de dispersão $\phi = 1$. Mostra-se que

$$\mathrm{Var}(Y_i) = \mathrm{E}(Y_i) = \mu_i$$

e, portanto, $\mathrm{Var}(Y_i)/\mathrm{E}(Y_i) = 1$ o que é um resultado muito restritivo. Para o modelo clássico de Poisson, assume-se $Y_i \sim \mathrm{Poisson}(\mu_i)$, com função de ligação logarítmica, isto é, $\eta_i = \log(\mu_i) = \boldsymbol{x_i}^T\boldsymbol{\beta}$. Esse modelo pode ser ajustado, usando-se a função `glm()` do pacote **stats** do software `R` em conjunto com o argumento `family=poisson()`.

Para o modelo Poisson, a "deviance" residual e a estatística X^2 de Pearson são

$$D_p = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]$$

e

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

em que $\hat{\mu}_i$, $i = 1, \ldots, n$, são os valores ajustados para o modelo corrente. Assintoticamente ($n \to \infty$), D_p e X^2 têm distribuição que converge para uma distribuição χ^2 com $n - p$ graus de liberdade, sob certas condições de regularidade (Jørgensen, 2006).

Então, para um modelo bem ajustado é esperado que D_P e X_P^2 sejam, aproximadamente, iguais ao número de graus de liberdade do resíduo. Quando isso não acontece é porque a variância observada é diferente daquela predita pelo modelo Poisson (Hinde e Demétrio, 1998; Demétrio et al., 2014).

Na prática, a estimativa da razão $\mathrm{Var}(Y_i)/\mathrm{E}(Y_i)$ é usada como índice de dispersão, podendo indicar problemas de superdispersão (> 1) ou subdispersão (< 1), necessitando de extensões. As causas mais comuns para a ocorrência da superdispersão são: heterogeneidade entre as observações, especificação incorreta do modelo, omissão de termos importantes no preditor linear, correlação entre as respostas, combinação de um ou mais das anteriores (Hinde e Demétrio, 1998; Demétrio et al., 2014; Moral et al., 2017; Agresti, 2019). A seguir, são apresentados alguns modelos que levam em conta superdispersão (Demétrio et al., 2014).

Modelo quase-Poisson

A forma mais simples de levar em conta a variabilidade extra é assumir um modelo de superdispersão constante, isto é, permitir que o parâmetro de dispersão ϕ seja maior do que 1, de modo que

$$\mathrm{Var}(Y_i) = \phi\mathrm{Var}(\mu_i) = \phi\mu_i. \tag{2.2}$$

Essa abordagem requer apenas a especificação de primeiro e segundo momentos de uma distribuição e as estimativas dos parâmetros é obtida por quase verossimilhança. A estimativa do parâmetro ϕ é obtida por

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

em que X^2 é a estatística generalizada de Pearson para o modelo Poisson (McCullagh e Nelder, 1989).

A estimativa do vetor de parâmetros $\boldsymbol{\beta}$ é a mesma que a do modelo Poisson. Entretanto, os erros-padrão de $\hat{\boldsymbol{\beta}}$ ficam inflacionados pelo fator de superdispersão, isto é, ficam multiplicados por $\sqrt{\hat{\phi}}$.

Esse modelo pode ser ajustado, usando-se a função `glm()` do pacote **stats** do software `R` em conjunto com o argumento `family=quasipoisson()`.

Modelo binomial negativo

O modelo de Poisson tem como pressuposições independência dos eventos e taxa constante de ocorrência dos eventos, o que na prática é raro. Assumindo que a taxa de ocorrência dos eventos pode variar de acordo com uma determinada distribuição, pode-se pensar um modelo de dois estágios para distribuição da variável resposta Y_i. Inicialmente, supõe-se que Y_i condicional a T_i tem distribuição de Poisson, isto é, $Y_i|T_i \sim \text{Poisson}(T_i)$. Assumindo que T_i é uma variável aleatória com distribuição gama, isto é, $T_i \sim \text{gama}(\alpha_1, \alpha_{2i})$, a distribuição marginal de Y_i é binomial negativa com média e variância dadas, respectivamente, por $\text{E}(Y_i) = \alpha_1/\alpha_{2i} = \mu_i$ e

$$\text{Var}(Y_i) = \mu_i + \mu_i^2/\alpha_1 = \mu_i(1 + \mu_i/\alpha_1) \tag{2.3}$$

permitindo incorporar variabilidade maior do que a média. Para α_1 conhecido, a distribuição binomial negativa pertence à família exponencial (2.1) e o algoritmo de estimação dos MLGs pode ser usado (Venables e Ripley, 2002; Demétrio et al., 2014).

Esse modelo pode ser ajustado usando-se a função `glm.nb()` do pacote **MASS** (Venables e Ripley, 2002) ou da função `aodml()` do pacote **aods3** (Lesnoff e Lancelot, 2013) no software `R`.

Modelo Poisson-normal

Uma forma alternativa de incorporar a superdispersão é pela adição de um efeito aleatório no preditor linear η_i em nível de observação, que represente efeitos latentes adicionais não explicados (Hinde, 1982). Supondo que a distribuição condicional Y_i dado Z_i é um modelo de Poisson, isto é, $Y_i|Z_i \sim \text{Poisson}(\lambda_i)$, com função de ligação logarítmica e preditor linear $\log(\lambda_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} + Z_i$ em que Z_i é uma variável aleatória com distribuição normal, isto é, $Z_i \sim \text{N}(0, \sigma^2)$, tem-se o modelo Poisson-normal, que é o caso (univariado) mais simples de modelo linear generalizado misto (MLGM). Não há forma fechada para a distribuição marginal de Y_i mas a média e a variância são dadas, respectivamente, por

$$\text{E}(Y_i) \;=\; \text{E}[\text{E}(Y_i|Z_i)] = \text{E}[\exp(x_i^T\boldsymbol{\beta} + Z_i)] = \text{e}^{x_i^T\boldsymbol{\beta} + \frac{1}{2}\sigma^2} = \mu_i$$

e

$$\begin{aligned}\text{Var}(Y_i) &= \text{E}[\text{Var}(Y_i|Z_i)] + \text{Var}[\text{E}(Y_i|Z_i)] = \text{E}[\exp(x_i^T\boldsymbol{\beta} + Z_i)] + \text{Var}[\exp(x_i^T\boldsymbol{\beta} + Z_i)] \\ &= \text{e}^{x_i^T\boldsymbol{\beta} + \frac{1}{2}\sigma^2} + \text{e}^{2x_i^T\boldsymbol{\beta} + \sigma^2}(\text{e}^{\sigma^2} - 1) = \mu_i + \phi\mu_i^2.\end{aligned}$$

Vê-se que a forma da expressão da variância é a mesma que a do modelo binomial negativo dada pela expressão (2.3).

Em muitas situações, as mensurações de diferentes atributos são feitas na mesma planta e/ou animal, por serem coletadas ao longo do tempo ou em grupos genéticos, e, portanto, exibem alguma forma de dependência, caracterizando uma variável aleatória multivariada.

Seja Y_{ij} a variável aleatória que representa a j-ésima medida longitudinal no i-ésimo indivíduo, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$, e o grupo de n_i medidas em um vetor $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$ com vetor de médias $\boldsymbol{\lambda_i} = (\lambda_{i1}, \ldots, \lambda_{in_i})^T$. Assumindo-se que, condicionalmente a um vetor q-dimensional de efeitos aleatórios $\boldsymbol{b}_i$ com distribuição $N_q(\boldsymbol{0}, \boldsymbol{D})$, em que $\boldsymbol{0}$ é um vetor de zeros e $\boldsymbol{D}$ a matriz de variâncias e covariâncias, as variáveis respostas Y_{ij} são consideradas independentes e, têm distribuição que pertence à família exponencial

$$f(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \phi) = \exp\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij}, \phi)\},$$

com

$$g(\mu_{ij}) = g[\mathrm{E}(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta})] = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i$$

para uma função de ligação conhecida $g(\cdot)$, com os vetores $\boldsymbol{x}_{ij}$ e $\boldsymbol{z}_{ij}$ p-dimensional e q-dimensional associados às covariáveis e aos efeitos aleatórios, respectivamente, $\boldsymbol{\beta}$ é o vetor p-dimensional de efeitos fixos e ϕ o parâmetro de dispersão. Denota-se, também, por $f(\boldsymbol{b}_i|\boldsymbol{D})$ a função densidade dos efeitos aleatórios $\boldsymbol{b}_i$. Assumindo que os dados são contagens, tem-se o modelo Poisson-normal multivariado

$$\begin{aligned} Y_{ij}|\boldsymbol{b}_i &\sim \mathrm{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i) \\ \boldsymbol{b}_i &\sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D}). \end{aligned} \tag{2.4}$$

Tem-se que $\mathrm{E}(\lambda_{ij}) = \exp\left(x_{ij}^T\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij}\right)$, $\mathrm{Var}(\lambda_{ij}) = \exp\left(x_{ij}^T\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij}\right) + \exp(2x_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij})\{\exp(\boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij}) - 1\}$ e $\mathrm{Cov}(\lambda_{ij}, \lambda_{ik}) = \exp[x_{ij}^T\boldsymbol{\beta} + x_{ik}^T\boldsymbol{\beta} + \frac{1}{2}(\boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij} + \boldsymbol{z}_{ik}^T\boldsymbol{D}\boldsymbol{z}_{ik})]\{\exp(\boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ik}) - 1\}$. Logo, marginalmente, o vetor de médias de $\boldsymbol{Y}_i$ tem elementos

$$\begin{aligned} \mathrm{E}(Y_{ij}) &= \mathrm{E}[\mathrm{E}(Y_{ij}|\boldsymbol{b_i})] = \mathrm{E}[\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)] \\ &= \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})\mathrm{E}[\exp(\boldsymbol{z}_{ij}^T\boldsymbol{b}_i)] = \exp\left(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij}\right) = \mu_{ij} \end{aligned}$$

e a matriz de variâncias-covariâncias $\boldsymbol{Y}_i$ é dada por

$$\begin{aligned} \mathrm{Var}(\boldsymbol{Y}_i) &= \mathrm{E}[\mathrm{Var}(\boldsymbol{Y}_i|\boldsymbol{b_i})] + \mathrm{Var}[\mathrm{E}(\boldsymbol{Y}_i|\boldsymbol{b_i})] \\ &= \mathrm{E}[\exp(\boldsymbol{X}_i^T\boldsymbol{\beta} + \boldsymbol{Z}_i^T\boldsymbol{b}_i)] + \mathrm{Var}[\exp(\boldsymbol{X}_i^T\boldsymbol{\beta} + \boldsymbol{Z}_i^T\boldsymbol{b}_i)] \\ &= \boldsymbol{M}_i + \boldsymbol{M}_i[\exp(\boldsymbol{Z}_i^T\boldsymbol{D}\boldsymbol{Z}_i) - \boldsymbol{J}_{n_i}]\boldsymbol{M_i}, \end{aligned}$$

em que $\boldsymbol{M}_i$ é uma matriz diagonal com $\boldsymbol{\mu}_i$ ao longo da diagonal.

Esse modelo pode ser ajustado por diversos pacotes do `R`, no entanto, o pacote **lme4** (Bates et al., 2015) é o mais tradicional para análise de MLGMs. Esse pacote fornece a função de ajuste `glmer()` que permite especificar os preditores lineares com efeitos fixos e aleatórios.

Modelos combinados

Os modelos combinados (MC) incorporam as duas principais estratégias utilizadas para introduzir efeitos aleatórios na estrutura GLM. A primeira envolve uma distribuição conjugada para o parâmetro, isto é, refere-se ao fato que as densidades hierárquicas e de efeitos aleatórios têm formas algébricas semelhantes (Lee et al., 2017). Para dados de contagem, por exemplo, a mistura da distribuição Poisson e gama, origina o modelo conjugado binomial negativo. A segunda é pela inserção de efeitos aleatórios no preditor linear.

A formulação do MC é semelhante àquela apresentada para os MLGMs. Supõe-se, inicialmente, que, condicional aos efeitos aleatórios independentes $\boldsymbol{b}_i$ e θ_{ij}, a variável aleatória Y_{ij} tem distribuição pertencente à família exponencial

$$f(y_{ij}|\boldsymbol{b}_i, \theta_{ij}, \phi) = \exp\{\phi^{-1}[y_{ij}\lambda_{ij} - \psi(\lambda_{ij})] + c(y_{ij}, \phi)\},$$

com média dada por

$$\mathrm{E}(Y_{ij}|\boldsymbol{b}_i, \lambda_{ij}) = \theta_{ij}k_{ij}.$$

Assume-se que $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{in_i})^T$ tem distribuição com vetor de médias $\mathrm{E}(\boldsymbol{\theta}_i) = \boldsymbol{\Phi}_i$ e matriz de variâncias e covariâncias $\mathrm{Var}(\boldsymbol{\theta}_i) = \boldsymbol{\Sigma}_i$, isto é, $\mathrm{E}(\theta_{ij}) = \phi_{ij}$; $\mathrm{Var}(\theta_{ij}) = \sigma_{i,jj}$ e $\mathrm{Cov}(\theta_{ij}, \theta_{ik}) = \sigma_{i,jk}$. Além disso, $k_{ij} = g(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)$ e supõe-se que $\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$.

No contexto dos dados de contagens, assume-se que $Y_{ij}|\boldsymbol{b}_i, \theta_{ij} \sim \mathrm{Poisson}(\lambda_{ij})$, isto é, como Molenberghs et al. (2007) apresentam de forma resumida

$$\begin{aligned} Y_{ij}|\boldsymbol{b}_i &\sim \mathrm{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \theta_{ij}k_{ij} = \theta_{ij}\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i) \\ \boldsymbol{b}_i &\sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D}) \\ \mathrm{E}(\boldsymbol{\theta}_i) &= \boldsymbol{\Phi}_i \\ \mathrm{Var}(\boldsymbol{\theta}_i) &= \boldsymbol{\Sigma}_i. \end{aligned} \tag{2.5}$$

Marginalmente, o vetor de médias tem elementos

$$\begin{aligned} \mathrm{E}(Y_{ij}) &= \mathrm{E}\{\mathrm{E}[\mathrm{E}(Y_{ij}|\theta_{ij}, \boldsymbol{b}_i)]\} \\ &= \phi_{ij}\exp\left(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij}\right) = \mu_{ij} \end{aligned}$$

e a matriz de variâncias-covariâncias dada por

$$\begin{aligned} \mathrm{Var}(\boldsymbol{Y}_i) &= \mathrm{E}\{\mathrm{E}[\mathrm{Var}(\boldsymbol{Y}_i|\theta_{ij}, \boldsymbol{b}_i)]\} + \mathrm{E}\{\mathrm{Var}[\mathrm{E}(\boldsymbol{Y}_i|\theta_{ij}, \boldsymbol{b}_i)]\} + \mathrm{Var}\{\mathrm{E}[\mathrm{E}(\boldsymbol{Y}_i|\theta_{ij}, \boldsymbol{b}_i)]\} \\ &= \boldsymbol{M}_i + \boldsymbol{M}_i[\boldsymbol{P}_i - \boldsymbol{J}_{n_i}]\boldsymbol{M}_i, \end{aligned}$$

em que $\boldsymbol{M}_i$ é uma matriz diagonal com $\boldsymbol{\mu}_i$ ao longo da diagonal e $\boldsymbol{P}_i$ é uma matriz com elementos

$$p_{i,jk} = \exp\left(\frac{1}{2}\boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ik}\right)\frac{\sigma_{i,jk}+\phi_{ij}\phi_{ik}}{\phi_{ij}\phi_{ik}}\exp\left(\frac{1}{2}\boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ik}\right).$$

Se $\boldsymbol{\theta}_i$ tiver distribuição gama, resulta o modelo Poisson-gama-normal.

Os efeitos aleatórios $\boldsymbol{b}_i$ com distribuição normal para acomodar a correlação e os efeitos θ_{ij} com distribuição gama para modelar a superdispersão tornam o MC flexível para a análise de dados na forma de contagens. Esse modelo tem como casos especiais os modelos binomial negativo e Poisson-normal, multivariados e univariados, bem como o modelo Poisson padrão.

As estimativas do vetor de parâmetros fixos $\boldsymbol{\beta}$ e dos componentes de variância da matriz $\boldsymbol{D}$, são obtidas pela maximização da função de verossimilhança marginal que resulta da integração da função de verossimilhança conjunta em relação aos efeitos aleatórios $\boldsymbol{b}_i$ e θ_{ij}. Molenberghs et al. (2007) obtiveram a expressão para a distribuição de probabilidades conjunta de $\boldsymbol{Y}_i$ e dos efeitos aleatórios independentes $\boldsymbol{b}_i$ e θ_{ij}, com elementos dados por

$$f_i(y_{ij}|\boldsymbol{\beta},\boldsymbol{b}_i,\theta_{ij}) \;=\; f_{ij}(y_{ij}|\boldsymbol{\beta},\boldsymbol{b}_i,\boldsymbol{\theta}_i)f(\boldsymbol{b}_i|\boldsymbol{D})f(\boldsymbol{\theta}_i|\alpha_{1j},\alpha_{2j}).$$

A contribuição do indivíduo i para a função de verossimilhança é dada por

$$f_i(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{D},\alpha_{1j},\alpha_{2j}) \;=\; \int\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\beta},\boldsymbol{b}_i,\boldsymbol{\theta}_i)f(\boldsymbol{b}_i|\boldsymbol{D})f(\boldsymbol{\theta}_i|\alpha_{1j},\alpha_{2j})d\boldsymbol{b}_i d\boldsymbol{\theta}_i.$$

em que $\boldsymbol{\beta}$ agrupa todos os parâmetros modelo condicional para $\boldsymbol{Y}_i$. Logo, a função de verossimilhança marginal é dada por

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{D},\alpha_1,\alpha_2) &= \prod_{i=1}^{N} f_i(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{D},\alpha_1,\alpha_2) \\ &= \prod_{i=1}^{N}\int\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\beta},\boldsymbol{b}_i,\boldsymbol{\theta}_i)f(\boldsymbol{b}_i|\boldsymbol{D})f(\boldsymbol{\theta}_i|\boldsymbol{\alpha}_{1j},\boldsymbol{\alpha}_{2j})d\boldsymbol{b}_i d\boldsymbol{\theta}_i \qquad (2.6) \end{aligned}$$

O problema de maximizar (2.6) é a presença de N integrais. Molenberghs et al. (2007), combinando técnicas analíticas e numéricas, propuseram uma marginalização parcial, para o caso do MC (2.5) pela integração apenas dos efeitos aleatórios gama assumidos independentes, resultando em

$$f(y_{ij}|\boldsymbol{b}_i) = \binom{\alpha_{1j}+y_{ij}-1}{\alpha_{1j}-1}\left(\frac{\alpha_{2j}}{1+k_{ij}\alpha_{2j}}\right)^{y_{ij}}\left(\frac{1}{1+k_{ij}\alpha_{2j}}\right)^{\alpha_{1j}} k_{ij}^{y_{ij}}, \qquad (2.7)$$

em que $k_{ij} = \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\boldsymbol{z}_{ij}^T\boldsymbol{b}_i)$.

Por integração numérica de (2.7) em relação aos efeitos normais, empregando a quadratura adaptativa de Gauss-Hermite ou a aproximação de Laplace, obtém-se a função de verossimilhança marginal (Molenberghs et al., 2007, 2010). Essa técnica pode ser adotada pelo uso da função `glmer.nb()` do pacote **lme4** (Bates et al., 2015) ou do procedimento SAS NLMIXED (SAS Institute Inc, 2015).

Verificação de ajuste de modelos e diagnósticos

Como etapa importante para o estabelecimento de um MLG ou de qualquer outro modelo é a detecção de possíveis falhas, que pode ser feita por meio de análises gráficas. Em geral, são utilizados resíduos, sendo o componente da "deviance" o mais comum. Um gráfico de fácil compreensão é o gráfico semi-normal de probabilidades com envelope de simulação para os resíduos que foi, originalmente, proposto por Atkinson (1987) para verificar a qualidade de ajuste. Pode ser obtido no `R`, utilizando o pacote **hnp** (Moral et al., 2017) e para uma quantidade de pontos inferior a 5% dos pontos, o ajuste do modelo é considerado satisfatório. Quando esses gráficos não são obtidos diretamente, é necessário implementá-los como uma nova classe de modelo declarando três funções: `dfun`, `sfun` e `ffun`, para obter o diagnóstico, simular as variáveis aleatórias e reajustar o modelo para as variáveis simuladas, respectivamente (Moral et al., 2017).

Seleção de modelos - inferência sobre efeitos aleatórios

Para a seleção da parte aleatória de um modelo, são feitos testes de hipóteses sobre os componentes de variância/covariância, do tipo $H_0 : \sigma_{ij} = 0$. Pode-se usar o teste da razão de verossimilhanças (LRT) que se baseia na comparação dos valores das funções de verossimilhanca de dois modelos aninhados, tendo o mesmo número de parâmetros de efeito fixo e diferentes números de parâmetros de efeito aleatório. É dado por

$$LRT = -2[\text{logLik(modelo reduzido) - logLik(modelo completo)}],$$

em que logLik é o logaritmo da função de verossimilhança.

Quando o teste não é no limite do espaço paramétrico, assintoticamente, $LRT \sim \chi^2_\nu$, em que ν é a diferença entre o número de parâmetros dos dois modelos. Quando o teste é no limite do espaço paramétrico ($H_0 : \sigma_{ii} = 0$ versus $H_a : \sigma_{ii} > 0$), LRT tem distribuição que é uma mistura de χ^2's (Zhang e Lin, 2008).

No caso de modelos de componentes de variância com independência entre os efeitos aleatórios a mistura de χ^2's é dada por

$$\sum_{m=0}^{k'} 2^{-k'} \binom{k'}{m} \chi^2_m,$$

sendo k' o número de componentes de variância sob H_0.

No entanto, na maioria dos casos, especificar a distribuição nula da estatística LRT é complicado e, para contornar esse problema empregam-se métodos numéricos baseados em simulação, ou seja, pode-se usar o método "bootstrap" paramétrico. De forma resumida, primeiro geram-se B ($B = 10.000$) amostras "bootstrap" $y^1, \ldots, y^B$, por reamostragem de $\hat{f}_0(y)$ (em que $\hat{f}_0$ denota o modelo ajustado sob a hipótese) com parâmetros estimados do conjunto original de dados. Em seguida, calculam-se os valores da estatística

$LRT^* = \{lrt^1, \ldots, lrt^B\}$ que fornecem uma distribuição empírica para LRT_{obs}. Então, pode-se obter um valor de p ("p-value"), numericamente, conforme a expressão apresentada por Davison e Hinkley (1997)

$$p = \frac{n_{extreme} + 1}{B + 1}, \quad \text{com} \quad n_{extreme} = \sum_{k=1}^{B} I(lrt^k \geq LRT_{obs}),$$

em que $I(x)$ é uma função indicadora, valendo 1 se x for verdadeiro e 0, caso contrário.

Seleção de modelos - inferência sobre efeitos fixos

Frequentemente, na avaliação de experimentos planejados ou ainda em estudos observacionais, o interesse está em verificar se a variável resposta é afetada por covariáveis, contínuas ou categóricas, ou seja, quais covariáveis devem entrar no preditor linear. Para avaliar a significância dos termos no preditor linear, nos MLGs usa-se a análise de "deviance", proposta por Nelder e Wedderburn (1972), que é uma generalização da ANOVA padrão. Dados dois modelos aninhados, com p e q, $(p < q)$, parâmetros e "deviances" residuais D_p e D_q, respectivamente, se ϕ é conhecido, tem-se que

$$\frac{D_p - D_q}{q - p} \sim \phi \chi^2_{q-p}$$

que nada mais é do que um teste de razão de verossimilhanças.

Se ϕ é desconhecido, deve-se obter uma estimativa $\hat{\phi}$ consistente, de preferência usando o modelo maximal, e a inferência pode ser baseada na estatística F expressa por $F = \frac{(D_p - D_q)/(q-p)}{\hat{\phi}} \sim F_{q-p,n-m}$, sendo m o número de parâmetros no modelo maximal e n o número de observações, como ocorre para o modelo quase-Poisson.

Para o modelo binomial negativo, um problema surge pela estimação do parâmetro de superdispersão. Há duas estratégias que podem ser adotadas: (i) reestimar o parâmetro de superdispersão para cada submodelo, ou (ii) fixar o valor da estimativa do parâmetro de superdispersão, usando o modelo maximal (Venables e Ripley, 2002; Demétrio et al., 2014). A estratégia (i) corresponde à estimação padrão por máxima verossimilhança de diferentes submodelos que são comparados usando o teste de razão de verossimilhanças. Entretanto, o valor da estimativa do parâmetro de dispersão pode aumentar por absorver diferenças entre tratamentos. A estratégia (ii) está mais de acordo com o que é feito nos modelos clássicos de regressão e usado, também, para o modelo quase-Poisson.

Para os MLGM, depois de escolher os termos aleatórios do modelo, a seleção do preditor linear, em geral, envolve comparações de modelos aninhados e diferenças de "deviances" ("Analysis of deviance"), isto é, testes de razão de verossimilhanças. Ela envolve avaliar o valor da função de verossimilhança para o modelo completo e para o

modelo sob H_0 (modelo reduzido), usando o método de máxima verossimilhança (MV)

$$\begin{aligned} LRT &= -2[\text{logLik(modelo reduzido) - logLik(modelo completo)}] \\ &= \text{deviance(modelo reduzido) - deviance(modelo completo)}, \end{aligned}$$

em que logLik é o logaritmo da função de verossimilhança. Os modelos aninhados e o modelo de referência têm o mesmo número de parâmetros de covariância e diferentes conjuntos de parâmetros de efeito fixo. Assintoticamente, $LRT \sim \chi^2_\nu$, sendo ν a diferença em número de parâmetros de efeito fixo dos dois modelos.

Outro critério bastante empregado é o critério de informação de Akaike (AIC) (Akaike, 1974), que permite a comparação de múltiplos modelos aninhados ou não (Bolker et al., 2009). O AIC foi obtido, usando-se a seguinte fórmula:

$$AIC = -2\log(L) + 2k,$$

em que $\log(L)$ é o logaritmo da função de verossimilhança para o modelo e k é o número de parâmetros do modelo. De acordo com esse critério o melhor modelo é aquele que tem o menor valor de AIC (Akaike, 1974).

Para auxiliar na seleção de um modelo pode-se usar a função `dropterm()` do pacote **MASS** (Venables e Ripley, 2002), que considera cada variável, individualmente, verificando qual será a mudança no AIC se essa variável for excluída. Para se acelerar o ajuste dos diferentes modelos estatísticos por remoção ou adição de variáveis pode-se usar a função de atualização `update()`. Pode-se, também, especificar qual distribuição será usada, χ^2 ou F.

Valores preditos

Em estudos de caráter prático, os pesquisadores têm interesse em fazer comparações múltiplas entre os tratamentos no caso de variáveis qualitativas ou obter intervalos de confiança. Para isso há necessidade de se obterem preditores lineares das médias marginais de tratamentos e erros-padrão.

Para os MLGs, tem-se que as estimativas dos preditores lineares são dadas por $\hat{\eta}$ com os respectivos erros-padrão $s(\hat{\eta})$. Intervalos de confiança assintóticos para os preditores lineares podem ser obtidos por

$$\hat{\eta} \pm z * (\hat{\eta}),$$

e re-transformados para a escala da variável resposta $\hat{\mu} = g^{-1}(\hat{\eta})$.

A partir dos preditores lineares podem ser obtidas as médias para os tratamentos e intervalos de confiança na escala do preditor linear e re-transformados para a escala da variável resposta. Todavia, as estimativas de máxima verossimilhança podem ser tendenciosas, principalmente, em pequenas amostras (Thorson e Kristensen, 2016) e, dependendo do caso, há necessidade de se fazerem correções para ajuste dos vieses.

A função `emmeans()` do pacote **emmeans** (Lenth, 2022) permite a obtenção de estimativas das médias marginais (ou médias por mínimos quadrados) para a variável resposta, considerando combinações de fatores para modelos lineares, modelos lineares generalizados, modelos lineares generalizados mistos etc. Além disso, intervalos de confiança também podem ser obtidos.

A função `emmeans()`, também, permite fazer a correção de vieses com base em uma aproximação de Taylor de segunda ordem. Suponha que U é uma transformação não linear de uma variável aleatória Y. Para fazer retro-transformação, considera-se $Y = h(U)$; como a $\mathrm{E}(Y)$ não é simples de ser obtida, pode-se obter uma aproximação por meio da expansão da série de Taylor de segunda ordem

$$\mathrm{E}(Y) \approx h(\eta) + 0,5h''(\eta)\mathrm{Var}(U) \Rightarrow \mathrm{E}(Y) \approx h(\eta) + 0,5h''(\eta)\sigma^2, \tag{2.8}$$

em que h é a transformação inversa, h'' derivada de segunda ordem dessa função, η é a parte fixa do preditor linear e $\sigma = \sqrt{\mathrm{Var}(U)}$. Para os MLGM, $U = \eta + Z$, sendo Z um efeito aleatório ou a soma de vários efeitos aleatórios.

2.4 Estudo de caso: ajuste e avaliação dos modelos para a variável resposta OT

Considera-se que Y_{jklr} representa a variável resposta OT de um estudo observacional descrito na Seção 2.2, com valores y_{jklr} obtidos no j-ésimo ($j = P1, P2$) período para uma fêmea doadora no k-ésimo ($k = D, H, M$) status, no l-ésimo ($l = 2wks, 3wks, na$) intervalo da OPU realizada na mesma doadora e r-ésima ($r = 0, 1, 5$) injeção de FSH. A seguir, mostra-se como ajustar e selecionar modelos para analisar adequadamente esses dados, usando-se o software `R` (R Core Team, 2022).

Modelo Poisson

Assume-se, inicialmente, o processo mais simples que os oócitos são obtidos independemente, aleatoriamente e com taxa constante (Hinde e Demétrio, 1998; Demétrio et al., 2014), sendo assumida a distribuição padrão de Poisson, isto é, $Y_{jklr} \sim \mathrm{P}(\mu_{jklr})$ com função de ligação logarítmica. As variáveis explicativas referentes ao período do ano e às doadoras entram no preditor linear

$$\eta_{jklr} = \beta_0 + \tau_j + \delta_k + \gamma_{jk} + \varphi_l + \alpha_r, \tag{2.9}$$

em que β_0 é a constante de efeito fixo, τ_j é efeito fixo do j-ésimo período do ano, δ_k é o efeito fixo do k-ésimo status da fêmea doadora, γ_{jk} é o efeito fixo da interação entre os fatores período do ano e status da fêmea doadora, φ_l é o efeito fixo do l-ésimo intervalo da OPU realizada na mesma doadora, α_r é o efeito fixo da r-ésima injeção de FSH.

Ajustando-se o modelo Poisson log-linear com preditor linear dado por (2.9) aos dados de OT (código `R` apresentado no Quadro 1) obtém-se "deviance" residual igual a

5627,072 e a estatística $X^2 = 6052{,}912$, com 1138 graus de liberdade, evidenciando a falta de ajuste do modelo o que é confirmado pelo gráfico semi-normal de probabilidades com envelope de simulação apresentado na Figura 2.2(a). Isso mostra que há mais variabilidade ($\hat{\phi} > 1$) que o modelo Poisson pode acomodar, uma evidência clara de superdispersão. A seguir, serão usados modelos alternativos que levam em conta a superdispersão presente nos dados (Demétrio et al., 2014).

Quadro 1. Código R para ajuste do modelo Poisson com o preditor linear (2.9).

```
# Examinando as primeiras 6 linhas do conjunto de dados
head(oocytes)
  Period Interval    Breed Injections status Donor OT
1     P2       na Holstein          5      M   780 16
2     P2     3wks Holstein          5      H  1375  9
3     P2       na Holstein          5      H 37181  9
4     P2     3wks Holstein          5      M 50505 13
5     P2       na Holstein          5      M 62439  9
6     P2       na Holstein          5      M 63886 12
# Modelo Poisson (model_P) usando Period, status, Interval e Injections
# como fator,
# OT com o preditor linear (2.9)
model_P <- glm(OT ~ Period+status+Period:status+Interval+Injections,
                family = poisson(link = "log"), data = oocytes)
# Deviance residual
summary(model_P)$deviance
# Estatística X2
(X2 <- sum(residuals(model_P, type="pearson")^2))
# Verificando a adequação do modelo com a função hnp
library(hnp)
hnp(model_P, conf = 0.95, print.on = TRUE)
```

Modelo quase-Poisson

A maneira mais simples de levar a variabilidade extra em consideração é assumir que ela é constante e independente do número de oócitos obtidos, substituindo a função de variância da Poisson pela forma mais geral (2.2). Ajustando-se um modelo quase-Poisson com função de ligação logarítmica e preditor linear de equação (2.9) aos dados de OT, o valor estimado de ϕ é $\hat{\phi} = 5{,}32$ $(6052{,}912/1138 = X^2/df)$. Esse ajuste é, facilmente, obtido em `R`, usando a família `quasipoisson` em `glm`, como mostrado no Quadro 2. Não há um teste de verificação de ajuste do modelo, pois o resíduo foi usado para estimar ϕ. Entretanto, o gráfico semi-normal de probabilidades com envelope de simulação pode ser

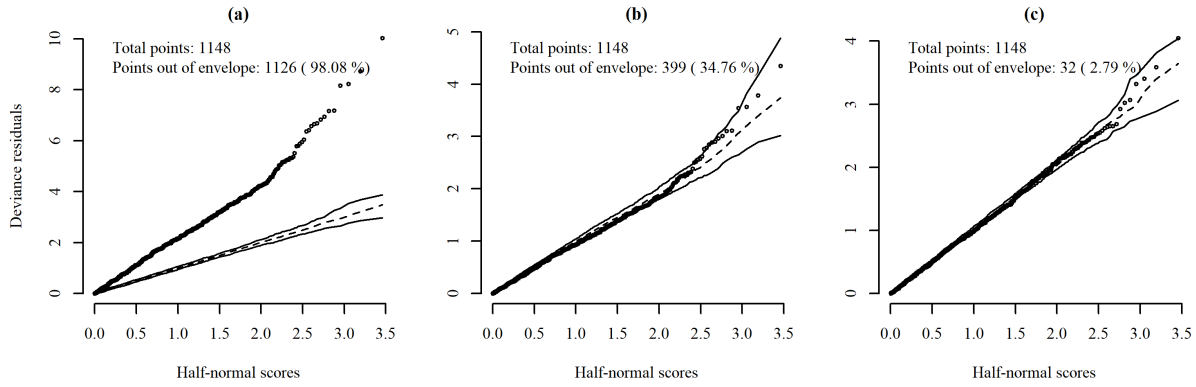

Figura 2.2. Variável resposta OT - Gráficos semi-normal para os modelos: (a) Poisson, (b) quase-Poisson e (c) binomial negativo, com o preditor linear de equação (2.9).

usado levando em conta a variação extra assumida em (2.2). O gráfico apresentado na Figura 2.2(b) mostra evidência forte de falta de ajuste do modelo, com mais de 30% dos resíduos observados fora do envelope de simulação.

Quadro 2. Código R para ajuste do modelo quase-Poisson com o preditor linear (2.9).

```
# Modelo quase-Poisson (model_QP) com o preditor linear (2.9)
model_QP <- glm(OT ~ Period+status+Period*status+Interval+Injections,
                            family =quasipoisson(link="log"), data =oocytes)
# Verificando a adequação do modelo com a função hnp
hnp(model_QP, conf = 0.95, print.on = TRUE)
# Obtendo o parâmetro de dispersão
summary(model_QP)$dispersion
```

Modelo binomial negativo

Assumindo que os oócitos são obtidos com taxas variáveis (por exemplo, diferenças em fertilidade das vacas doadoras), trazendo variabilidade extra para as contagens observadas, o que pode ser levado em conta pelo uso de um modelo em dois estágios como o binomial negativo. Pode-se ajustar um modelo binomial negativo com função de ligação logarítmica e preditor linear de equação (2.9) aos dados de OT (código `R` apresentado no Quadro 3), usando-se a função `glm.nb()` do pacote **MASS** (Venables e Ripley, 2002) ou da função `aodml()` do pacote **aods3** (Lesnoff e Lancelot, 2013). O valor estimado de θ é $\hat{\theta} = 4,3596$ (e.p. $= 0,2284$) o que indica forte superdispersão nos dados.

O gráfico semi-normal de probabilidades com envelope de simulação (Figura 2.2(c)) evidencia que o modelo binomial negativo ajusta-se bem aos dados.

Quadro 3. Código R para ajuste do modelo binomial negativo com o preditor linear (2.9).

```
library(MASS)
# Modelo binomial negativo (model_NB) usando Period, status, Interval e
# Injections  como fator
# Com o preditor linear (2.9)
 model_NB <- glm.nb(OT~ Period+status+Period*status+Interval+Injections,
                                                  data = oocytes)
# Estimativa de teta e seu  erro padrão
(theta<-c(theta = model_NB$theta, SE = model_NB$SE.theta))
# Verificação da adequação do modelo com a função hnp
set.seed(2144)
hnp(model_NB , conf = 0.95, print.on = TRUE)
-------------------------------------------------------------------------------
                        Parte 2: Análise de deviance

# Análise de deviance, usando o modelo binomial negativo com theta fixado
anova(model_NB,test = "Chisq")

# Modelo sem a covariável  interval
model_NB1 <- glm(OT ~ Period+Status+Period*Status+Injections,
                   family =negative.binomial(4.3596), data =oocytes)
```

Após reajustar o modelo binomial negativo, fixando-se $\hat{\theta} = 4,3596$, obtém-se a tabela de análise de "deviance" (Tabela 2.1). Usando-se o teste da razão de verossimilhanças, verifica-se que existem evidências de não significância para a variável "Interval" e significância das demais variáveis, ao nível de 5% de significância, evidenciando que os intervalos entre as OPUs não influenciam na obtenção de oócitos. Portanto o preditor linear de equação (2.9) pode ser reduzido para

$$\eta_{jkr} = \beta_0 + \tau_j + \delta_k + \gamma_{jk} + \alpha_r. \tag{2.10}$$

Modelo Poisson-normal

Uma forma alternativa de levar em conta a superdispersão é supor que há várias fontes não explicadas que afetam a obtenção de oócitos que podem ser representadas pela adição de um efeito aleatório em nível de observação, $Z_{jklr} \sim N(0, \sigma^2_O)$, no preditor linear (Fatoretto et al., 2018; Demétrio et al., 2020)

$$\eta_{jklr} = \beta_0 + \tau_j + \delta_k + \gamma_{jk} + \varphi_l + \alpha_r + Z_{jklr}. \tag{2.11}$$

A inclusão do efeito aleatório em nível de observação tem sido indicada em muitos trabalhos como uma forma de controlar a heterogeneidade advinda de um elemento de

Tabela 2.1. Análise de "deviance" para os modelos Poisson, quase-Poisson e binomial negativo com o preditor linear (2.9) e Poisson-normal e Combinado com o preditor linear (2.13).

Fontes de variação	GL	Poisson	Quase-Poisson	Binomial negativo
Period	1	79,3600 (<0,0001)	79,3600 (<0,0001)	15,872 (<0,0001)
Status	2	947,1800 (<0,0001)	947,1800 (0,6514)	188,406 (<0,0001)
Interval	2	4,5600 (0,1022)	4,5600 (0,6514)	2,287 (0,3187)
Injections	2	33,1000 (<0,0001)	33,1000 (0,0449)	6,597 (0,0369))
Period:Status	2	44,2300 (<0,0001)	44,2300 (0,01588)	7,603 (0,0223)
$-2\log(L)$	—	10830,6000	—	8161,5390
AIC	—	10851,0000	—	8183,5000
Fontes de variação	GL	Poisson-normal	Combinado	—
Period	1	0,09250 (0,7610)	0,09190 (0,7618)	—
Status	2	50,9935 (<0,0001)	50,7893 (<0,0001)	—
Interval	2	1,0681 (0,5862)	1,0584 (0,5891)	—
Injections	2	1,6598 (0,4360)	1,6473 (0,4388)	—
Period:Status	2	9,9966 (0,0067)	9,9240 (0,0069)	—
$-2\log(L)$	—	7595,1280	7595,0910	—
AIC	—	7619,1280	7621,0910	—

GL - graus de liberdade; $\log(L)$ é o logaritmo da função de verossimilhança e valores dentro dos parenteses são os valor de p com base na estatística χ^2 para os modelos Poisson e binomial negativo, Poisson-normal e Combinado e F para o modelo quase-Poisson.

confundimento, que foi omitido por ser desconhecido ou por não ter sido avaliado, por exemplo, no caso da produção *in vitro* de embriões, uma simples mudança no estado metabólico/hormonal da doadora causa heterogeneidade em nível de observação (Silva et al., 2017).

Outra maneira, para acomodar a hierarquia dos dados, pois as doadoras foram aspiradas em mais de uma ocasião, é adicionar ao preditor linear um efeito aleatório em nível de doadora, $\xi_{d(jklr)} \sim N(0, \sigma_d^2)$ (Molenberghs et al., 2007, 2010; Garcia et al., 2020)

$$\eta_{jklr} = \beta_0 + \tau_j + \delta_k + \gamma_{jk} + \varphi_l + \alpha_r + \xi_{d(jklr)},\ d = 1, \ldots, 318. \tag{2.12}$$

Alternativamente, podem-se incluir ambos os efeitos aleatórios no mesmo preditor linear

$$\eta_{jklr} = \beta_0 + \tau_j + \delta_k + \gamma_{jk} + \varphi_l + \alpha_r + \xi_{d(jklr)} + Z_{jklr},\ d = 1, \ldots, 318. \tag{2.13}$$

Esses modelos são exemplos de modelo Poisson log-linear que é um modelo linear generalizado misto conforme apresentado na equação (2.4). Para o ajuste desses modelos, pode-se usar a função `glmer()` do pacote **lme4** (Bates et al., 2015) com código `R` apresentado no Quadro 4.

Quadro 4. Código R para ajuste do modelo Poisson-normal com os preditores lineares dados nas equações (2.11), (2.12) e (2.13) para a variável resposta OT.

```
library(lme4)
# Modelo Poisson-normal (model_PN1_id) com o preditor linear (2.11)

id <- factor(1:nrow(oocytes))
model_PN1_id <- glmer(OT~ Period+status+Period*status+Interval+Injections+
(1|id),family=poisson, data=oocytes, control=glmerControl(optimizer = "bobyqa",
optCtrl=list(maxfun=600000)))

# Modelo Poisson-normal (model_PN1) com o preditor linear (2.12)
model_PN1 <- glmer(OT~ Period+status+Period*status+Interval+Injections+
(1|Donor),family=poisson, data=oocytes,
control=glmerControl(optimizer = "bobyqa", optCtrl=list(maxfun=600000)))

# Modelo Poisson-normal (model_PN2) com o preditor linear (2.13)
model_PN2 <- glmer(OT~ Period+status+Period*status+Interval+Injections+
(1|Donor)+ (1|id), family=poisson, data=oocytes,
control=glmerControl(optimizer = "bobyqa",
optCtrl=list(maxfun=600000)))

# Verificando a adequação do modelo com a função hnp
hnp(model_PN1, conf = 0.95, print = TRUE)
hnp(model_PN1_id, conf = 0.95, print = TRUE)
hnp(model_PN2, conf = 0.95, print = TRUE)

# AIC dos modelos
AIC(model_PN1_id)
[1] 8177.476
AIC(model_PN1)
[1] 7798.079
AIC(model_PN2)
[1] 7619.128
```

Ajustando-se o modelo Poisson-normal com os preditores lineares de equações (2.11), (2.12) e (2.13), o gráfico semi-normal de probabilidades com envelopes de simulação evidencia que o modelo com efeito em nível de observação (Figura 2.3(a)) se ajusta aos dados de OT, enquanto que o modelo com efeito aleatório em nível de doadora (Figura 2.3(b)) não se ajusta. Por outro lado, o modelo com ambos os efeitos aleatórios (Figura 2.3(c)) também se ajusta aos dados de OT, com menor valor de AIC (= 7619,128).

Verifica-se que, à medida que novos efeitos aleatórios são incluídos no preditor linear, a convergência torna-se mais demorada podendo em alguns casos não ocorrer. Uma

das maneiras de minimizar os problemas numéricos é por meio da utilização do método de otimização `bobyqa` que obtém o mínimo de funções de muitas variáveis, empregando métodos de região de confiança e formando modelos quadráticos por interpolação (Powell, 2009). Para o exemplo em estudo, definiu-se `maxfun` = 600000 como o número máximo de avaliações da função, pois permitiu obter melhores resultados quanto à convergência.

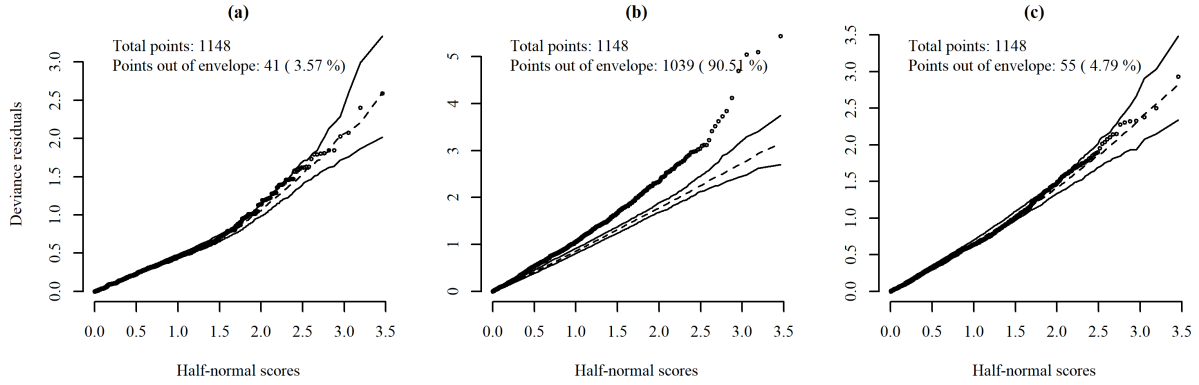

Figura 2.3. Variável resposta OT - Gráficos semi-normais para os modelos: (a) Poisson-normal com o preditor linear (2.11), (b) Poisson-normal com o preditor linear (2.12) e (c) Poisson-normal com o preditor linear (2.13).

A fim de verificar se o modelo Poisson-normal com preditor linear de equação (2.13) pode ser simplificado, a seguir são feitos testes sobre os efeitos aleatórios σ_d^2 e σ_O^2. Note que esses testes estão no limite do espaço paramétrico e que, portanto, a estatística dos testes tem distribuição dada pela mistura $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 = \frac{1}{2}\chi_1^2$.

Para testar a hipótese $H_0 : \sigma_O^2 = 0$ versus $H_0 : \sigma_O^2 > 0$, o teste da razão de verossimilhanças é feito pela comparação dos modelos `model_PN1` versus `model_PN2`, com preditores lineares de equações (2.11) e (2.13). Tem-se que $LRT = -2 \times (-3888,039 + 3797,564) = 180,9512$ com 1 grau de liberdade. Com base na distribuição nula do LRT, obtida numericamente via "bootstrap" paramétrico (código R apresentado no Quadro 5), obtém-se o valor $p < 0,0001$, sugerindo que há fortes evidências para se rejeitar H_0.

Quadro 5. Código R para "bootstrap" paramétrico.

```
# Valor da estatística (LRT_obs)
LRT_obs <- as.numeric(-2*(logLik(model_PN1)-logLik(model_PN2)))
valor_p <- pchisq(LRT_obs,df=1,lower=FALSE)
data.frame(LRT_obs, valor_p)
# Bootstrap paramétrico
y <- simulate(model_PN1) #Modelo nulo
B=10000 #Amostras bootstrap
lrstat <- numeric(B)

set.seed(0821)
```

```
for(i in 1:B){
y <- unlist(simulate(model_PN1))
null_model<-glmer(y~Period+status+Period*status+Interval+Injections+
(1|Donor),family=poisson, data=oocytes, control=glmerControl(optimizer = "bobyqa",
optCtrl=list(maxfun=600000)))

alterna_model<- glmer(y~Period+status+Period*status+Interval+Injections+
(1|Donor)+(1|id),family=poisson, data=oocytes,
control=glmerControl(optimizer = "bobyqa",
optCtrl=list(maxfun=600000)))
lrstat[i]<-round(as.numeric(-2*(logLik(null_model)-logLik(alterna_model))),4)
print(paste0("Amostra bootstrap:",i))
}
# Cálculo direto do valor de p
(p <- (sum(lrstat>=LRT_obs)+1)/(B+1))
```

Para testar a hipótese $H_0 : \sigma_d^2 = 0$ versus $H_0 : \sigma_d^2 > 0$ o teste da razão de verossimilhanças é feito pela comparação dos modelos `model_PN1_id` versus `model_PN2` com preditores lineares de equações (2.12) e (2.13). Tem-se que $LRT = -2\times(-4077,738+3797,564) = 560,3485$ com 1 grau de liberdade. Com base na distribuição nula do LRT, obtida numericamente via "bootstrap" paramétrico, obtém-se o valor de $p < 0,0001$, sugerindo que há fortes evidências para se rejeitar H_0.

Entretanto, é necessário testar ambos as componentes, $H_0 : \sigma_d^2 = \sigma_O^2 = 0$, no modelo normal de Poisson. O teste de razão de verossimilhança é realizado comparando `model_P` versus `model_PN2` com preditores de equação linear (2.9) e (2.13). Neste caso a distribuição da mistura é dada por $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2 = \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2$. Temos $LRT = -2\times(-5415,301 + 3797,564) = 3235,474$ com 2 graus de liberdade com um valor $p < 0,0001$, o que sugere que há fortes evidências para rejeitar H_0.

Após a seleção da parte aleatória do modelo Poisson-normal, a próxima questão de interesse é se a parte fixa do preditor linear de equação (2.13) (modelo `model_PN2`) pode ser reduzida. Usando-se, as funções `dropterm()` e `update()`, conforme indicado no Quadro 6, há evidência de não significância das covariáveis "Injections" e "Interval", resultando no modelo (`model_PN4`) com preditor linear

$$\eta_{jk} = \beta_0 + \tau_j + \delta_k + \gamma_{jk} + \xi_{d(jk)} + Z_{jk}, \quad d = 1, \ldots, 318. \tag{2.14}$$

Tem-se que AIC = 7613,832 é menor do que o valor de AIC (7619,128) para o modelo Poisson-normal com preditor linear de equação (2.13). O teste da razão de verossimilhanças entre estes dois modelos $LRT = -2 \times (-3798,9 + 3797,6) = 2,60$ com 4 graus de liberdade é não significativo (valor $p = 0,6085$), indicando por parcimônia a escolha do modelo Poisson-normal com preditor linear de equação (2.14). O gráfico meio-normal de probabilidades com envelope de simulação confirma que existem evidências de bom ajuste (Figura 2.4) desse modelo.

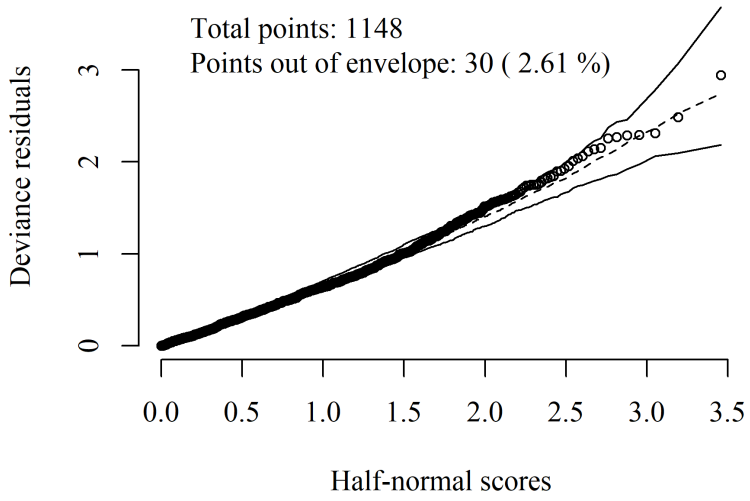

Figura 2.4. Variável resposta OT - Gráficos semi-normal para o modelo Poisson-normal com o preditor linear (2.14).

Quadro 6. Código R para seleção de variáveis.

```
dropterm(model_PN2,test = "Chisq")
Single term deletions
Model:
OT ~ Period+status+Period*status+Interval+Injections +
(1|Donor)+(1|id)
              Df    AIC    LRT  Pr(Chi)
<none>           7619.1
Interval       2 7616.2 1.0592 0.588847
Injections     2 7616.8 1.6443 0.439480
Period:status  2 7625.0 9.8866 0.007131 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
model_PN3 <-update(model_PN2,.~.-Interval)
dropterm(model_OT_PN3,test = "Chisq")
model_PN4 <-update(model_PN3,.~.-Injections)
dropterm(model_PN4,test = "Chisq")
model_PN5 <-update(model_PN4,.~.-Period:status)
AIC(model_PN2)
[1] 7619.128
AIC(model_PN3)
[1] 7616.187
AIC(model_PN4)
[1] 7613.832
AIC(model_PN5)
[1] 7618.81
```

Modelo combinado

Assumindo agora que a superdispersão pode ser causada, simultaneamente, por variabilidade da média e por hierarquia, dois conjuntos de efeitos aleatórios podem ser adicionados ao modelo, θ_{ij} seguindo uma distribuição gama e $\boldsymbol{b}_i$ tendo distribuição normal, dando origem ao modelo Poisson-gama-normal ou modelo combinado, como resumido em (2.5).

De forma semelhante ao modelo Poisson-Normal, ajustou-se o MC (modelo binomial negativo-normal) com os efeitos aleatórios em nível de observação ($Z_{jklr} \sim N(0, \sigma_O^2)$) e/ou de fêmea doadora ($\xi_{d(jklr)} \sim N(0, \sigma_d^2)$, de acordo com os preditores lineares de equações (2.11), (2.12) e (2.13), usando-se a função `glmer.nb()` do pacote **lme4** (Quadro 7).

Quadro 7. Código R para ajustar o MC com os preditores lineares dados nas equações (2.11), (2.12) e (2.13) para a variável resposta OT.

```
library(lme4)
# Modelo combinado (model_CMid) com o preditor linear (2.11)
id <- factor(1:nrow(oocytes))
model_CMid <- glmer.nb(OT~ Period+Status+Period*Status+Interval+Injections+
                           (1|id),data=oocytes,
control=glmerControl(optimizer = 'bobyqa', optCtrl=list(maxfun=600000)))
summary(model_CMid)

# Modelo combinado (model_CM1)
# Com o preditor linear (2.12)
model_CM1 <- glmer.nb(OT~ Period+status+Period*status+Interval+Injections+
(1|Donor),data=oocytes,
control=glmerControl(optimizer = 'bobyqa',optCtrl=list(maxfun=600000)))
summary(model_CM1)

# Parâmetro teta (binomial negativo)
getME(model_CM1, "glmer.nb.theta")

# Modelo combinado (model_CM1)
# Com o preditor linear dado em  (2.13)
model_CM2 <- glmer.nb(OT~ Period+status+Period*status+Interval+Injections+
(1|Donor)+(1|id),data=oocytes,
control=glmerControl(optimizer = 'bobyqa',optCtrl=list(maxfun=600000)))
summary(model_CM2)

# Parâmetro teta (binomial negativo)
getME(model_CM2, "glmer.nb.theta")

# Verificação da adequação do modelo com a função hnp
```

```
# A implementação de uma nova classe de modelo fornecendo três funções para hnp.
# Variável-resosta:OT
resp<-oocytes$OT

dfun <- function(obj) resid(obj,type="deviance")
sfun <- function(n, obj) simulate(obj)[[1]]
ffunid <- function(resp)glmer.nb(resp~ Period+Status+Period*Status+Interval+
Injections+(1|id),data=oocytes,
control=glmerControl(optimizer= 'bobyqa',optCtrl=list(maxfun=600000)))
ffun1 <- function(resp)glmer.nb(resp~ Period+status+Period*status+Interval+
Injections+(1|Donor), data=oocytes,
control=glmerControl(optimizer = 'bobyqa',optCtrl=list(maxfun=600000)))
ffun2 <- function(resp)glmer.nb(resp~ Period+status+Period*status+Interval+
Injections+(1|Donor)+(1|id), data=oocytes,
control=glmerControl(optimizer = 'bobyqa',optCtrl=list(maxfun=600000)))

hnp(model_CMid, conf = 0.95,newclass = TRUE,
diagfun = dfun, simfun = sfun, fitfun = ffunid,print = TRUE)
hnp(model_CM1, conf = 0.95,newclass = TRUE,
diagfun = dfun, simfun = sfun, fitfun = ffun1,print = TRUE)
hnp(model_CM2, conf = 0.95,newclass = TRUE,
diagfun = dfun, simfun = sfun, fitfun = ffun2,print = TRUE)
```

O gráfico meio-normal de probabilidades com envelope de simulação (Figura 2.5(b)) evidencia que o MC com o preditor linear de equação (2.12), isto é, somente com efeito aleatório normal de doadora não se ajusta. Entretanto, quando se adiciona, somente o efeito aleatório normal de indivíduo (equação (2.11)) ou ambos efeitos aleatórios no mesmo preditor linear (equação (2.13)), há evidência de bom ajuste (Figura 2.5(a) e 2.5(c)).

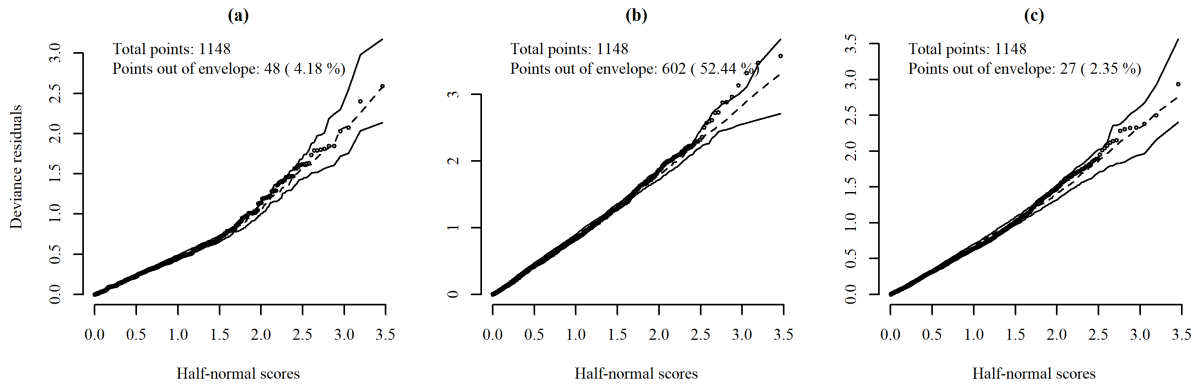

Figura 2.5. Variável resposta OT - Gráficos semi-normais para os modelos: (a) MC com o preditor linear (2.11), (b) MC com o preditor linear (2.12) e (c) MC com o preditor linear (2.13).

A fim de verificar se o modelo binomial negativo-normal com preditor linear de

equação (2.13) pode ser simplificado, a seguir são feitos testes sobre os efeitos aleatórios σ_d^2 e σ_O^2. Note que esses testes estão no limite do espaço paramétrico e que, portanto, a estatística dos testes tem distribuição dada pela mistura $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 = \frac{1}{2}\chi_1^2$.

Para testar a hipótese $H_0 : \sigma_O^2 = 0$ versus $H_0 : \sigma_O^2 > 0$, o teste da razão de verossimilhanças é feito pela comparação dos modelos `model_CMid` versus `model_CM2` com preditores lineares de equações (2.11) e (2.13). Tem-se que $LRT = -2 \times (-4077,70 + 3797,50) = 560,40$ com 1 grau de liberdade. Com base na distribuição nula do LRT, obtida numericamente via "bootstrap" paramétrico, obtém-se o valor $p < 0,0001$, sugerindo que há fortes evidências para se rejeitar H_0.

Para testar a hipótese $H_0 : \sigma_d^2 = 0$ versus $H_0 : \sigma_d^2 > 0$, o teste da razão de verossimilhança é feito pela comparação dos modelos `model_CM1` versus `model_CM2` com preditores lineares de equações (2.12) e (2.13). Tem-se que $LRT = -2 \times (-3897,1 + 3797,50) = 199,20$ com 1 grau de liberdade. Com base na distribuição nula do LRT, obtida numericamente via "bootstrap" paramétrico, obtém-se o valor $p < 0,0001$, sugerindo que há fortes evidências para se rejeitar H_0.

Testando ambos os componentes $H_0 : \sigma_d^2 = \sigma_O^2 = 0$, o teste de razão de verossimilhança é realizado comparando `model_NB` versus `model_CM2` com preditores lineares (2.9) e (2.13). Temos $LRT = -2 \times (-4080,77 + 3797,546) = 566,4481$ com 2 graus de liberdade com valor $p < 0,0001$, sugerindo que há fortes evidências para rejeitar H_0.

Portanto, há necessidade ambos os efeitos aleatórios no modelo binomial negativo-normal, que possui o menor valor de AIC. O valor estimado para o parâmetro θ é $\hat{\theta} = 3462,1190$, indicando superdispersão fraca. Os valores estimados para σ_d^2 e σ_O^2 são, respectivamente, $\hat{\sigma}_d^2 = 0,2058$ e $\hat{\sigma}_O^2 = 0,0438$, muito semelhantes às obtidas para o modelo Poisson-normal com ambos os efeitos aleatórios.

A obtenção do MC mais parcimonioso em termos dos efeitos fixos foi feito utilizando as funções `dropterm()` e `update()`, considerando o modelo `model_CM2` com o mesmo procedimento apresentando no Quadro 6 e na mesma sequência de exclusão das covariáveis, obtendo o mesmo preditor do modelo Poisson-normal com preditor linear de equação (2.14).

Seleção do modelo e cálculo de médias marginais

Pelos resultados obtidos, verificou-se que os modelos Poisson e quase-Poisson com preditor linear de equação (2.9) não se ajustaram enquanto que o modelo binomial negativo com preditor linear de equação (2.9) e os modelos Poisson-normal e combinado com preditor linear de equação (2.13) ajustaram-se aos dados de OT. Pela análise de "deviance" apresentada na Tabela 2.1, vê-se que os níveis de significância e os valores de p dos testes para os fatores mudam, dependendo do modelo usado. Ao nível de 5% de significância, a interação "Period:Status" foi significativa e o fator "Interval" não significativo para todos os modelos, enquanto o fator "Status" foi não significativo apenas para

o modelo quase-Poisson e o fator “Injections” foi não significativo apenas para os modelos Poisson-normal e combinado.

Usando-se, testes de razão de verossimilhança, pôde-se fazer uma redução do preditor linear e os três modelos que melhor se ajustaram aos dados de OT foram o binomial negativo com preditor linear de equação (2.10), o modelo Poisson-normal com preditor linear de equação (2.14) e o MC com preditor linear de equação (2.14). Na Tabela 2.2, são apresentadas as estimativas dos parâmetros com seus erros-padrão e os valores das estatísticas $-2\log(L)$ e AIC. O modelo com menor valor de AIC é o Poisson-normal com preditor linear de equação (2.14).

Tabela 2.2. Estimativas dos parâmetros com seus respectivos erros-padrão (e.p.) para os modelos, Poisson (eq. 2.9), binomial negativo (eq. 2.10), Poisson-normal e Combinado (eq. 2.14) para a variável resposta OT.

	Poisson		Binomial negativo	
Parâmetros†	Estimativa (e.p.)		Estimativa (e.p.)	
Intercept	3,2894 (0,0339)	***	3,2873 (0,0732)	***
Period - P2	−0,0268 (0,0253)		−0,0222 (0,0609)	
status - H	−0,5936 (0,0259)	***	−0,5947 (0,0574)	***
status - M	−0,4013 (0,0354)	***	−0,4037 (0,0792)	***
Period - P2:status - H	0,1884 (0,0334)	***	0,1722 (0,0730)	*
Period - P2:status - M	0,2451 (0,0443)	***	0,2360 (0,0995)	*
Interval - 3wks	−0,0453 (0,0247)		—	
Interval - na	0,0136 (0,0174)		—	
Injections1	−0,1638 (0,0252)	***	−0,1554 (0,0575)	**
Injections5	−0,1691 (0,0275)	***	−0,1662 (0,0626)	**
Parâmetro-Superdispersão	—		4,3596 (0,2280)	
$-2\log(L)$	10830,6000		8164,2510	
AIC	10851,0000		8180,2510	
	Poisson-normal		Combinado	
Parâmetros†	Estimativa (e.p.)		Estimativa (e.p.)	
Intercept	2,9933 (0,0555)	***	2,9934 (0,0554)	***
Period - P2	−0,0012 (0,0355)		−0,0017 (0,0356)	
status - H	−0,4990 (0,0713)	***	−0,4990 (0,0712)	***
status - M	−0,3721 (0,0886)	***	−0,3721 (0,0887)	***
Period - P2:status - H	0,0987 (0,0534)		0,0987 (0,0534)	
Period - P2:status - M	0,2190 (0,0774)	**	0,2190 (0,0776)	**
Parâmetro-Superdispersão	—		3462,1190	
Observation (σ_O^2)	0,0442		0,0439	
Donor (σ_d^2)	0,2072		0,2071	
$-2\log(L)$	7597,832		7597,7960	
AIC	7613,8320		7615,7960	

†Efeitos dos parâmetros são calculados usando a codificação padrão para contrastes, o primeiro valor é usado como referência; significância (***valor $p < 0,0001$, **valor $p < 0,01$ e *valor $p < 0,05$) e $\log(L)$ é o logaritmo da função de verossimilhança.

O teste da razão de verossimilhanças para a hipótese H_0 : modelo Poisson-normal ($\theta \to \infty$) versus H_a : modelo combinado é um teste no limite do espaço paramétrico com distribuição dada pela mistura $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 = \frac{1}{2}\chi_1^2$ (Lawless, 1987). Tem valor $LRT = -2 \times (-3798,916 + 3798,898) = 0,03541$ com 1 grau de liberdade e valor $p = 0,4254$, sugerindo que o modelo Poisson-normal com preditor linear de equação (2.14) é o mais

parcimonioso, sendo, portanto, o escolhido e, como já visto pelo gráfico da Figura 2.4, ajusta-se bem aos dados.

Como a interação "Period:Status" foi significativa, calculam-se as médias marginais estimadas de OT para as combinações dos níveis de "Period" e "Status", usando a retro-transformação definida em (2.8). Nesse caso, os efeitos aleatórios independentes de doadoras e de indivíduos no preditor linear (2.14) devem ser levados em consideração para estimar as médias marginais. Portanto, para o uso da função `emmeans()`, calcula-se $\hat{\sigma}^2 = \hat{\sigma}_d^2 + \hat{\sigma}_O^2$. Adicionalmente, calcularam-se os intervalos de confiança simultâneos, empregando a desigualdade de Bonferroni (Miller, 1981), com nível de confiança conjunto de 84%, para um nível de significância de, aproximadamente, 5% (Goldstein e Healy, 1995; Payton et al., 2003). O código `R` para a obtenção das estimativas das médias marginais e dos respectivos intervalos de confiança, é apresentado no Quadro 8.

Quadro 8. Código R para calcular as médias marginais e intervalos de confiança.

```
# Efeito combinado dos componentes de variância
 (total.SD<-sqrt(0.0442+0.2072))
marginal.means<- emmeans(model_PN4,~ Period*status, bias.adjust = TRUE,
sigma = total.SD,  type="response", level=0.84, adjust = "bonferroni")
marginal.means
Period status rate    SE  df asymp.LCL asymp.UCL
P1     D      22.5 1.245 Inf      19.9      25.4
P2     D      22.4 1.137 Inf      20.0      25.1
P1     H      13.6 0.636 Inf      12.3      15.1
P2     H      15.0 0.646 Inf      13.7      16.5
P1     M      15.5 1.131 Inf      13.2      18.2
P2     M      19.2 1.163 Inf      16.8      22.0
```

Na Figura 2.6, têm-se as estimativas das médias marginais e dos respectivos intervalos de confiança. Claramente, no período de temperaturas mais elevadas (P1), as doadoras de status "vacas secas" (D) apresentam a maior média para o número de OT por OPU, diferindo estatisticamente das doadoras de status "novilhas" (H) e "vacas em lactação" (M), enquanto que H e M não diferem estatisticamente. No período de temperaturas mais amenas (P2) as doadoras D embora também com maior média diferem de H mas não diferem de M. De maneira geral, a obtenção de OT é favorecida em temperaturas menos elevadas, especialmente em doadoras de status H e M.

2.5 Considerações finais e conclusões

Neste trabalho, é apresentada uma revisão de modelos que podem ser usados para analisar dados na forma de contagens e levar em conta os diversos aspectos de falhas das suposições do modelo Poisson, devido à superdispersão e/ou ocorrência de

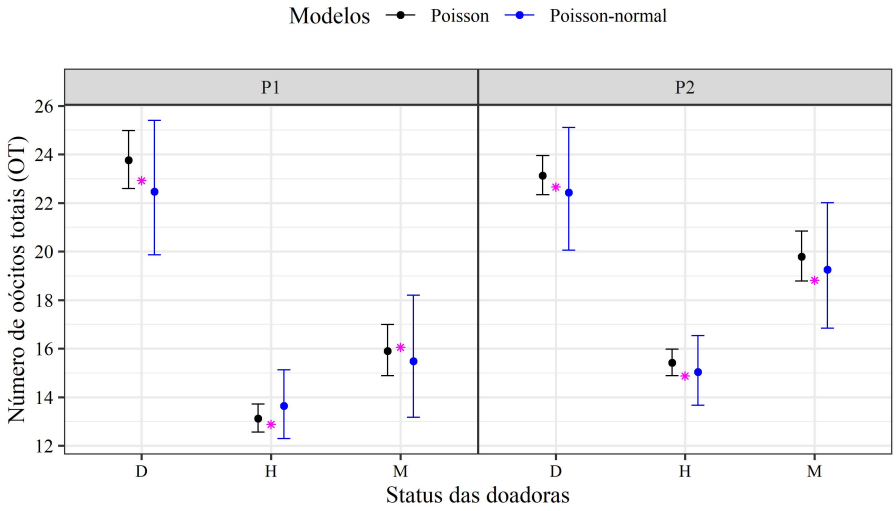

Figura 2.6. Variável resposta OT - Intervalos de confiança simultâneos para médias marginais para as combinações dos níveis de períodos do ano (P1: Junho a Outubro e P2: sete meses restantes) e status das doadoras (H: novilhas, M: vacas em lactação e D: vacas secas), usando a desigualdade de Bonferroni com nível de confiança conjunto de 84%, considerando os modelos Poisson (−) com preditor linear de equação (2.9) e Poisson-normal (−) com preditor linear de equação (2.14), as médias observadas são representadas pelo símbolo (∗).

correlação causada por uma estrutura hierárquica. Um tutorial, utilizando o software `R`, mostra como ajustar os modelos, usando como ilustração a análise de dados referentes à quantidade de oócitos de um estudo observacional. Destaca-se, também, o uso do gráfico meio-normal de probabilidades com envelope de simulação para verificação de ajuste dos modelos. Apresentam-se, ainda, outras técnicas para avaliação e comparação dos modelos bem como seleção de efeitos fixos e aleatórios, como: teste LRT, simulação "bootstrap" paramétrica, para seleção de efeitos aleatórios e o critério de informação AIC, para seleção de efeitos fixos em modelos MLGM e MC.

Os testes das hipóteses sobre os componentes de variância ($H_0 : \sigma_{ii} = 0$ versus $H_a : \sigma_{ii} > 0$) estão no limite do espaço paramétrico e LRT tem distribuição que é uma mistura de χ^2's (Zhang e Lin, 2008). Todavia, para os casos mais complexos, a distribuição nula da estatística LRT é complicada ou desconhecida. Uma saída viável é a utilização de métodos numéricos como o "bootstrap" paramétrico, embora, aumente o custo computacional (Bates et al., 2015). Neste trabalho, assumindo suposições assintóticas do teste LRT ou métodos de simulação para testar a significância dos componentes de variância, as conclusões foram as mesmas, sugerindo fortes evidências contra H_0, conclusões similares às obtidas por Sakamoto (2019).

Após a seleção do modelo, foi proposta a utilização de intervalos de confiança simultâneos, aplicando a desigualdade de Bonferroni com nível de confiança conjunto de 84% (Goldstein e Healy, 1995; Payton et al., 2003) para verificar a diferença estatística entre as médias. A desigualdade de Bonferroni distribui a taxa fixada α para os m intervalos, α/m, fornecendo inferências simultâneas prevenindo a ocorrência do erro tipo I

(Miller, 1981). A avaliação visual da significância estatística fornece uma maneira simples e eficaz de compreender os resultados estatísticos (Noguchi e Marmolejo-Ramos, 2016).

Conclusões de caráter prático sob a ótica dos dados de contagem de oócitos são apresentadas. Ressalta-se, ainda, que não existe uma regra geral para analisar dados desse tipo, uma vez que existem muitos fatores envolvidos para cada variável, desde características ambientais até aquelas de cunho genético.

2.5.1 Implicações estatísticas

As observações da variável resposta OT foram analisadas, usando-se diversos modelos, sendo que há evidências de bom ajuste para os modelos binomial negativo, Poisson-normal e binomial negativo-normal, conforme gráficos meio-normais de probabilidades com envelopes de simulação (menos de 5% dos pontos ficaram fora dos envelopes de simulação). Entretanto, o modelo Poisson-normal incluindo efeitos aleatórios em nível de observação e doadoras foi o que melhor se ajustou, além de ter interpretação prática. A inclusão de efeito aleatório de doadoras modela a correlação existente entre observações repetidas no mesmo animal, uma vez que oócitos provenientes da mesma doadora são geneticamente semelhantes. A inclusão de efeito aleatório em nível de observação leva em conta heterogeneidade entre as observações devido a uma combinação de fontes de variação não explicadas.

Conforme mostrado, o modelo Poisson com preditor linear de equação (2.9) não se ajustou aos dados de OT e se usado indica a significância inadequada do efeito do fator "Injection", o que poderia elevar os custos de produção na PIVE. Além disso, os intervalos de confiança simultâneos obtidos (Figura 2.6) são muito menores do que aqueles produzidos usando o modelo Poisson-normal com preditor linear de equação (2.14), levando a conclusões erradas.

2.5.2 Implicações práticas

Pelos resultados obtidos, o estresse induzido por temperaturas mais elevadas reduz a quantidade de oócitos produzidos o que pode refletir diretamente na quantidade e qualidade dos embriões. O estresse térmico afetou de maneira diferente as doadoras, tendo efeito mais pronunciado em animais de status H e M, que apresentaram as maiores reduções no número de OT, quando comparadas com vacas de status D.

Quando os animais estão fora da zona de conforto térmico, os desequilíbrios fisiológicos são potencializados e a produtividade diminui (Souza-Cácares et al., 2019). O efeito deletério do estresse térmico interfere na maturação dos oócitos degradando as proteínas de choque térmico (HSPs), que representam as respostas primárias de proteção celular (Castro et al., 2013). Em situações adversas, as concentrações de espécies reativas de oxigênio são aumentadas e, em contrapartida, os níveis de HSP também se elevam

numa tentativa de sintetizar novas proteínas. Quando esse mecanismo de defesa falha, o número de oócitos que atingem a metáfase II diminui e assim a competência oocitária é reduzida (Castro et al., 2013; Pöhland et al., 2020).

O presente estudo, também, evidencia que os intervalos entre aspirações e o uso de protocolos hormonal FSH não alterou o número de OT produzidos. Nessa mesma linha, alguns autores relataram que o número de sessões de OPUs não tem influência sobre a média de oócitos obtidos por doadora (Pontes et al., 2011; Gimenes et al., 2015). Já em relação ao protocolo hormonal, é importante ressaltar que o tratamento hormonal melhora a qualidade dos oócitos recuperados na OPU, mas não aumenta significativamente a quantidade total de oócitos. Os principais benefícios dos protocolos hormonais é a melhora na competência oocitária, aumentando o número de oócitos viáveis e o números de embriões que são produzidos (Demétrio et al., 2020; Ongaratto et al., 2020).

Diante dos resultados, se o produtor/laboratório estiver simplesmente interessado em aumentar a média de embriões produzidos, usar as doadoras de status D seria a melhor escolha, pois produzem o maior número de oócitos, o que aumenta as chances de sucesso de obter um bezerro no final do processo. Entretanto, sabe-se que deve haver consideração do mérito genético de quais vacas coletar oócitos para produção *in vitro* de embriões. Por fim, em relação às implicações práticas, ressalta-se que as conclusões obtidas nesse artigo são restritivas às variáveis de contagem, sendo necessárias outras variáveis para obter conclusões mais amplas como: a taxa de oócitos grau I e II/número total de oócitos viáveis, taxa de clivagem e embriões/número de oócitos cultivados *in vitro*.

Referências

Agresti, A. (2019). *An introduction to categorical data analysis*. John Wiley & Sons, Hoboken, NJ, 3 edition.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Atkinson, A. C. (1987). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. OXFORD UNIV PR.

Bates, D., Mächler, M., Bolker, B., e Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1).

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., e White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3):127–135.

Castro, S. V., Lobo, C. H., Figueiredo, J. R., e Rodrigues, A. P. R. (2013). Proteínas de choque térmico hsp 70: estrutura e atuação em resposta ao estresse celular. *Acta Veterinaria Brasilica*, 7(4):261–271.

Cavalieri, F. L. B., Morotti, F., Seneda, M. M., Colombo, A. H. B., Andreazzi, M. A., Emanuelli, I. P., e Rigolon, L. P. (2018). Improvement of bovine in vitro embryo production by ovarian follicular wave synchronization prior to ovum pick-up. *Theriogenology*, 117:57–60.

Chebel, R. C., Demétrio, D. G. B., e Metzger, J. (2008). Factors affecting success of embryo collection and transfer in large dairy herds. *Theriogenology*, 69(1):98–106.

Davison, A. C. e Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

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.

Demétrio, D. G. B., Benedetti, E., Demétrio, C. G. B., Fonseca, J., Oliveira, M., Magalhaes, A., e Santos, R. M. (2020). How can we improve embryo production and pregnancy outcomes of holstein embryos produced in vitro? (12 years of practical results at a california dairy farm). *Animal Reproduction*, 17(3).

Fatoretto, M. B., de Andrade Moral, R., Demétrio, C. G. B., de Pádua, C. S., Menarin, V., Rojas, V. M. A., D'Alessandro, C. P., e Delalibera, I. (2018). Overdispersed fungus germination data: statistical analysis using R. *Biocontrol Science and Technology*, 28(11):1034–1053.

Garcia, S. M., Morotti, F., Cavalieri, F. L. B., Lunardelli, P. A., de Oliveira Santos, A., Membrive, C. M. B., Castilho, C., Puelker, R. Z., Silva, J. O. F., Zangirolamo, A. F., e Seneda, M. M. (2020). Synchronization of stage of follicle development before OPU improves embryo production in cows with large antral follicle counts. *Animal Reproduction Science*, 221:106601.

Gimenes, L. U., Ferraz, M. L., Fantinato-Neto, P., Chiaratti, M. R., Mesquita, L. G., Filho, M. F. S., Meirelles, F. V., Trinca, L. A., Rennó, F. P., Watanabe, Y. F., e Baruselli, P. S. (2015). The interval between the emergence of pharmacologically synchronized ovarian follicular waves and ovum pickup does not significantly affect in vitro embryo production in *Bos indicus*, *Bos taurus*, and *Bubalus bubalis*. *Theriogenology*, 83(3):385–393.

Goldstein, H. e Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(1):175.

Hinde, J. (1982). Compound poisson regression models. In *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*, pages 109–121. Springer New York.

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

Hoef, J. M. V. e Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772.

Iddi, S. e Molenberghs, G. (2012). A combined overdispersed and marginalized multilevel model. *Computational Statistics & Data Analysis*, 56(6):1944–1951.

Jørgensen, B. (2006). Generalized linear models. *Encyclopedia of environmetrics*, 3.

Kosma, M., Studnicki, M., Wójcik-Seliga, J., Michalska-Klimczak, B., Wyszyński, Z., e Wójcik-Gront, E. (2019). Over-dispersed count data in crop and agronomy research. *Journal of Agronomy and Crop Science*, 205(4):414–421.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 209–225.

Lee, Y., Nelder, J. A., e Pawitan, Y. (2017). *Generalized linear models with random effects: unified analysis via H-likelihood*, volume 153. CRC Press.

Lenth, R. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.1-1.

Lesnoff, M. e Lancelot, R. (2013). Aods3: Analysis of overdispersed data using s3 methods. aods3 package version 0.4-1.

McCullagh, P. e Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, Boca Raton London New York.

Miller, R. G. (1981). *Simultaneous Statistical Inference*. Springer New York.

Molenberghs, G., Verbeke, G., e Demétrio, C. G. B. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, 13(4):513–531.

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

Moral, R. A., Hinde, J., e Demétrio, C. G. B. (2017). Half-normal plots and overdispersed models in R: The hnp package. *Journal of Statistical Software*, 81(10).

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

Noguchi, K. e Marmolejo-Ramos, F. (2016). Assessing equality of means using the overlap of range-preserving confidence intervals. *The American Statistician*, 70(4):325–334.

Ongaratto, F. L., Cedeño, A. V., Rodriguez-Villamil, P., Tríbulo, A., e Bó, G. A. (2020). Effect of FSH treatment on cumulus oocyte complex recovery by ovum pick up and in vitro embryo production in beef donor cows. *Animal Reproduction Science*, 214:106274.

Payton, M. E., Greenstone, M. H., e Schenker, N. (2003). Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *Journal of Insect Science*, 3(1).

Pontes, J. H. F., Sterza, F. A. M., Basso, A. C., Ferreira, C. R., Sanches, B. V., Rubin, K. C. P., e Seneda, M. M. (2011). Ovum pick up, in vitro embryo production, and pregnancy rates from a large-scale commercial program using nelore cattle (*Bos indicus*) donors. *Theriogenology*, 75(9):1640–1646.

Powell, M. J. D. (2009). The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, pages 26–46.

Pöhland, R., Souza-Cácares, M. B., Datta, T. K., Vanselow, J., Martins, M. I. M., da Silva, W. A. L., Cardoso, C. J. T., e de Andrade Melo-Sterza, F. (2020). Influence of long-term thermal stress on the in vitro maturation on embryo development and heat shock protein abundance in zebu cattle. *Animal Reproduction*, 17(3).

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sakamoto, W. (2019). Inference on variance components near boundary in linear mixed effect models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(6).

SAS Institute Inc (2015). *SAS/STAT® 14.1 user's guide*. Cary, NC: SAS Institute Inc.

Silva, J. C. B., Ferreira, R. M., Filho, M. M., de Rezende Naves, J., Santin, T., Pugliesi, G., e Madureira, E. H. (2017). Use of FSH in two different regimens for ovarian superstimulation prior to ovum pick up and in vitro embryo production in holstein cows. *Theriogenology*, 90:65–73.

Sirard, M.-A. (2018). 40 years of bovine ivf in the new genomic selection context. *Reproduction*, 156(1):R1–R7.

Souza-Cácares, M. B., Fialho, A. L. L., Silva, W. A. L., Cardoso, C. J. T., Pöhland, R., Martins, M. I. M., e Melo-Sterza, F. A. (2019). Oocyte quality and heat shock proteins in oocytes from bovine breeds adapted to the tropics under different conditions of environmental thermal stress. *Theriogenology*, 130:103–110.

Stroup, W. W. (2015). Rethinking the analysis of non-normal data in plant and soil science. *Agronomy Journal*, 107(2):811–827.

Thorson, J. T. e Kristensen, K. (2016). Implementing a generic method for bias correction in statistical models using random effects, with spatial and population dynamics examples. *Fisheries Research*, 175:66–74.

Venables, W. N. e Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer-Verlag GmbH.

Wang, J., Li, J., Wang, F., Xiao, J., Wang, Y., Yang, H., Li, S., e Cao, Z. (2020). Heat stress on calves and heifers: a review. *Journal of Animal Science and Biotechnology*, 11(1).

Zhang, D. e Lin, X. (2008). Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In *Random Effect and Latent Variable Model Selection*, pages 19–36. Springer New York.

3 COMBTMB: UM PACOTE R PARA AJUSTE DE MODELOS A DADOS LONGITUDINAIS E SUPERDISPERSOS

Resumo

A classe dos modelos combinados (MC) pode ser usada para modelar superdispersão e correlação em dados não normais medidos no contexto longitudinal ou hierárquico. O MC tem como característica o uso de dois conjuntos de efeitos aleatórios para capturar a superdispersão e correlação. Por exemplo, no contexto de dados de contagem, o MC reduz-se, naturalmente, ao modelo Poisson-normal, uma instância do modelo linear generalizado misto, e, também, reduz-se ao modelo binomial negativo na ausência de correlação. Apesar da flexibilidade de modelagem oferecida, os MCs ainda não possuem uma ferramenta computacional padronizada no software de código-aberto `R`. Em função disso foi desenvolvido o pacote **combTMB**, que fornece a classe dos MCs no sistema `R` para computação estatística. O método de estimação dos parâmetros é por máxima verossimilhança, usando como ferramenta o "Template Model Builder" (**TMB**). A teoria e a estrutura de modelagem são brevemente delineadas, além de apresentar alguns detalhes de implementação e suas principais características. As funcionalidades do pacote são ilustradas com duas aplicações em dados na forma de contagens e de proporções. Em geral, as características mais atraentes são a combinação de velocidade e uma interface flexível e familiar aos usuários de `glm()` e **lme4**.

Palavras-chave: Modelos lineares generalizados; Superdispersão; Modelos mistos; Dados longitudinais; `R`.

3.1 Introdução

Na pesquisa científica, dados não normais são frequentemente mensurados em contextos longitudinais ou hierárquicos em áreas como ciências agrárias, medicina, ciências sociais, economia e outras. Exemplos comuns desses tipos de variáveis resposta incluem dados de contagem, binomiais e binários que, geralmente, são analisados usando modelos lineares generalizados (MLG) (Nelder e Wedderburn, 1972; McCullagh e Nelder, 1989; Agresti, 2019), uma estrutura unificadora baseada na família exponencial.

A classe dos MLGs apresenta como característica chave a relação *média-variância*, sendo que a variância é uma função determinística da média (Molenberghs et al., 2007, 2010). Por exemplo, para resultados Bernoulli com probabilidade de sucesso $\mu = \pi$, a variância é $V(\mu) = \pi(1 - \pi)$, e para contagens seguindo o modelo Poisson, a relação é $V(\mu) = \mu$. No entanto, em situações práticas, essa relação é muito restritiva, uma vez que a variância observada pode ser superior ou inferior ao que é prescrito pelo modelo, levando à chamada superdispersão e subdispersão, respectivamente (Molenberghs et al.,

2017; Agresti, 2019). Em geral, a superdispersão é o fenômeno predominante, sendo comumente encontrada em dados na forma de contagens ou proporções, mas também pode ocorrer em dados binários desde que as hierarquias estejam presentes (Molenberghs et al., 2012; Demétrio et al., 2014).

Associadas ao fenômeno de superdispersão, as hierarquias impostas pelas medidas repetidas ou longitudinais resultam em dados correlacionados, uma vez que as medições de diferentes atributos são realizadas no mesmo indivíduo, local ou período de tempo, introduzindo algum grau de associação. Dessa forma, ambos, superdispersão e correlação, podem ocorrer simultaneamente. A utilização de modelos que considerem somente a superdispersão, por exemplo, o modelo binomial negativo para dados de contagem e o modelo beta-binomial para dados na forma de proporções (Hinde e Demétrio, 1998; Demétrio et al., 2014), implica em independência entre medidas repetidas, o que não é realista. Por outro lado, os modelos lineares generalizados mistos (MLGM), propostos por Breslow e Clayton (1993), acomodam as correlações, mas capturam apenas parcialmente a superdispersão, comprometendo as conclusões (Molenberghs et al., 2007; Kassahun et al., 2012; Iddi e Molenberghs, 2012).

Molenberghs et al. (2007) e Molenberghs et al. (2010) desenvolveram modelos combinados (MC), uma estrutura geral e flexível que permite acomodar a superdispersão e a correlação em dados não normais medidos longitudinalmente, por meio da introdução de efeitos aleatórios tanto no preditor linear quanto na média da distribuição. Ao longo dos anos, essa metodologia tem se mostrado eficiente quando ambas as características estão presentes, apresentando resultados mais robustos do que as metodologias anteriores (Molenberghs et al., 2017).

Todavia, passados mais de 15 anos, não existe um software padrão que implemente os MCs e suas extensões, a exemplo do pacote `R` **lme4** para MLGMs (Bates et al., 2015). Basicamente, a maioria das aplicações que utilizaram a metodologia dos MCs empregaram o procedimento NLMIXED do SAS (SAS Institute, 2014), que é limitado a poucos efeitos aleatórios e de difícil acesso para a maioria dos pesquisadores de áreas aplicadas. Portanto, para preencher essa lacuna, apresenta-se o pacote `R` **combTMB**, que implementa os MCs e, além disso, permite o ajuste dos tradicionais MLGs e MLGMs, que poderá facilitar a seleção de modelos.

O restante do artigo está estruturado como se segue. A Seção 3.2 apresenta uma revisão breve dos métodos padrões e suas principais extensões para a análise de dados com superdispersão. Na Seção 3.3, é descrita a formulação dos MCs com seus casos especiais e modelagem multinível marginalizada e processo de estimação. As Seções 3.6 e 3.7 são dedicadas à discussão de algumas questões sobre a implementação do pacote **combTMB** em `R` e à apresentação de suas funcionalidades, analisando dados de contagem e proporção. As Seções 3.8 e 3.9 tratam do desempenho computacional e de considerações gerais sobre o pacote **combTMB**.

3.2 Modelagem padrão

3.2.1 Modelos lineares generalizados

Um MLG é definido por três componentes básicos (Nelder e Wedderburn, 1972). O primeiro, chamado de componente aleatório, é representado pelas variáveis aleatórias independentes, $Y_1, \ldots, Y_n$, todas com função densidade ou de probabilidade $f(y_i; \theta_i, \phi)$ pertencentes à família exponencial na forma canônica dada por:

$$f(y_i; \theta_i, \phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\}, i = 1, 2, \ldots, n, \tag{3.1}$$

em que $b(\cdot)$ e $c(\cdot)$ são funções conhecidas, $\phi > 0$ é o parâmetro de dispersão e θ_i é o parâmetro canônico. A média e variância são, respectivamente, $\mathrm{E}(Y_i) = \mu_i = b'(\theta_i)$ e $\mathrm{Var}(Y_i) = \phi b''(\theta_i) = \phi V_i$, sendo que $V_i = V(\mu_i) = d\mu_i/d\theta_i$ é a função de variância. A função de variância na família exponencial exerce um papel central, isto é, caracteriza a distribuição. Assim, por exemplo, a função de $V(\mu_i) = \mu_i$, $\mu_i > 0$, caracteriza a classe de distribuições Poisson com média μ_i.

O segundo componente, chamado sistemático, é representado pelo preditor linear $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$, sendo $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})^T$ a matriz do modelo, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ o vetor de parâmetros desconhecidos.

Finalmente, o terceiro componente, é a função de ligação $g(\cdot)$, monotônica e diferenciável, que relaciona a média da distribuição ao preditor linear, isto é, $\eta_i = g(\mu_i)$.

Após o estabelecimento do trinômio: distribuição da variável resposta, matriz do modelo ($\boldsymbol{X}$) e função de ligação, o próximo passo é a estimação do vetor de parâmetros $\boldsymbol{\beta}$ para o MLG. Para isso, métodos como a máxima verossimilhança ou quase verossimilhança podem ser utilizados (Molenberghs et al., 2012).

3.2.2 Superdispersão

Um dos principais problemas que surgem ao analisar dados não gaussianos, como dados binomiais, binários agregados e dados de contagem, é a superdispersão. Isso pode ocorrer quando a variância observada da variável resposta é maior do que a variância esperada pelo modelo probabilístico adotado (McCullagh e Nelder, 1989; Hinde e Demétrio, 1998). As causas potenciais para ocorrência de superdispersão incluem má especificação do preditor linear, excesso de zeros nas observações, variabilidade do material experimental, correlação entre as respostas individuais, amostragem por conglomerado e outras causas subjacentes (Demétrio et al., 2014).

Os problemas referentes à ocorrência de superdispersão podem ser contornados por diferentes estratégias, mas particularmente, duas se destacam (Hinde e Demétrio, 1998; Demétrio et al., 2014). Para a primeira, considera-se uma forma mais geral para a função de variância com adição de um parâmetro extra $\phi \neq 1$, permitindo maior

flexibilidade para a variância, $\text{Var}(Y_i) = \phi V(\mu_i)$, ou seja, uma abordagem de quase-verossimilhança. Para dados na forma de proporções, assume-se $\text{Var}(Y_i) = \phi n_i \pi_i (1 - \pi_i)$, enquanto que para contagens, assume-se $\text{Var}(Y_i) = \phi \mu_i$.

Para a segunda forma de incorporar a superdispersão no modelo considera-se o modelo em dois estágios, isto é, admite-se que o parâmetro do modelo para a resposta tem uma distribuição, levando a um modelo composto. Para observações na forma de proporções, inicialmente, pode-se assumir que Y_i condicional a S_i tem distribuição binomial, isto é, $Y_i|S_i \sim \text{Binomial}\,(n_i, S_i)$ e que S_i é uma variável aleatória com $\text{E}(S_i) = \pi_i$ e $\text{Var}(S_i) = \phi \pi_i (1 - \pi_i)$. Usando propriedades de esperança condicional, segue-se que

$$\text{E}(Y_i) \;=\; \text{E}[\text{E}(Y_i|S_i)] = n_i \pi_i,$$

e

$$\begin{aligned} \text{Var}(Y_i) &= \text{E}[(\text{Var}(Y_i|S_i)] + \text{Var}[(\text{E}(Y_i|S_i)] \\ &= n_i \pi_i (1 - \pi_i)[1 + \phi(n_i - 1)]. \end{aligned} \tag{3.2}$$

Especificamente, se S_i segue uma distribuição $\text{Beta}(\alpha_{1i}, \alpha_{2i})$ com $\alpha_{1i} + \alpha_{2i}$ constante, então, marginalmente, Y_i tem uma distribuição beta-binomial, em que $\pi = \alpha_{1i}/(\alpha_{1i} + \alpha_{2i})$ e $\phi = 1/(\alpha_{1i} + \alpha_{2i} + 1)$. A ocorrência de superdispersão se dá quando $\phi > 0$, sendo que para $\phi = 0$ a variância da distribuição binomial é restabelecida. Além disso, quando $n_i = 1$, o modelo Bernoulli é obtido. É importante destacar que os modelos de superdispersão para dados univariados de Bernoulli são irrelevantes, sendo aplicáveis apenas quando existem hierarquias nos dados ou quando os dados binários são agregados em dados binomiais (Molenberghs et al., 2012).

Analogamente, para dados de contagem, supõe-se que uma variável Y_i condicional a T_i possui distribuição Poisson, ou seja, $Y_i|T_i \sim \text{Poisson}(T_i)$. Considerando que T_i é uma variável aleatória com distribuição gama, isto é, $T_i \sim \text{gama}(\alpha_1, \alpha_{2i})$, a distribuição marginal de Y_i é binomial negativa, com média e variância dadas por

$$\text{E}(Y_i) \;=\; \alpha_1/\alpha_{2i} = \mu_i$$

e

$$\text{Var}(Y_i) = \mu_i + \mu_i^2/\alpha_1 = \mu_i(1 + \mu_i/\alpha_1), \tag{3.3}$$

permitindo incorporar variabilidade maior do que a média. Para α_1 conhecido, a distribuição binomial negativa pertence à família exponencial no formato (3.1) (Demétrio et al., 2014), e o algoritmo de estimação dos MLGs pode ser utilizado (Venables e Ripley, 2002; Demétrio et al., 2014).

Outro tipo de modelo em dois estágios resulta pela inclusão de efeitos aleatórios no preditor linear η_i do MLG. Esses efeitos aleatórios são, geralmente, assumidos como provenientes de uma distribuição normal, obtendo assim a classe dos modelos lineares generalizados mistos (Breslow e Clayton, 1993).

3.2.3 Modelos lineares generalizados mistos

A classe dos modelos lineares generalizados mistos (MLGM) possui aplicações importantes na análise de dados discretos, especialmente quando as observações exibem alguma forma de dependência e/ou superdispersão (Breslow e Clayton, 1993; Stroup, 2016). Além dos efeitos fixos no preditor linear, ele acomoda variabilidade extra por meio da adição de efeitos aleatórios. Pode ser escrito em uma estrutura hierárquica que é especialmente útil para a análise de dados provenientes de medições repetidas e estudos longitudinais.

Seja Y_{ij} a variável aleatória que representa a j-ésima medida longitudinal no i-ésimo indivíduo, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. Seja $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in_i})^T$, o vetor das n_i medidas feitas no mesmo indivíduo. Assumindo-se que, condicionalmente a um vetor q-dimensional de efeitos aleatórios $\boldsymbol{b}_i$ com distribuição $N_q(\boldsymbol{0}, \boldsymbol{D})$, em que $\boldsymbol{0}$ é um vetor de zeros e $\boldsymbol{D}$ a matriz de variâncias e covariâncias, as variáveis respostas Y_{ij} são consideradas independentes e têm distribuição que pertence à família exponencial

$$f(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \phi) = \exp\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij}, \phi)\},$$

com

$$g(\mu_{ij}) = g[\mathrm{E}(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta})] = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i$$

para uma função de ligação conhecida $g(\cdot)$, com os vetores $\boldsymbol{x}_{ij}$ e $\boldsymbol{z}_{ij}$ de dimensões p e q, associados às covariáveis e aos efeitos aleatórios, respectivamente, $\boldsymbol{\beta}$ é o vetor de dimensão p de parâmetros de efeitos fixos e ϕ é o parâmetro de dispersão. Denota-se, também, por $f(\boldsymbol{b}_i|\boldsymbol{D})$ a função densidade dos efeitos aleatórios $\boldsymbol{b}_i$.

3.3 Modelos combinados: efeitos aleatórios conjugados e normais

Superdispersão (Seção 3.2.2) e correlação entre observações (Seção 3.2.3) podem ocorrer simultaneamente e ser incorporadas nos chamados modelos combinados (MC) (Molenberghs et al., 2007, 2010). Dois tipos de efeitos aleatórios são usados, um para o parâmetro do modelo, considerado aleatório, que tem uma distribuição conjugada à distribuição condicional da variável resposta (Lee et al., 2017) e o outro, em geral com distribuição normal, adicionado ao preditor linear.

Supõe-se, inicialmente, que, condicional aos efeitos aleatórios independentes $\boldsymbol{b}_i$ e θ_{ij}, a variável aleatória Y_{ij} tem distribuição pertencente à família exponencial

$$f_i(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \theta_{ij}, \phi) = \exp\{\phi^{-1}[y_{ij}\lambda_{ij} - \psi(\lambda_{ij})] + c(y_{ij}, \phi)\},$$

com média dada por

$$\mathrm{E}(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \theta_{ij}) = \mu_{ij}^c = \theta_{ij}k_{ij}.$$

Assume-se que $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{in_i})^T$ tem distribuição com vetor de médias $\mathrm{E}(\boldsymbol{\theta}_i) = \boldsymbol{\Phi}_i$ e matriz de variâncias e covariâncias $\mathrm{Var}(\boldsymbol{\theta}_i) = \boldsymbol{\Sigma}_i$, isto é, $\mathrm{E}(\theta_{ij}) = \phi_{ij}$; $\mathrm{Var}(\theta_{ij}) = \sigma_{i,jj}$ e $\mathrm{Cov}(\theta_{ij}, \theta_{ik}) = \sigma_{i,jk}$. Além disso, $k_{ij} = g(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)$ e supõe-se que $\boldsymbol{b}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{D})$.

3.3.1 Casos específicos: para dados de contagem e binários

Para dados de contagens, Molenberghs et al. (2007) assumem que

$$\begin{aligned} Y_{ij}|\boldsymbol{b}_i, \theta_{ij} &\sim \mathrm{Poisson}(\lambda_{ij}), && (3.4)\\ \lambda_{ij} &= \theta_{ij}k_{ij} = \theta_{ij}\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i), && (3.5)\\ \boldsymbol{b}_i &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{D}), && (3.6)\\ \mathrm{E}(\boldsymbol{\theta}_i) &= \boldsymbol{\Phi}_i, && (3.7)\\ \mathrm{Var}(\boldsymbol{\theta}_i) &= \boldsymbol{\Sigma}_i. && (3.8) \end{aligned}$$

Marginalmente, o vetor de médias de $\boldsymbol{Y}_i$ tem elementos

$$\begin{aligned} \mathrm{E}(Y_{ij}) &= \mathrm{E}\{\mathrm{E}[\mathrm{E}(Y_{ij}|\theta_{ij}, \boldsymbol{b_i})]\}\\ &= \phi_{ij}\exp\left(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij}\right) = \mu_{ij} \qquad (3.9) \end{aligned}$$

enquanto que a matriz de variâncias-covariâncias é dada por

$$\begin{aligned} \mathrm{Var}(\boldsymbol{Y}_i) &= \mathrm{E}\{\mathrm{E}[\mathrm{Var}(\boldsymbol{Y}_i|\theta_{ij}, \boldsymbol{b}_i)]\} + \mathrm{E}\{\mathrm{Var}[\mathrm{E}(\boldsymbol{Y}_i|\theta_{ij}, \boldsymbol{b}_i)]\} + \mathrm{Var}\{\mathrm{E}[\mathrm{E}(\boldsymbol{Y}_i|\theta_{ij}, \boldsymbol{b}_i)]\}\\ &= \boldsymbol{M}_i + \boldsymbol{M}_i[\boldsymbol{P}_i - \boldsymbol{J}_{n_i}]\boldsymbol{M}_i, \end{aligned}$$

em que $\boldsymbol{M}_i$ é uma matriz diagonal com $\boldsymbol{\mu}_i$ ao longo da diagonal e $\boldsymbol{P}_i$ é uma matriz com elementos

$$p_{i,jk} = \exp\left(\frac{1}{2}\boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ik}\right)\frac{\sigma_{i,jk} + \phi_{ij}\phi_{ik}}{\phi_{ij}\phi_{ik}}\exp\left(\frac{1}{2}\boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ik}\right).$$

Se θ_{ij} tiver distribuição gama, resulta o modelo Poisson-gama-normal que tem como casos especiais os modelos binomial negativo e Poisson-normal, multivariados e univariados, bem como o modelo Poisson padrão.

De forma semelhante, para dados de binários, tem-se que

$$\begin{aligned} Y_{ij}|\boldsymbol{b}_i &\sim \mathrm{Bernoulli}(\pi_{ij}), && (3.10)\\ \pi_{ij} &= \theta_{ij}k_{ij} = \theta_{ij}\frac{\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)}{1 + \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)}, && (3.11) \end{aligned}$$

a especificação é complementada com (3.6) - (3.8). Ao contrário do caso Poisson, não existem formas fechadas para média e variância quando os efeitos aleatórios normais estão presentes e a função de ligação logística é usada.

Segundo Molenberghs et al. (2010), pode-se assumir que os efeitos aleatórios de superdispersão $\theta_{ij} = \theta_i$ obtendo a distribuição beta-binomial na ausência de efeitos

aleatórios normais. Ao considerar explicitamente que $\theta_{ij} \sim \text{Beta}(\alpha_1, \alpha_2)$, então $\phi_{ij} = \text{E}(\theta_{ij}) = \alpha_1/(\alpha_1 + \alpha_2)$, e as variâncias σ^2_{ij} e covariâncias $\sigma_{i,jk}$ mensuradas no mesmo "elemento-específico" são

$$\sigma^2_{ij} = \sigma_{i,jj} = \frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)}, \quad \sigma_{i,jk} = \rho_{ijk}\frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)}.$$

Na expressão para as covariâncias existem duas correlações: ρ_{ijk} capturando a correlação entre sorteios da distribuição beta e $(\alpha_1+\alpha_2+1)^{-1}$. Os parâmetros α_1 e α_2 podem variar em i e/ou j; com isso os momentos marginais mudam um pouco, ainda assim os cálculos são diretos.

As expressões para os momentos marginais obtidos anteriormente podem ser utilizadas para obter expressões aproximadas para média e elementos de variância-covariância para o caso especial dado em (3.11), desde que somente os efeitos fixos sejam admitidos, ou seja,

$$k_{ij} = \frac{\exp(\boldsymbol{x}^T_{ij}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^T_{ij}\boldsymbol{\beta})},$$

dessa forma,

$$\begin{aligned}
\text{E}(Y_{ij}) &= \frac{\alpha_1}{\alpha_1+\alpha_2}k_{ij}, \\
\text{Var}(Y_{ij}) &= \frac{\alpha_1}{\alpha_1+\alpha_2}k_{ij} - \left(\frac{\alpha_1}{\alpha_1+\alpha_2}\right)^2 k^2_{ij}, \\
\text{Cov}(Y_{ij}, Y_{ik}) &= \rho_{ijk}\frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)}k_{ij}k_{ik}.
\end{aligned} \tag{3.12}$$

Suposições adicionais de permutabilidade podem ser feitas, ou seja, $k_{ij} = k_{ik} \equiv k_i$ e $\rho_{ijk} = \rho_i$. Com isso, pode-se definir $k_i = 1$ e, finalmente, obter o modelo beta-binomial. Para generalizar para a versão binomial, define-se: $Z_i = \sum_{j=1}^{n_i} Y_{ij}$. Então, pode-se mostrar que $\text{E}(Z_i)$ e $\text{Var}(Z_i)$ resultam no modelo beta-binomial, conforme discutido na abordagem de dois estágios na Seção 3.2.2.

Ainda no âmbito de dados binários, Molenberghs et al. (2010) mostraram que o uso da função de ligação probit no lugar de logit resulta em expressões de forma fechada para média e variância. A função de ligação probit, também, possibilita o uso da fórmula de aproximação da distribuição normal para logística (Zeger et al., 1988), dada por

$$\frac{e^y}{1+e^y} \approx \Phi_1(cy),$$

em que $c = (16\sqrt{3}/15\pi)$. Aplicando a (3.10)-(3.11), tem-se

$$\begin{aligned}
\pi_{ij} &\sim \theta_{ij}\frac{\exp(\boldsymbol{x}^T_{ij}\boldsymbol{\beta} + \boldsymbol{z}^T_{ij}\boldsymbol{b}_i)}{1+\exp(\boldsymbol{x}^T_{ij}\boldsymbol{\beta} + \boldsymbol{z}^T_{ij}\boldsymbol{b}_i)}, \\
&\approx \theta_{ij}\Phi_1[c(\boldsymbol{x}^T_{ij}\boldsymbol{\beta} + \boldsymbol{z}^T_{ij})].
\end{aligned} \tag{3.13}$$

cujo uso abre uma janela de oportunidades para o caso logístico.

Nos casos específicos, mostrou-se que os MCs possuem dois tipos de efeitos aleatórios: (a) $\boldsymbol{b}_i$ com distribuição normal no preditor linear para acomodar a correlação entre as medidas repetidas (e alguma superdispersão); e (b) θ_{ij} para acomodar a superdispersão adicional (assumindo que os θ_{ij} são independentes). Alternativamente, pode-se supor que os θ_{ij} sejam correlacionados de modo que $\boldsymbol{\Sigma}_i$ possa assumir estruturas mais gerais. Isso implica no uso de distribuições multivariadas, por exemplo, a distribuição gama multivariada para o caso Poisson. Além dos casos específicos aqui apresentados, existem outras extensões dos MCs. Por exemplo, os modelos Weibull e Exponencial para medidas repetidas na análise de sobrevivência com efeitos aleatórios gama e normais (Molenberghs et al., 2010, 2017).

3.3.2 Modelo combinado marginalizado

Os aspectos da modelagem da superdispersão e correlação sob a ótica dos MCs e MLGM permitem a interpretação condicional ou "elemento-específico" (Iddi e Molenberghs, 2012). Todavia, isso está em contraste com as estimativas marginais ou de "média populacional" que também podem ser de interesse dos analistas. Tratando-se de inferência baseada na população, o modelo multinível marginalizado (MMM) é a abordagem mais robusta, pois combina a força das equações de estimativa generalizada (GEE) e MLGM (Zeger et al., 1988; Heagerty, 1999).

O MMM oferece a possibilidade de analisar um determinado problema sob a ótica marginal e condicional ao mesmo tempo, sem ajustar modelos separados (Iddi e Molenberghs, 2012). Mesmo quando há dados ausentes, aleatoriamente, ele fornece inferência válida, embora isso suponha que a estrutura de variância-covariância seja especificada corretamente (Hedeker et al., 2018). Assim, a formulação geral para um MMM de acordo com Heagerty (1999) é dada por

$$\begin{aligned} \mathrm{g}(\mu_{ij}^m) &= \boldsymbol{x}_{ij}^T\boldsymbol{\beta}^m, && (3.14)\\ \mathrm{g}(\mu_{ij}^c) &= \Delta_{ij} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i, \\ \boldsymbol{b}_i &\sim F_b(\boldsymbol{0}, \boldsymbol{D}), && (3.15)\\ \mathrm{Y}_{ij}^c &= \mathrm{Y}_{ij}|\boldsymbol{b}_i \sim F_{Y^c}(\mu_{ij}^c, \upsilon). && (3.16) \end{aligned}$$

em que υ é o parâmetro de dispersão, semelhante ao parâmetro ϕ na família exponencial (equação 3.1). A média marginal $\mu_{ij}^m = \mathrm{E}(Y_{ij})$ sob o contexto longitudinal depende de uma matriz $n_i \times p$ de preditores lineares $\boldsymbol{X}_i$ para uma função de ligação conhecida $g(\cdot)$. Já a média condicional $\mu_{ij}^c = \mathrm{E}(Y_{ij}|\boldsymbol{b}_i)$ relaciona-se com os efeitos aleatórios $\boldsymbol{b}_i$ distribuídos segundo (3.15) e a função Δ_{ij} conecta-se às médias marginal e condicional por meio de uma função de ligação comum. A variável resposta condicional possui distribuição dada por F_{Y^c}. A função Δ_{ij} é obtida por meio da solução da integral

$$\mu_{ij}^m = g^{-1}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}^m) = \int_b g^{-1}(\Delta_{ij} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)dF_b.$$

Por exemplo, para dados de contagem uma especificação log-log-normal leva a

$$\Delta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}^m - \boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij}/2. \tag{3.17}$$

Já para dados binomiais a expressão para Δ_{ij} pode ser encontrada em Griswold e Zeger (2004).

A estrutura de modelagem do MMM, considerando apenas a presença dos efeitos aleatórios $\boldsymbol{b}_i$, pode falhar em capturar com flexibilidade a superdispersão e a correlação simultaneamente. Iddi e Molenberghs (2012) propuseram combinar a estratégia de modelagem dos MCs com (3.14), (3.15) e (3.16) do MMM, originando o modelo combinado marginalizado (MCM) com formulação geral dada por

$$\begin{aligned}
\mathrm{g}(\mu_{ij}^m) &= \boldsymbol{x}_{ij}^T\boldsymbol{\beta}^m, \\
\mathrm{g}(k_{ij}) &= \Delta_{ij} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i, \\
\mu_{ij}^c &= \theta_{ij}k_{ij}, \\
\theta_{ij} &= \Theta_{ij}(\tau_{ij}, \sigma_{ij}^2), \\
\boldsymbol{b}_i &\sim F_b(\boldsymbol{0}, \boldsymbol{D}), \\
\mathrm{Y}_{ij}^c &= (\mathrm{Y}_{ij}|\theta_{ij}, \boldsymbol{b}_i) \sim F_{Y^c}(\mu_{ij}^c, \upsilon).
\end{aligned}$$

A distribuição da variável resposta está condicionada aos efeitos aleatórios $\boldsymbol{\theta}_i$ e $\boldsymbol{b}_i$ que, por conveniência, são considerados independentes. Isso implica em uma mudança na função Δ_{ij}, pois a média condicional é dada por $\mu_{ij}^c = \mathrm{E}(Y_{ij}|\theta_{ij}, \boldsymbol{b}_i)$. Assim, a função Δ_{ij} é obtida por meio da solução da integral dupla

$$\begin{aligned}
\mu_{ij}^m &= g^{-1}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}^m) = \int_b\int_\theta \theta_{ij}g^{-1}(\Delta_{ij} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)d\Theta_\theta dF_b \\
&= \int_b \mathrm{E}(\theta_{ij})g^{-1}(\Delta_{ij} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)dF_b.
\end{aligned} \tag{3.18}$$

Tomando, como exemplo, dados de contagem, utilizando o modelo MMM log-log-normal com distribuição gama para o parâmetro de superdispersão, ou seja, $\theta_{ij} \sim$ gama$(\alpha_{1j}, \alpha_{2j})$, a função Δ_{ij} em (3.18) é dada por

$$\Delta_{ij} = -\log(\alpha_{1j}\alpha_{2j}) + \boldsymbol{x}_{ij}^T\boldsymbol{\beta}^m - \boldsymbol{z}_{ij}^T\boldsymbol{D}\boldsymbol{z}_{ij}/2.$$

Para dados binários a expressão para função Δ_{ij}, assumindo $\theta_{ij} \sim \mathrm{Beta}(\alpha_{1j}, \alpha_{2j})$ para o modelo logístico-probit-normal, pode ser encontrada em Molenberghs et al. (2012).

3.3.3 Estimação

Os parâmetros dos MCs são estimados usando-se o método da máxima verossimilhança, sendo necessária a integração em relação aos efeitos aleatórios. A contribuição do indivíduo i para a função de verossimilhança é dada por

$$f_i(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{\Phi}_i, \boldsymbol{\Sigma}_i) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\theta}_i) f(\boldsymbol{b}_i|\boldsymbol{D}) f(\boldsymbol{\theta}_i|\boldsymbol{\Phi}_i, \boldsymbol{\Sigma}_i) d\boldsymbol{b}_i d\boldsymbol{\theta}_i. \tag{3.19}$$

em que $\boldsymbol{\beta}$ agrupa todos os parâmetros modelo condicional para $\boldsymbol{Y}_i$. Logo, a função de verossimilhança marginal é dada por

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{\Phi}_i, \boldsymbol{\Sigma}_i) &= \prod_{i=1}^{N} f_i(y_i|\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{\Phi}_i, \boldsymbol{\Sigma}_i) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\theta}_i) f(\boldsymbol{b}_i|\boldsymbol{D}) f(\boldsymbol{\theta}_i|\boldsymbol{\Phi}_i, \boldsymbol{\Sigma}_i) d\boldsymbol{b}_i d\boldsymbol{\theta}_i \end{aligned} \quad (3.20)$$

O problema de maximizar (3.20) é a presença de N integrais em relação aos efeitos aleatórios $\boldsymbol{b}_i$ e $\boldsymbol{\theta}_i$. Esse problema pode ser simplificado com uso de marginalização parcial, isto é, integrar analiticamente em relação ao efeito aleatório conjugado e numericamente sobre o efeito aleatório normal (Molenberghs et al., 2007, 2010, 2012, 2017).

A marginalização parcial para o caso de contagens (Seção 3.3.1), ou seja, para o caso particular do MC especificado nas equações (3.4) - (3.8) foi obtida por Molenberghs et al. (2007) pela integração dos efeitos aleatórios θ_{ij} com distribuição gama assumidos independentes, resultando em

$$f(y_{ij}|\mathbf{b_i}) = \binom{\alpha_{1j} + y_{ij} - 1}{\alpha_{1j} - 1} \left(\frac{\alpha_{2j}}{1 + k_{ij}\alpha_{2j}}\right)^{y_{ij}} \left(\frac{1}{1 + k_{ij}\alpha_{2j}}\right)^{\alpha_{1j}} k_{ij}^{y_{ij}}. \quad (3.21)$$

Para o caso binário, Molenberghs et al. (2012) obtiveram

$$f\left(y_{ij}|\boldsymbol{b_i}\right) = \frac{1}{\alpha_{1j} + \alpha_{2j}} (k_{ij}\alpha_{1j})^{y_{ij}} [(1 - k_{ij})\alpha_{1j} + \alpha_{2j}]^{1-y_{ij}}. \quad (3.22)$$

A seguir, faz-se a integração numérica das expressões (3.21) e (3.22) em relação ao efeitos aleatórios normais $\boldsymbol{b_i}$, via, por exemplo, método de Laplace (Molenberghs e Verbeke, 2005; Kristensen et al., 2016). Os parâmetros α_1 e α_2 em (3.21) e (3.22) não são identificáveis. Uma solução para o primeiro caso é supor que $\alpha_2 = 1/\alpha_1$, enquanto que, no segundo caso, pode-se definir, $\alpha_2/\alpha_1 = c$ em que c é uma constante.

A interpretação do vetor de parâmetros fixos $\boldsymbol{\beta}$ nos MCs é a mesma que em um MLGM clássico, uma vez que θ_{ij} segue uma distribuição conjugada (Kassahun et al., 2012). Isso significa que o efeito nos parâmetros de regressão vem apenas dos efeitos aleatórios normais no preditor linear (Molenberghs e Verbeke, 2005).

A estimação dos parâmetros dos modelos MMM, também, é feita pelo método da máxima verossimilhança (Griswold e Zeger, 2004; Iddi e Molenberghs, 2012; Hedeker et al., 2018). Já os parâmetros do modelo MCM são estimados utilizando as mesmas expressões parcialmente marginalizadas, (3.21) e (3.22), necessitando apenas substituir k_{ij} por $k_{ij}^* = g^{-1}(\Delta_{ij} + \boldsymbol{z}_{ij}^T \boldsymbol{b}_i)$. É importante mencionar que o uso do método de Laplace pode apresentar problemas de convergência e precisão para os modelos MCM em dados binários, pois a expressão para função Δ_{ij} exibe uma relação não linear com a matriz de componentes de variância $\boldsymbol{D}$ e o parâmetro de regressão $\boldsymbol{\beta}^m$ (Iddi e Molenberghs, 2012). Já para o caso de contagens, o conector em (3.17) é linear e portanto, não ocorre mudanças significativas nos componentes de variância.

3.4 Seleção de modelos - inferência sobre efeitos aleatórios

Para a seleção da parte aleatória de um modelo, são feitos testes de hipóteses sobre os componentes de variância/covariância, do tipo $H_0 : \sigma_{ij} = 0$. Pode-se usar o teste da razão de verossimilhanças (LRT) que se baseia na comparação dos valores das funções de verossimilhança de dois modelos aninhados, tendo o mesmo número de parâmetros de efeito fixo e diferentes números de parâmetros de efeito aleatório. É dado por

$$LRT = -2[\text{logLik(modelo reduzido) - logLik(modelo completo)}],$$

em que logLik é o logaritmo da função de verossimilhança.

Quando o teste não é no limite do espaço paramétrico, assintoticamente, $LRT \sim \chi^2_\nu$, em que ν é a diferença entre o número de parâmetros dos dois modelos. Quando o teste é no limite do espaço paramétrico ($H_0 : \sigma_{ii} = 0$ versus $H_a : \sigma_{ii} > 0$), LRT tem distribuição que é uma mistura de χ^2's (Zhang e Lin, 2008).

No caso de modelos de componentes de variância com independência entre os efeitos aleatórios a mistura de χ^2's é dada por

$$\sum_{m=0}^{k'} 2^{-k'} \binom{k'}{m} \chi^2_m, \tag{3.23}$$

sendo k' o número de componentes de variância sob H_0.

3.5 Seleção de modelos - inferência sobre efeitos fixos

Depois de escolher os termos aleatórios do modelo, a seleção do preditor linear, em geral, envolve comparações de modelos aninhados e diferenças de "deviances" ("Analysis of deviance"), isto é, testes de razão de verossimilhanças. Ela envolve avaliar o valor da função de verossimilhança para o modelo completo e para o modelo sob H_0 (modelo reduzido), usando o método de máxima verossimilhança

$$\begin{aligned} LRT &= -2[\text{logLik(modelo reduzido) - logLik(modelo completo)}] \\ &= \text{deviance(modelo reduzido) - deviance(modelo completo)}, \end{aligned}$$

em que logLik é o logaritmo da função de verossimilhança. Os modelos aninhados e o modelo de referência têm o mesmo número de parâmetros de covariância e diferentes conjuntos de parâmetros de efeito fixo. Assintoticamente, $LRT \sim \chi^2_\nu$, sendo ν a diferença em número de parâmetros de efeito fixo dos dois modelos.

Outro critério bastante empregado é o critério de informação de Akaike (AIC) (Akaike, 1974), que permite a comparação de múltiplos modelos aninhados ou não (Bolker et al., 2009),

$$AIC = -2\log(L) + 2p.$$

Em que $\log(L)$ é o logaritmo da função de verossimilhança para o modelo e p é o número de parâmetros do modelo. De acordo com esse critério o melhor modelo é aquele que tem o menor valor de AIC (Akaike, 1974).

3.6 Pacote combTMB

3.6.1 Implementação

Com a finalidade de ajustar os MLGs, MLGMs e MCs no software de código-aberto `R` (R Core Team, 2022), foi desenvolvido o pacote **combTMB**. Para maximizar a flexibilidade e a velocidade no **combTMB**, a estimação por máxima verossimilhança é realizada, tendo com ferramenta o "Template Model Builder" (pacote **TMB**) (Kristensen et al., 2016). O **TMB** permite o ajuste de modelos com e sem efeitos aleatórios. Foi construído utilizando bibliotecas C++ de alto desempenho, como **CppAD** (Bell, 2005), **Eigen** (Guennebaud et al., 2010), **BLAS** (Blackford et al., 2002), entre outras bibliotecas, que são responsáveis por obter as derivadas usando diferenciação automática, cálculos de álgebra linear e paralelização, respectivamente.

A interface do pacote **combTMB** (por exemplo, a sintaxe da fórmula) é similar à do pacote **lme4** (Bates et al., 2015), e exibe mensagens informativas no padrão do pacote **cli** (Csárdi, 2022). Assim como o pacote **lme4**, o **combTMB** utiliza a aproximação de Laplace para para fazer a integração numérica em relação aos efeitos aleatórios. A otimização interna da aproximação Laplace é realizada pelo método de Newton, com a primeira e segunda derivadas fornecidas por diferenciação automática a partir do pacote **TMB**. Em seguida, a verossimilhança marginal pode ser otimizada para obter as estimativas dos parâmetros por máxima verossimilhança ou, opcionalmente, por máxima verossimilhança restrita (REML) utilizando algoritmos como PORT ou Broyden–Fletcher–Goldfarb–Shanno (BFGS), implementados nas rotinas `stats::nlminb()` e `stats::optim()` do `R`. Uma das principais vantagens do pacote **combTMB**, em comparação com o pacote **lme4**, é a velocidade na estimativa de modelos não Gaussianos (Figura 3.5).

A principal função para o ajuste de modelos no pacote **combTMB** é a função `combTMB()`. Os principais argumentos dessa função são:

```
combTMB(formula, data = NULL, family = gaussian(link = "identity"),
  dformula = ~1, REML = FALSE, doMarginal = FALSE, weights = NULL,
  offset = NULL, contrasts = NULL, na.action, starting_val = "zero",
  control = combTMBcontrol(), map = NULL, ...)
```

que retorna um objeto da classe `combTMB`, contendo várias informações referentes ao ajuste do modelo. Os três primeiros argumentos dessa função são responsáveis pela descrição simbólica do preditor linear, especificação do quadro de dados e tipo de modelo com a fun-

ção de ligação adotada. Os demais argumentos são utilizados quando o usuário necessita de um modelo mais complexo ou quando as configurações padrões não são adequadas.

De maneira geral, no pacote **combTMB**, um modelo é formado por três componentes principais: uma fórmula bilateral (com a variável resposta à esquerda e preditores à direita, incluindo efeitos aleatórios potenciais), uma distribuição para a variável resposta e uma fórmula para o modelo de dispersão. Atualmente, apenas interceptos aleatórios independentes e identicamente distribuídos (IID) são permitidos. MLGs simples são obtidos quando os efeitos aleatórios não são declarados e as configurações padrões são mantidas. Para ajustar um MLGM ou MC, utiliza-se a mesma sintaxe do **lme4**. Por exemplo, para o ajuste de um modelo ao conjunto de dados "embryos" (Seção 3.7) pode-se considerar o número de oócitos totais como variável resposta (OT), o período do ano (`Period`) como efeito fixo e doadora (`Donor`) como efeito aleatório resultando a fórmula

```
OT ~ Period + (1|Donor)
```

A distribuição para a variável resposta é especificada usando o argumento `family`. Ao todo, 13 distribuições (Tabela 3.1) podem ser utilizadas diretamente para construção de modelos usando o pacote **combTMB**, abrangendo variáveis respostas contínuas e discretas. A distribuição padrão adotada é a normal, $N(\mu, \sigma^2)$, com função de ligação identidade. Outras distribuições podem ser obtidas com uso dos argumentos `map` e `start_params` (informações sobre esses argumentos são obtidas digitando `?combTMB()` e `?combTMBcontrol()` no `R`).

Tabela 3.1. Tipos de distribuições implementadas no **combTMB** e funções de ligação padrões.

Distribuição	Código de exemplo
Normal	`family = gaussian(link = "identity")`
Poisson	`family = poisson(link = "log")`
Poisson-gama	`family = poigamma(link = "log")`
Geométrica	`family = geometric(link = "log")`
Poisson-generalizada	`family = gpoisson(link = "log")`
COM-Poisson	`family = cmp(link = "log")`
Binomial negativa generalizada	`family = gnb(link = "log")`
Beta binomial negativa	`family = bbn(link = "log")`
Beta	`family = beta_fam (link = "logit")`
Binomial	`family = binomial(link = "logit")`
Beta-binomial	`family = betabinomial(link = "logit")`
Beta-Bernoulli	`family = betabernoulli(link = "logit")`
Beta-Bernoulli-aproximada	`family = betabernoulli_al(link = "probit")`

As distribuições Poisson-gama, beta-binomial, beta-Bernoulli e beta-Bernoulli-aproximada são casos particulares dos MCs formulados na Seção 3.3, utilizadas para a análise de dados na forma de contagens, proporções e binários, respectivamente (Tabela 3.1). Quando a distribuição Poisson-gama é especificada somente com efeitos fixos no argumento `formula`, ou seja, $k_{ij} = g(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})$, a distribuição binomial negativa com variância

dada em (3.3) é obtida. De forma semelhante, a distribuição beta-binomial, reduz-se ao caso especial com variância dada em (3.2). As distribuições beta-Bernoulli e beta-Bernoulli-aproximada são recomendadas somente para dados binários mensurados em um contexto longitudinal ou hierárquico, portanto, necessitam da inclusão dos efeitos aleatórios no argumento `formula`. Vale mencionar que a distribuição beta-Bernoulli-aproximada é uma aproximação do caso `family = betabernoulli(link = "logit")` utilizando a função de ligação "probit", conforme a equação (3.13), relacionando as funções de densidades normal e logística (Molenberghs et al., 2012).

As distribuições normal, Poisson e binomial estão implementadas na maioria dos pacotes do `R` que ajustam MLGs e MLGMs. Por outro lado, as distribuições geométrica, Poisson generalizada, COM-Poisson (Conway-Maxwell-Poisson), binomial negativa generalizada, beta binomial negativa e Beta possuem pouco suporte computacional, especialmente em softwares livres. O pacote **combTMB** agrega todas essas distribuições em uma única ferramenta computacional, permitindo a comparação rápida e eficiente entre os diversos modelos.

A distribuição geométrica pode ser utilizada quando o parâmetro de dispersão da distribuição binomial negativa é próximo de 1 (Hilbe, 2011), ou seja, é uma distribuição intermediária entre as distribuições Poisson e binomial negativa. Já para situações mais gerais, em que existe subdispersão ou superdispersão nos dados, as distribuições Poisson generalizada e COM-Poisson podem ser utilizadas (Consul e Famoye, 1992; Huang, 2017). O parâmetro de dispersão ϕ para a distribuição Poisson generalizada é baseado na distribuição lognormal (Hilbe, 2011). Quanto à distribuição COM-Poisson, foi considerada a versão reparametrizada de acordo com Huang (2017).

A distribuição binomial negativa generalizada foi desenvolvida por Greene (2008) para permitir mais flexibilidade na variância do modelo binomial negativo (3.3). A variância para essa distribuição é dada por

$$\text{Var}(Y_i) = \mu_i(1 + \alpha\mu_i^{Q-1}),$$

em que Q é um parâmetro a ser estimado. O modelo obtido tem três parâmetros: μ, α e Q. Os modelos binomiais negativos tipo I e II são obtidos quando $Q = 1$ e $Q = 2$, respectivamente. No pacote **combTMB**, essa distribuição foi implementada de acordo com Hilbe (2011).

Ainda no contexto dos dados de contagem, pode-se modelar a superdispersão, assumindo que a variância seja dividida em três partes (puro acaso (ou aleatoriedade), diferentes exposições ao risco no processo de contagem (responsabilidade) e diferenças apenas devido a características individuais (propensão)) por meio da distribuição beta binomial negativa, também conhecida como distribuição Waring Generalizada (Irwin, 1968; Rodríguez-Avi et al., 2009; Vílchez-López et al., 2016). No pacote **combTMB**, essa distribuição foi implementada considerando as condições $k > 0$ e $\rho > 1$ para garantir a

existência da média do modelo (Rodríguez-Avi et al., 2009), sendo os parâmetros k e ρ estimados, considerando a parametrização

$$k = \exp(k_0), \quad \rho = 1 + \exp(\rho_0),$$

com k_0 e $\rho_0 \in \mathbb{R}$. O pacote **GWRM** (Vílchez-López et al., 2016) implementa essa distribuição mas não permite a inclusão de efeitos aleatórios normais no preditor linear. Da mesma forma que o pacote **GWRM**, o pacote **combTMB** contém a função `partvar()` que divide a variância em três componentes.

Variáveis respostas que assumem valores no intervalo padrão de unidade (0,1), como taxas, proporções ou índices de concentração, podem ser modeladas por meio da regressão beta. No pacote **combTMB**, a versão da regressão beta é a mesma do pacote **betareg** (Cribari-Neto e Zeileis, 2010). Além disso, com o uso do pacote **combTMB**, é possível adicionar efeitos aleatórios no preditor linear para a regressão beta, o que não é possível com os recursos atuais do **betareg**.

O modelo para o parâmetro de dispersão adotado no pacote **combTMB** tem como padrão a função de ligação logarítmica e preditor linear especificado pelo argumento

```
dformula = ~1
```

Nesse caso, o parâmetro de dispersão (por exemplo, α_1 para distribuição binomial negativa) é idêntico para todas as observações. Alternativamente, podem-se incorporar regressores de efeitos fixos no preditor linear para explicar a heteroscedasticidade (Hilbe, 2011). Por exemplo, se a contagem de OT for mais variável (em relação a média) à medida que o intervalo entre coletas (`Interval`) de oócitos aumenta em uma mesma doadora, então um modelo com distribuição Poisson-gama pode usar a fórmula unilateral

```
dformula = ~Interval
```

Ademais, os regressores de efeitos fixos podem ser os mesmos nos argumentos `formula` e `dformula`, porém, isso pode potencialmente levar a problemas de não convergência.

3.6.2 Instalação

O código-fonte e o site de ajuda do pacote **combTMB** com todas as informações sobre suas fonções e parâmetros estão disponíveis no repositório do GitHub (`https://github.com/deoclecioamorim/combTMB`). Para sua instalação é necessário que o usuário tenha um compilador C++ no seu computador. Então, executam-se os comandos no `R`:

```
#install.packages("remotes")
remotes::install_github("deoclecioamorim/combTMB", dependencies = TRUE)
```

3.7 Aplicações

O uso do pacote **combTMB** é ilustrado com duas aplicações, utilizando o conjunto de dados "embryos" (para acessá-lo no R, digite `data(embryos, package = "combTMB")`), composto por 15 variáveis e 1148 observações, referentes à produção *in vitro* de embriões bovinos da raça holandesa (*Bos taurus*). Esse conjunto de dados é proveniente de um estudo observacional da fazenda RuAnn and Maddox Dairy Farms in Riverdale, California, USA (Demétrio et al., 2020). Cada observação corresponde a uma sessão de coleta dos óvulos (OPU), realizada em 318 fêmeas bovinas doadoras de oócitos de três status: novilhas (H), vacas em lactação (M) e vacas secas (D). As observações das sessões de OPU foram classificadas por períodos do ano (P1 -período de temperatura mais elevada, junho, julho, agosto, setembro e outubro; P2 - período de temperatura menos elevada, sete meses restantes) no estado da Califórnia (EUA). A fertilização *in vitro* foi feita a partir do sêmen de 66 touros.

As variáveis respostas consideradas foram o número de oócitos totais (OT) e a taxa de clivagem (D3, total de embriões no terceiro dia de cultivo *in vitro*, dividido por IVC, total de oócitos em cultivo *in vitro*), usadas como exemplos de dados na forma de contagens e proporções, respectivamente.

3.7.1 Dados de contagem

Considera-se que Y_{jk} representa a variável resposta OT, com valores y_{jk} obtidos no j-ésimo ($j = P1, P2$) período para uma doadora no k-ésimo ($k = D, H, M$) status. Pode-se supor que Y_{jk} tem distribuição Poisson, binomial negativo (Poisson-gama), Poisson-normal ou Poisson-gama-normal com média μ_{jk} e função de ligação logarítmica, $\eta_{jk} = \log(\mu_{jk})$, conforme especificado na Tabela 3.1. Para os modelos Poisson-normal e Poisson-gama-normal (modelo combinado), consideram-se os seguintes preditores lineares

$$\eta_{jk} = \beta_0 + \tau_j + \xi_{d(jk)},\ d = 1, \ldots, 318 \tag{3.24}$$

$$\eta_{jk} = \beta_0 + \tau_j + \delta_k + \xi_{d(jk)},\ d = 1, \ldots, 318 \tag{3.25}$$

$$\eta_{jk} = \beta_0 + \tau_j + \delta_k + \gamma_{jk} + \xi_{d(jk)},\ d = 1, \ldots, 318 \tag{3.26}$$

em que β_0 é a constante de efeito fixo, τ_j é efeito fixo do j-ésimo período do ano, δ_k é o efeito fixo do k-ésimo status da doadora, γ_{jk} é o efeito fixo da interação entre período do ano status da doadora e $\xi_{d(jk)} \sim N(0, \sigma_d^2)$ é o efeito aleatório de doadora. A inclusão do efeito aleatório de doadora é para acomodar a hierarquia pois as doadoras foram aspiradas em mais de uma ocasião e, portanto, suas observações têm alguma forma de dependência (Molenberghs et al., 2007).

Considerando o preditor linear maximal para a parte fixa do modelo, uma das primeiras etapas é a seleção da parte aleatória do modelo, na equação (3.26). Desta forma, testa-se a hipótese sobre o componente de variância relacionado à doadora ($H_0 : \sigma_d^2 = 0$),

utilizando o teste LRT, comparando os modelos Poisson (M1) versus Poisson-normal (M2) e binomial negativo (M3) versus Poisson-gama-normal (M4). Para os modelos M1 e M3, elimina-se $\xi_{d(jk)}$ do preditor linear (3.26). Esses modelos são ajustados no pacote **combTMB**, usando os seguintes comandos:

```
##--Modelo Poisson (MLG)
M1 <- combTMB(OT ~ Period * Status, embryos, family=poisson)
##--Modelo Poisson-normal (MLGM)
M2 <- combTMB(OT ~ Period * Status + (1|Donor), embryos, family=poisson)
##--Modelo Binomial negativo (MLG)
M3 <- combTMB(OT ~ Period * Status, embryos, family=poigamma)
##--Modelo Poisson-gama-normal (MC)
M4 <- combTMB(OT ~ Period * Status + (1|Donor), embryos, family=poigamma)
```

Como é comum nas declarações simbólicas do preditor linear no argumento `formula`, o símbolo $*$ indica uma interação, além dos efeitos principais. Nos modelos, também forneceu-se o conjunto de dados `embryos` e especificaram-se as distribuições Poisson e Poisson-gama usando o argumento `family`. O logaritmo é assumido como a função de ligação padrão, logo, não é necessário declará-lo no **combTMB**. Observa-se que a sintaxe para ajustar os MLGs é semelhante à usada na função `glm()`, enquanto a sintaxe para ajustar os MLGMs e MCs é semelhante à usada na função `glmer()` do pacote **lme4**.

Utilizando as funções auxiliares `logLik()`, `df.residual()` e `AIC()` os resultados dos ajustes dos modelos M1-M4 foram sumarizados na Tabela 3.2. Para comparar os modelos da Tabela 3.2 obtiveram-se os valores da diferença da estatística $-2\log(L)$ para os modelos encaixados e aplicou-se o teste LRT ao nível de 5% de significância para testar a hipótese $H_0 : \sigma_d^2 = 0$ versus $H_0 : \sigma_d^2 > 0$. É importante ressaltar que esse teste está no limite do espaço paramétrico e, portanto, valor calculado é comparado com uma mistura de qui-quadrado (equação (3.23)) dado por

$$\sum_{m=0}^{1} 2^{-1}\binom{1}{m}\chi_m^2 = \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 = 1,9210.$$

O teste LRT sugere que há fortes evidências para se rejeitar H_0. Portanto, há necessidade do efeito aleatório de doadora nos modelos Poisson-normal e Poisson-gama-normal. No entanto, os valores da estatística $-2\log(L)$ e do critério AIC sugerem o modelo Poisson-gama-normal, M4 (modelo combinado), como o melhor modelo (Tabela 3.2). A escolha do modelo M4 também é confirmada pelo teste LRT que compara a hipótese H_0 : modelo Poisson-normal (M2) ($\theta \to \infty$) versus H_a : M4. Note que esse teste está no limite do espaço paramétrico com distribuição dada pela mistura $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 = \frac{1}{2}\chi_1^2$ (Lawless, 1987). Tem valor $LRT = 182,8831$ com 1 grau de liberdade e valor $p < 0,0001$.

A partir da seleção da parte aleatória, a próxima questão de interesse é obter o modelo mais parcimonioso, considerando os preditores lineares (3.24) e (3.25), ou seja,

Tabela 3.2. Resultados do ajuste de vários modelos aos dados de oócitos totais, usando os mesmos efeitos fixos no preditor linear de equação (3.26), com e sem efeito aleatório de doadora.

Poisson versus Poisson-normal					
Modelo	Parte fixa	Efeito aleatório	$-2\log(L)$	GL	AIC
M1	eq. (3.26)	nenhum	10883,2300	1142	10895,2300
M2	eq. (3.26)	Donor	7780,3870	1141	7794,3870
Diferença (M1−M2)	–	–	3102,8430	–	–
Binomial negativo versus Poisson-gama-normal					
Modelo	Parte fixa	Efeito aleatório	$-2\log(L)$	GL	AIC
M3	eq. (3.26)	nenhum	8172,5700	1141	8186,5700
M4	eq. (3.26)	Donor	7597,5040	1140	7613,5040
Diferença (M3−M4)	–	–	575,0660	–	–

$\log(L)$ é o logaritmo da função de verossimilhança e GL - número de graus de liberdade residual dos modelos ajustados.

o objetivo agora é verificar se a parte fixa do preditor linear de equação (3.26) pode ser reduzida. Os códigos para o ajuste dos modelos com os preditores lineares (3.24) e (3.25) são:

```
##--MCs
M5 <- combTMB(OT ~ Period + (1|Donor), embryos, family=poigamma)
M6 <- combTMB(OT ~ Period + Status + (1|Donor), embryos, family=poigamma)
```

Nessa etapa, o critério utilizado para comparar os modelos (M4, M5 e M6) foi o Critério de Informação de Akaike (AIC, Akaike (1974)) com o uso da função `AICtab()` do pacote **bbmle** (Bolker e R Development Core Team, 2022). Os códigos são apresentados no Apêndice A1.

Dos modelos considerados, o mais parcimonioso foi o modelo combinado M4, pois apresentou o menores valores para estatística $-2\log(L)$ e AIC. O modelo M4, pode ser representado conforme as equações (3.4) - (3.8) e a visão geral do ajuste pode ser obtido com o uso da função `summary()`:

```
R> summary(M4)
Family: poigamma
Link function: log
Formula: OT ~ Period * Status + (1 | Donor)
Dformula: ~1
Number of obs: 1148
-2 x logLik         AIC         BIC    df.resid
     7597.5      7613.5      7653.9        1140
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.016660   0.055702  54.157  < 2e-16 ***
```

```
PeriodP2          -0.002981   0.035812  -0.083  0.93366
StatusH           -0.499990   0.071490  -6.994 2.67e-12 ***
StatusM           -0.371633   0.088879  -4.181 2.90e-05 ***
PeriodP2:StatusH   0.100128   0.053751   1.863  0.06249 .
PeriodP2:StatusM   0.219397   0.077670   2.825  0.00473 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Random effects:
                   Estimate Std. Error
Donor (Intercept)    0.2069      0.021
Overdisp.(theta)    22.2973      2.569

Number of subjects: 318
```

A saída obtida com a função `summary()` pode ser dividida em três partes. A parte inicial dá uma visão geral contendo uma descrição do modelo (`Family`, `Link function`, `Formula`, `Dformula`, `Number of obs`) e resultados dos critérios de informação. A segunda parte, descreve a parte fixa do modelo com os valores dos coeficientes $\boldsymbol{\beta}$ incluindo a estatística Z de Wald e valores de p. As soluções para os parâmetros $\boldsymbol{\beta}$ são obtidas usando a codificação `R` padrão para contrastes (`contr.treatment`), em que o primeiro valor é usado como referência (`Intercept`). A última parte (`Random effects`) mostra as estimativas do componente de variância e do parâmetro de superdispersão com os respectivos erros-padrão, calculados pelo método delta (Kristensen et al., 2016).

A qualidade de ajuste do modelo M4 pode ser verificada, inicialmente, com a função `sanitycombTMB()` (`sanitycombTMB(M4)`, Apêndice A1). Essa função permite verificar se existem anomalias ou falta de convergência no ajuste do modelo. Outra etapa importante na validação do modelo é a análise de resíduos, complementando os teste de hipóteses e a seleção do modelo mais apropriado (Thygesen et al., 2017). A análise de resíduos pode ser feita com a função `residuals()` obtendo os resíduos quantílicos aleatorizados (Dunn e Smyth, 1996) como padrão. Outra alternativa é a utilização de resíduos baseados em simulação (`dharma_residuals()`) (Hartig, 2022). A seguir demonstra-se o uso da função (`dharma_residuals()`) para avaliar a qualidade de ajuste do modelo M4:

```
R> ##---Resíduos com a função dharma_residuals
R> resi1 <- dharma_residuals(M4, nsim=250, plot=FALSE)
R> head(resi1$out_1)
     Observed    Expected
1 0.000000000 0.000870322
2 0.001383623 0.001740644
```

```
3 0.001519784 0.002610966
R> resi2 <- dharma_residuals(M4, nsim=500, plot=FALSE)
R> head(resi2$out_1)
      Observed    Expected
1 0.0000000000 0.000870322
2 0.0006214178 0.001740644
3 0.0010475462 0.002610966
```

Os resíduos obtidos estão padronizados para valores entre 0 e 1 e podem ser interpretados da mesma forma que os resíduos de uma regressão linear simples (Hartig, 2022). A função executa três passos principais: i - cria n conjunto de dados simulados a partir do modelo ajustado; ii - calcula a distribuição cumulativa de valores simulados para cada valor observado e iii - retorna o valor do quantil que corresponde ao valor observado. Uma distribuição uniforme dos resíduos indicam que o modelo está especificado corretamente. Na Figura 3.1 os resíduos foram obtidos a partir de 250 e 500 simulações e ambas as situações indicam boa qualidade de ajuste para o modelo M4.

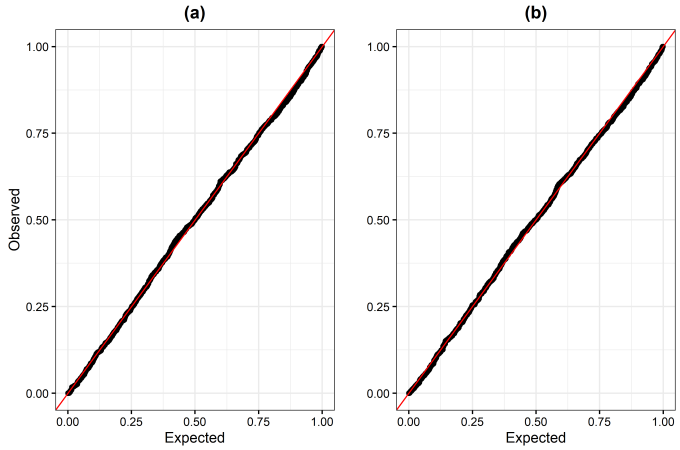

Figura 3.1. Resíduos quantílicos aleatorizados obtidos por meio da função `dharma_residuals()` para o modelo M4: (a) Resíduos obtidos a partir de 250 simulações e (b) Resíduos obtidos a partir de 500 simulações.

As estimativas dos parâmetros para os modelos Poisson, binomial negativo, Poisson-normal e Poisson-gama-normal, usando o preditor linear com a equação (3.26) e os valores para as estatísticas $-2\log(L)$ e AIC, estão apresentados na Tabela 3.3. O principal impacto observado pelo ajuste dos diferentes modelos é a mudança nos erros-padrão das estimativas dos parâmetros e na significância dos testes para os parâmetros.

Claramente, o modelo Poisson-normal é superior aos modelos Poisson simples e binomial negativo em termos de verossimilhança. No entanto, o modelo Poisson-normal ignora grande parte da superdispersão, levando em conta basicamente a correlação entre as medidas repetidas na mesma doadora. Por outro lado, o modelo Poisson-gama-normal (M4), que é um modelo combinado, utiliza dois conjuntos de efeitos aleatórios. O efeito

aleatório normal de doadora, σ_d^2, modela a correlação existente entre as medidas repetidas e parte da superdispersão, enquanto que os efeitos aleatórios θ_{ij}, com distribuição gama, são responsáveis por acomodar a superdispersão (Molenberghs et al., 2007, 2017). Portanto, é um modelo mais realista. Vale mencionar que os efeitos aleatórios θ_{ij} são considerados independentes e introduzidos em nível de observação, o que implica em independência entre as medidas repetidas (Molenberghs et al., 2007, 2010).

Tabela 3.3. Estimativas dos parâmetros, com seus respectivos erros-padrão (e.p.), para os modelos Poisson, binomial negativo, Poisson-normal e combinado com o preditor linear (3.26), para a variável resposta OT.

	Poisson		Binomial negativo	
Parâmetros†	Estimativa (e.p.)		Estimativa (e.p.)	
Intercept	3,1319 (0,0181)	***	3,1319 (0,0455)	***
Period - P2	−0,0116 (0,0223)		−0,0116 (0,0560)	
status - H	−0,5763 (0,0251)	***	−0,5763 (0,0573)	***
status - M	−0,3560 (0,0340)	***	−0,3560 (0,0775)	***
Period - P2:status - H	0,1550 (0,0320)	***	0,1550 (0,0730)	*
Period - P2:status - M	0,1702 (0,0422)	***	0,1702 (0,0974)	
Parâmetro-Superdispersão	—		4,3088	
$-2\log(L)$	10883,2000		8172,6000	
AIC	10895,2000		8186,6000	
	Poisson-normal		Combinado	
Parâmetros†	Estimativa (e.p.)		Estimativa (e.p.)	
Intercept	2,9839 (0,0510)	***	3,0167 (0,0557)	***
Period - P2	0,0076 (0,0249)		−0,0030 (0,0358)	
status - H	−0,4670 (0,0658)	***	−0,5000 (0,0715)	***
status - M	−0,3555 (0,0774)	***	−0,3716 (0,0889)	***
Period - P2:status - H	0,0811 (0,0405)	*	0,1001 (0,0538)	
Period - P2:status - M	0,2323 (0,0595)	***	0,2194 (0,0777)	**
Parâmetro-Superdispersão	—		22,2979	
Donor (σ_d^2)	0,2274		0,2069	
$-2\log(L)$	7780,4000		7597,5000	
AIC	7794,4000		7613,5000	

†Efeitos dos parâmetros são calculados usando a codificação padrão para contrastes, o primeiro valor é usado como referência; significância (***valor $p < 0,0001$, **valor $p < 0,01$ e *valor $p < 0,05$) e $\log(L)$ é o logaritmo da função de verossimilhança.

Adicionalmente, o pacote **combTMB**, no caso de dados na forma de contagens, possibilita analisar a contagem de OT sob a ótica marginal e condicional simultaneamente, por meio do argumento `doMarginal=TRUE`. O modelo M4 é marginalizado seguindo as suposições teóricas apresentadas na Seção 3.3.2 e é obtido da seguinte maneira:

```
M4MCM <- combTMB(OT ~ Period * Status + (1|Donor), embryos,
                 family=poigamma, doMarginal = TRUE)
```

As estimativas dos parâmetros e correspondentes erros-padrão, considerando o modelo M4MCM, estão apresentados na Tabela 3.4. Observa-se que apenas a estimativa do intercepto é afetada, enquanto que as estimativas dos demais parâmetros permanecem iguais às do modelo M4. As estimativas do componente de variância e do parâmetro de dispersão bem como os valores das estatísticas $-2\log(L)$ e AIC também não mudam. Assim, o modelo resultante possui três propriedades desejáveis: (i) modelagem flexível da superdispersão; (ii) modelagem flexível de associação e efeitos "elementos-específico"; (iii) interpretação marginal direta dos parâmetros de regressão (Iddi e Molenberghs, 2012).

Tabela 3.4. Estimativas dos parâmetros com seus respectivos erros-padrão (e.p.) para os modelos M4 e M4MCM para a variável resposta OT.

	M4	M4MCM
Parâmtros †	Estimativa (e.p.)	Estimativa (e.p.)
(Intercept)	3,0167(0,0557)	3,1201(0,0560)
PeriodP2	$-0,0030$(0,0358)	$-0,0030$(0,0358)
StatusH	$-0,5000$(0,0715)	$-0,5000$(0,0715)
StatusM	$-0,3716$(0,0889)	$-0,3716$(0,0889)
PeriodP2:StatusH	0,1001(0,0538)	0,1001(0,0538)
PeriodP2:StatusM	0,2194(0,0777)	0,2194(0,0777)
σ_d^2	0,2070(0,0200)	0,2070(0,0200)
θ_{ij}	22,2980(2,5700)	22,2980(2,5700)
$-2\log(L)$	7597,5000	7597,5000
AIC	7613,500	7613,5000

† Estimativas dos parâmetros obtidas por máxima verossimilhança empregando a aproximação de Laplace; σ_d^2 é o componente de variância referente à doadora; θ é parâmetro de dispersão e $\log(L)$ é o logaritmo da função de verossimilhança.

A interpretação marginal do modelo M4MCM pode ser complementada com o uso do pacote **emmeans** (Lenth, 2022) para o cálculo das médias marginais do efeito de interação `Status*Period`. Intervalos de confiança com 95% de confiança para médias marginais para as combinações dos níveis de períodos do ano (P1: Junho a Outubro e P2: sete meses restantes) e status das doadoras (H: novilhas, M: vacas em lactação e D: vacas secas) considerando os modelos Poisson e combinado (M4MCM) com o preditor linear de equação (3.26) podem ser observados na Figura 3.2. Nesse exemplo, no período de temperaturas mais elevadas (P1), as doadoras de status "vacas secas" (D) apresentam a maior média para o número de OT por OPU, diferindo estatisticamente das doadoras de status "novilhas" (H) e "vacas em lactação" (M), enquanto que H e M não diferem estatisticamente. No período de temperaturas mais amenas (P2) as doadoras D embora também com maior média diferem de H mas não diferem de M. De maneira geral, a obtenção de OT é favorecida em temperaturas menos elevadas, especialmente em doadoras de status H e M. Importante notar que os intervalos são maiores para o modelo combinado, pois leva em conta a variabilidade extra causada por superdispersão e correlaçao entre observações.

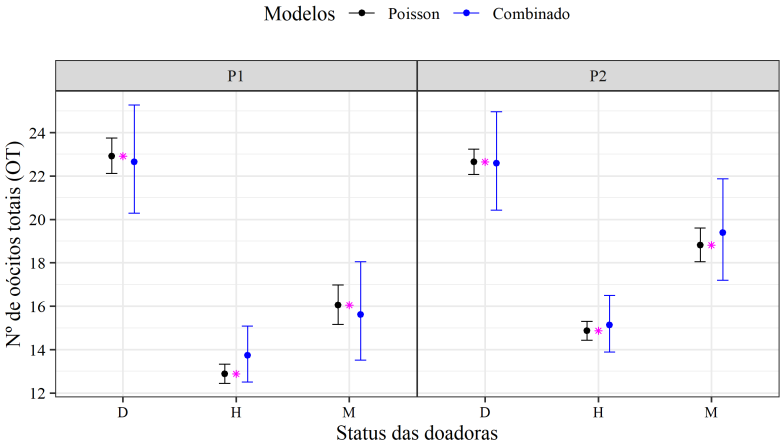

Figura 3.2. Variável resposta OT - Intervalos de confiança com 95% de confiança para médias marginais para as combinações dos níveis de períodos do ano (P1: Junho a Outubro e P2: sete meses restantes) e status das doadoras (H: novilhas, M: vacas em lactação e D: vacas secas) considerando os modelos Poisson (–) e combinado (M4MCM) (–) com o preditor linear de equação (3.26), as médias observadas são representadas pelo símbolo (∗).

3.7.2 Dados de proporção

Considera-se que Y_{jk} representa o número de embriões em estágio D3 obtido (valores observados y_{jk}) a partir de número n_{jk} de oócitos em IVC. O interesse está em saber como a taxa de clivagem, (y_{jk}/n_{jk}), varia de acordo com o período de coleta dos óvulos e o status da vaca.

Pode-se supor que Y_{jk} tem distribuição binomial, $Y_{jk} \sim \text{Binomial}(n_{jk}, \pi_{jk})$, beta-binomial, binomial-normal ou beta-binomial-normal, com função de ligação logística ($\eta_{jk} = \log[\pi_{jk}/(1-\pi_{jk})]$). Para os modelos logístico-normal e beta-binomial-normal (modelo combinado), consideram-se os seguintes preditores lineares

$$\eta_{jk} = \beta_0 + \tau_j + \xi_{d(jk)} + \psi_{s(jk)},\ d = 1, \ldots, 318 \text{ e } s = 1, \ldots, 66 \quad (3.27)$$

$$\eta_{jk} = \beta_0 + \tau_j + \delta_k + \xi_{d(jk)} + \psi_{s(jk)},\ d = 1, \ldots, 318 \text{ e } s = 1, \ldots, 66 \quad (3.28)$$

$$\eta_{jk} = \beta_0 + \tau_j + \delta_k + \gamma_{jk} + \xi_{d(jk)} + \psi_{s(jk)},\ d = 1, \ldots, 318 \text{ e } s = 1, \ldots, 66 \quad (3.29)$$

em que β_0 é a constante de efeito fixo, τ_j $(j = P1, P2)$ é o efeito fixo do j-ésimo período do ano, δ_k é o efeito fixo do k-ésimo status $(k = D, H, M)$ da fêmea doadora, γ_{jk} é o efeito fixo da interação entre os fatores, período do ano e status da fêmea doadora, $\xi_{d(jk)} \sim N(0, \sigma_d^2)$ é o efeito aleatório de doadora e $\psi_{s(jk)} \sim N(0, \sigma_s^2)$ é o efeito aleatório de touro. A inclusão do efeito aleatório de doadora e de touro é para acomodar a hierarquia, pois as doadoras foram aspiradas e touros forneceram sêmen em mais de uma ocasião e, portanto, suas observações têm alguma forma de dependência (Molenberghs et al., 2007). Para os modelos binomial e beta-binomial, eliminam-se $\xi_{d(jk)}$ e $\psi_{s(jk)}$ dos preditores lineares (3.27), (3.28) e (3.29).

De forma semelhante ao caso de contagens, pode-se fazer, inicialmente, a seleção da parte aleatória do modelo realizando testes sobre os efeitos aleatórios σ_d^2 e σ_s^2. Para

isso, foram ajustados os modelos binomial (M1) e binomial-normal (M2 a M4, variando os efeitos aleatórios incluídos) com o preditor linear descrito na equação (3.29). Em seguida, foram ajustados os modelos beta-binomial (M5) e beta-binomial-normal (M6 a M8, variando os efeitos aleatórios incluídos), também com o preditor linear descrito na equação (3.29). Os valores para as estatísticas $-2\log(L)$ e critério AIC para esses ajustes estão apresentados na Tabela 3.5.

Tabela 3.5. Resultados do ajuste de vários modelos aos dados de taxa de clivagem, usando os mesmos efeitos fixos no preditor linear de equação (3.29), mas variando os efeitos aleatórios incluídos.

Binomial versus binomial-normal					
Modelo	Parte fixa	Efeito aleatório	$-2\log(L)$	GL	AIC
M1	eq. (3.29)	nenhum	4852,0440	1142	4864,0440
M2	eq. (3.29)	Donor	4652,9560	1141	4666,9560
M3	eq. (3.29)	Sire	4785,3110	1141	4799,3110
M4	eq. (3.29)	Donor, Sire	4623,1990	1140	4639,1190
Beta-binomial versus beta-binomial-normal					
Modelo	Parte fixa	Efeito aleatório	$-2\log(L)$	GL	AIC
M5	eq. (3.29)	nenhum	4602,6880	1141	4616,6880
M6	eq. (3.29)	Donor	4569,9670	1140	4585,9670
M7	eq. (3.29)	Sire	4595,9370	1140	4611,9370
M8	eq. (3.29)	Donor, Sire	4562,6930	1139	4580,6930

$\log(L)$ é o logaritmo da função de verossimilhança e GL - número de graus de liberdade residual dos modelos ajustados.

No pacote **combTMB**, a variável resposta nos modelos apresentados na Tabela 3.5 pode ser especificada na forma de proporção (`prop ~ ..., weights = N`), ou na forma matricial, `cbind(sucesso,falha) ~ ...`, com o uso de duas colunas. Nas ilustrações, a forma matricial foi a preferida, pois exige menor volume de código. Os códigos para o ajuste dos modelos M1 a M4, são dados por:

```
##--Modelo binomial (MLG)
M1 <- combTMB(cbind(D3, IVC-D3) ~ Period*Status, embryos, family=binomial)
##--Modelo binomial-normal (MLGMs)
M2 <- combTMB(cbind(D3, IVC-D3) ~ Period*Status + (1|Donor), embryos,
                   family=binomial)
M3 <- combTMB(cbind(D3, IVC-D3) ~ Period*Status + (1|Sire), embryos,
                  family=binomial)
M4 <- combTMB(cbind(D3, IVC-D3) ~ Period*Status + (1|Donor) + (1|Sire),
              embryos, family=binomial)
```

Os códigos para ajustar os modelos M5 a M8 são apresentados no Apêndice A2.

Para comparar os modelos da Tabela 3.5 obtiveram-se as diferenças para os valores de $-2\log(L)$ para os modelos encaixados (Tabela 3.6) e aplicou-se o teste LRT ao nível de 5% de significância para testar os componentes de variância. Para testar apenas um componente de variância, por exemplo, o componente relacionado à doadoras ($H_0 : \sigma_d^2 = 0$), compara-se o valor calculado com uma mistura de qui-quadrado (equação (3.23)) dado por

$$\sum_{m=0}^{1} 2^{-1} \binom{1}{m} \chi_m^2 = \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 = 1,9210,$$

e para dois componentes ($H_0 : \sigma_1^2 = \sigma_2^2 = 0$), é

$$\sum_{m=0}^{1} 2^{-1} \binom{1}{m} \chi_m^2 = \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2 = 3,4190.$$

Tabela 3.6. Diferença nos valores da estatística $-2\log(L)$ para os modelos encaixados Tabela 3.5.

Binomial versus binomial-normal			
	M1	M2	M3
M2	199,0900 (1)		
M3	66,7330 (1)		
M4	228,9200 (2)	29,8370 (1)	162,1900 (1)
Beta-binomial versus beta-binomial-normal			
	M5	M6	M7
M6	32,7210 (1)		
M7	6,7514 (1)		
M8	39,9960 (2)	7,2743 (1)	33,2440 (1)

O teste LRT indica que é necessário incluir ambos os efeitos aleatórios nos modelos binomial-normal e beta-binomial-normal. Contudo, o modelo beta-binomial-normal (representado pelo modelo M8) apresenta os menores valores para as estatísticas $-2\log(L)$ e AIC, portanto, deve ser escolhido. Como próximo passo, pode-se ajustar o modelo beta-binomial-normal com os preditores lineares (3.27) e (3.28) e verificar se a parte fixa do preditor linear de equação (3.29) pode ser reduzida. No pacote **combTMB**, os modelos foram ajustados usando os seguintes comandos:

```
##--MCs
M9 <- combTMB(cbind(D3, IVC-D3) ~ Period + (1|Donor) + (1|Sire),
              embryos, family=betabinomial)
M10 <- combTMB(cbind(D3, IVC-D3) ~ Period + Status + (1|Donor) + (1|Sire),
               embryos, family=betabinomial)
```

Utilizando a função `AICtab` podem-se comparar os modelos M8, M9 e M10 (código no Apêndice A2). O modelo M8 especificado conforme o preditor linear de equação (3.29) é o que apresenta menor valor de AIC. No entanto, esse modelo apresenta uma diferença muito pequena, aproximadamente 1,4, quando comparado ao modelo M10. A escolha do modelo baseada somente no critério AIC em alguns casos pode levar o pesquisador a escolher um modelo menos parcimonioso. Nesse caso, pode-se empregar um teste *LRT* para complementar o processo de seleção da parte fixa. Para isso, o pacote **combTMB** possui a função `anova()`

```
R> anova(M10,M8)
Likelihood ratio test for combTMB regression models
 Model 1: cbind(D3, IVC - D3) ~ Period + Status + (1|Donor) + (1|Sire)
 Model 2: cbind(D3, IVC - D3) ~ Period * Status + (1|Donor) + (1|Sire)
           AIC    BIC Resid.df  logLik Chisq.df  Chisq Pr(>Chisq)
Model 1 4582.1 4617.4     1141 -2284.1
Model 2 4580.7 4626.1     1139 -2281.3        2 5.4078    0.06695 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

o valor $p = 0,06695$ indica que a interação `Period*Status` é não significativa e, portanto, o modelo M10 deve ser escolhido.

Sequencialmente, pode-se realizar a avaliação da qualidade de ajuste para o modelo M10 de forma semelhante ao caso de contagens. Inicialmente, utiliza-se a função `sanitycombTMB()` e, posteriormente, complementa-se com a análise de resíduos baseados em simulação, utilizando a função `dharma_residuals()`. A saída da função `sanitycombTMB()` não revelou anomalias no modelo e a análise de resíduos mostra os resíduos distribuídos de forma uniforme sobre a reta (Figura 3.3).

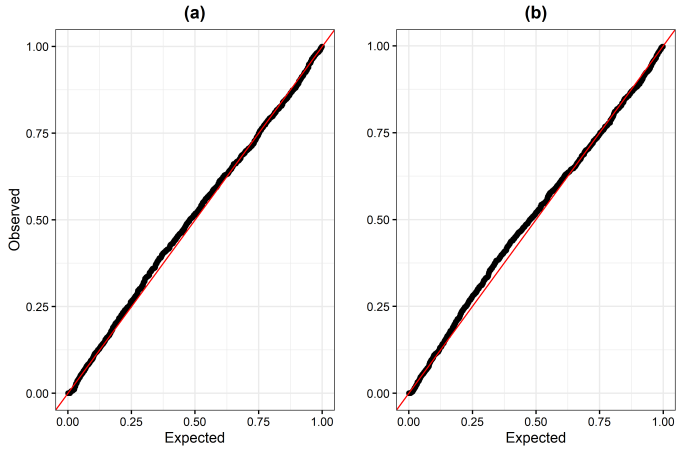

Figura 3.3. Resíduos quantílicos aleatorizados obtidos por meio da função `dharma_residuals()` para o modelo M10: (a) resíduos obtidos a partir de 250 simulações (b) Resíduos obtidos a partir de 500 simulações.

O resumo estatístico obtido com a função `summary()` para o modelo M10 possui a mesma estrutura que para o caso de contagem. Para complementar a análise, ajustaram-se os modelos binomial e beta-binomial (código no Apêndice A2), considerando apenas os efeitos fixos do preditor linear de equação (3.28).

As estimativas dos parâmetros com respectivos erros-padrão para os modelos binomial, beta-binomial, binomial-normal e combinado com preditor linear de equação (3.28) são apresentadas na Tabela 3.7. Nota-se, claramente, as diferenças em termos de erros-padrão das estimativas e significância dos parâmetros. Verifica-se que o modelo M10 tem os menores valores para as estatísticas $-2\log(L)$ e AIC e quando comparado com os modelos aninhados é significativamente superior. Portanto, o modelo combinado (M10) é o candidato mais viável, uma vez que os efeitos aleatórios normais σ_d^2 e σ_s^2 são responsáveis por modelar a correlação existente entre as medidas repetidas nos mesmos indivíduos (mais de uma sessão de OPU por doadora e efeito genético comum de alguns touros) e parte da superdispersão, enquanto que θ_{ij} seguindo uma distribuição Beta acomoda a maior parte da superdispersão (Kassahun et al., 2012; Molenberghs et al., 2012).

Tabela 3.7. Estimativas dos parâmetros, com seus respectivos erros-padrão (e.p.), para os modelos binomial, beta-binomial, binomial-normal e combinado com o preditor linear de equação (3.28), para a variável resposta taxa de clivagem.

	Binomial		Beta-binomial	
Parâmetros†	Estimativa (e.p.)		Estimativa (e.p.)	
Intercept	1,2889 (0,0377)	***	1,2944 (0,0526)	***
Period - P2	0,0814 (0,0377)	*	0,0808 (0,0508)	
status - H	$-0,1294$ (0,0407)	**	$-0,1334$ (0,0552)	*
status - M	$-0,3850$ (0,0506)	***	$-0,4044$ (0,0694)	***
Parâmetro-Superdispersão	—		0,0511	
$-2\log(L)$	4856,7000		4606,3000	
AIC	4864,7000		4616,3000	
	Binomial-normal		Combinado	
Parâmetros†	Estimativa (e.p.)		Estimativa (e.p.)	
Intercept	1,3038 (0,0824)	***	1,3060 (0,0737)	***
Period - P2	0,0614 (0,0498)		0,0732 (0,0548)	
status - H	$-0,1364$ (0,0858)		$-0,1244$ (0,0788)	
status - M	$-0,3760$ (0,0915)	***	$-0,3939$ (0,0887)	***
Parâmetro-Superdispersão	—		0,0306	
Donor (σ_d^2)	0,1740		0,1085	
Sire (σ_s^2)	0,0659		0,0181	
$-2\log(L)$	4630,8000		4568,1000	
AIC	4642,8000		4582,1000	

†Efeitos dos parâmetros são calculados usando a codificação padrão para contrastes, o primeiro valor é usado como referência; significância (***valor $p < 0,0001$, **valor $p < 0,01$ e *valor $p < 0,05$) e $\log(L)$ é o logaritmo da função de verossimilhança.

Usando-se o modelo combinado, existem evidências de que a interação entre os

fatores “Period” e “Status” não é significativa e, portanto, uma etapa natural para os pesquisadores é obter a média prevista para as médias marginais dos fatores “Period” e “Status” com os respectivos intervalos de confiança. Para isso o usuário pode utilizar a função `predict()`

```
data_pred <- predict(M10, re_form=~0, type="response", se_fit=TRUE)
```

Observe que nessa codificação estamos prevendo no modo de população, definindo os efeitos aleatórios iguais a zero por meio do argumento `re_form=~0`. A função retorna duas listas, uma com os valores preditos na escala da variável resposta e outra com os erros-padrão. O mesmo resultado poderia ser obtido definindo `re_form=NA`.

A Figura 3.4 mostra a média dos valores ajustados para os diferentes períodos do ano e status das doadoras, com intervalos de confiança de aproximadamente 90%, sugerindo que não há diferença na taxa de clivagem entre os períodos do ano. Contudo, em relação aos status das doadoras, a única diferença significativa foi observada para as doadoras com status M, apresentando uma taxa de clivagem menor conforme evidenciado na Tabela 3.7.

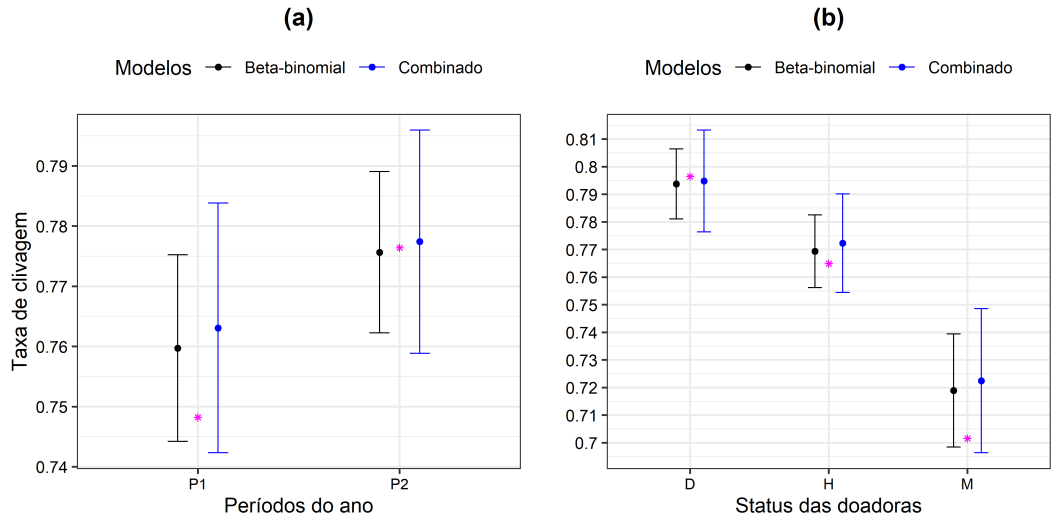

Figura 3.4. Variável resposta taxa de clivagem - Intervalos de confiança para as médias preditas dos níveis de períodos do ano (P1: Junho a Outubro e P2: sete meses restantes) e status das doadoras (H: novilhas, M: vacas em lactação e D: vacas secas) considerando os modelos beta-binomial (–) e combinado M10 (–) com o preditor linear de equação (3.28), as médias observadas são representadas pelo símbolo (∗).

3.8 Avaliação de tempo de processamento

Para estudo do desempenho do pacote **combTMB** foi feita uma avaliação comparativa (“benchmarking”) com o pacote **lme4**, pois é o principal pacote para análise de modelos com efeitos aleatórios. Os testes de desempenho foram executados em computador com sistema operacional Windows 10 Intel(R) Core i7 em um único núcleo e com auxílio do pacote **microbenchmark**.

Foram usados dois conjuntos de testes de desempenho. No primeiro cenário, foram simulados 10 conjuntos de dados com base no modelo M4, conservando a mesma estrutura dos dados originais. O tempo computacional foi medido 40 vezes para cada conjunto de dados, totalizando 400 ajustes. No segundo cenário, o conjunto de dados original foi replicado por 1, 2, 4, 6, 8 e 10 vezes para criar conjuntos de dados maiores, com mais observações por nível de efeito aleatório, porém mantendo o mesmo número de níveis de efeito aleatório. A esses conjuntos de dados foi ajustado o modelo Poisson-gama-normal (Seção 3.7.1) com preditor linear (3.26) usando o pacote **combTMB** e a função `glmer.nb()` do pacote **lme4**.

De acordo com a Figura 3.5, verifica-se que o pacote **combTMB** é mais rápido que o pacote **lme4** em ambos os cenários, para o ajuste dos modelos. Para o primeiro cenário, o pacote **combTMB** foi em média 5,2 vezes mais rápido que o **lme4** (Figura 3.5 (a)). No segundo cenário, o tempo de execução aumentou linearmente com o aumento do número de observações mas, também, foi menor para o pacote **combTMB** (Figura 3.5(b)).

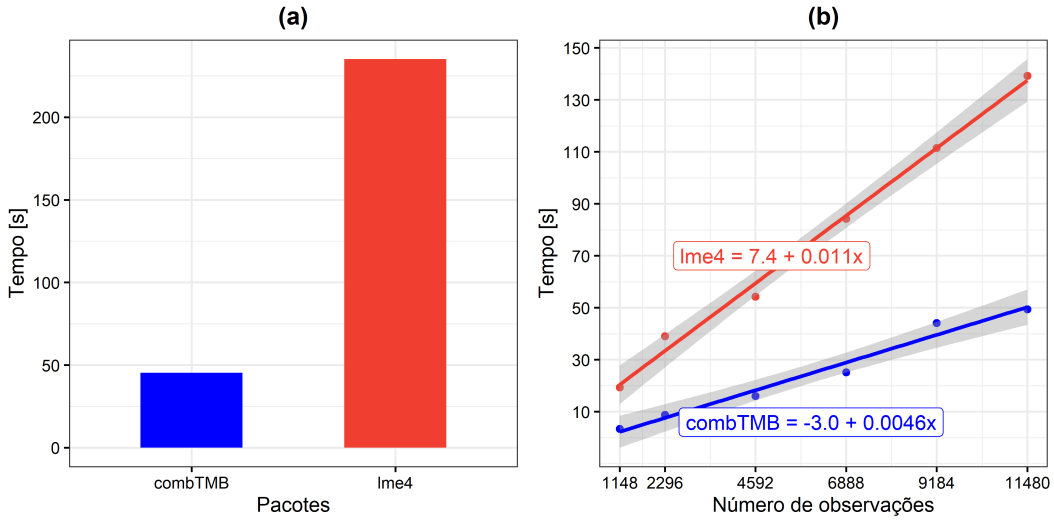

Figura 3.5. Avaliação comparativa: (a) tempo médio em segundos para o ajuste de 10 conjuntos de dados simulados com base no modelo M4 e (b) o conjunto de dados `embryos` foi replicado por 1, 2, 4, 6, 8, e 10 vezes para criar conjuntos de dados maiores.

3.9 Considerações finais

Este artigo descreve o pacote **combTMB**, desenvolvido em `R`, rápido e flexível para ajustar os MCs desenvolvidos por Molenberghs et al. (2007, 2010), que levam em conta a superdispersão e a dependência proveniente de agrupamentos, usando dois conjuntos distintos de efeitos aleatórios. Esses modelos enfatizam particularmente os chamados efeitos aleatórios conjugados ao nível da média para o primeiro aspecto e efeitos aleatórios normais incluídos no preditor linear para o segundo aspecto. Esse pacote permite, também, o ajuste dos tradicionais MLGs e MLGMs. Com essa flexibilidade em uma única ferramenta computacional, os modelos ajustados podem ser comparados usando métodos baseados

em probabilidade, incluindo critérios de informação. Nas aplicações apresentadas, os MCs foram superiores aos tradicionais MLGMs, tendo conseguido captar a variabilidade extra e a correlação entre as medidas longitudinais.

O pacote **combTMB** foi desenvolvido utilizando as ferramentas computacionais do **TMB**, e sua eficiência pode ser atribuída ao tratamento do parâmetro de dispersão pelo pacote **TMB** e à implementação da aproximação Laplace, para fazer a integração numérica em relação ao vetor de efeitos aleatórios. Os cálculos são projetados para serem rápidos em modelos com muitos parâmetros ($\approx 10^3$) e muitos efeitos aleatórios ($\approx 10^6$) (Kristensen et al., 2016), conforme evidenciado na Figura 3.8. Todavia, é importante destacar que o método de Laplace pode resultar em estimativas viesadas (Thygesen et al., 2017). Segundo Joe (2008), a ocorrência de viés assintótico está ligada à quantidade de discretização na variável resposta, sendo que há mais viés para respostas binárias do que para contagens, e diminui à medida que o tamanho do agrupamento aumenta.

No que diz respeito a problemas de convergência, o pacote **combTMB** é robusto para a maioria das aplicações, conforme mostrado pelo uso da função `sanitycombTMB()` nos exemplos das Seções 3.7.1 e 3.7.2. Entretanto, em modelos com alta dimensionalidade, uma única chamada para `stats::nlminb()` ou `stats::optim()` pode ser insuficiente para ocorrer a convergência (por exemplo, a grandeza do gradiente máximo da verossimilhança marginal em relação aos parâmetros de efeito fixo ainda não é pequena o suficiente), sendo necessário executar o algoritmo várias vezes. Outros problemas que podem ocorrer podem estar relacionados com instabilidade da matriz Hessiana.

Os problemas de convergência em sua grande maioria podem ser resolvidos usando-se o argumento `combTMB::combTMBcontrol()`, em que permite ao usuário modificar os valores padrões dos argumentos `nlminbLoops` e `newtonLoops`. O argumento `nlminbLoops`, reiniciará a otimização nos melhores valores anteriores. Assim, problemas relacionados com o gradiente máximo da função de verossimilhança marginal em relação aos parâmetros de efeitos fixos podem ser resolvidos com o aumento do número de "loops", por exemplo, para duas ou três vezes, sendo que o padrão é uma vez. Geralmente, valores de gradiente máximo $< 0,001$ indicam que o ajuste é consistente (`fit$Gradients`). O argumento `newtonLoops`, permite a reavaliação da matriz Hessiana com a rotina `stats::optimHess()` após o uso de `stats::nlminb()`. A função `stats::optimHess()` implementa a rotina de otimização de Newton para encontrar a matriz Hessiana. Por padrão, essa etapa adicional não está incluída e `newtonLoops` é definido como 0. É importante salientar que, em alguns casos, a não convergência pode ser uma indicação de que o modelo especificado está superparametrizado e pode ser necessário simplificá-lo.

Outra etapa importante da modelagem estatística é a de validação e seleção de modelos. Todavia, isso pode ser desafiador, particularmente, quando se usar a aproximação de Laplace em modelos que envolvem efeitos aleatórios, sejam eles normais ou não (Thygesen et al., 2017). Neste artigo, foram usadas diferentes estratégias com esse propó-

sito, dentre elas: (1) A função `sanitycombTMB()` que faz verificações iniciais no modelo ajustado, por exemplo, se a matriz Hessiana é positiva definida. (2) O Critério de Informação de Akaike (AIC, Akaike (1974)) que pode ser calculado com `AIC()`, caso o valor retornado seja `NA` significa que o modelo não convergiu. (3) Teste da razão de verossimilhanças empregando a função `anova` para modelos encaixados. (4) Resíduos quantílicos aleatorizados obtidos por meio de simulação com o emprego da função (`dharma_residuals()`) (Dunn e Smyth, 1996; Hartig, 2022). Os resultados do ajuste de um modelo usando o pacote **combTMB** podem ser passados para o pacote **tmbstan** (Monnahan e Kristensen, 2018) a fim de se realizarem inferências do ponto de vista Bayesiano.

Atualizações futuras do pacote **combTMB** poderão incluir a adição de outros modelos combinados, como Weibull e Exponencial para medidas repetidas na análise de sobrevivência com efeitos aleatórios gama e normais, modelos inflacionados de zeros, diagnóstico de influência, outras estruturas de variâncias-covariâncias, modelagem marginal para dados na forma de contagens e de proporções quando mais de um efeito aleatório normal está presente e geração de contagens correlacionadas e/ou superdispersas. O arquivo `combTMB_TMBExports.cpp` incluído, também, fornece um modelo de template de alta qualidade e testado que pode ser modificado para adicionar recursos adicionais.

Referências

Agresti, A. (2019). *An introduction to categorical data analysis*. John Wiley & Sons, Hoboken, NJ, 3 edition.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Bates, D., Mächler, M., Bolker, B., e Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1).

Bell, B. M. (2005). *CppAD: a package for C++ algorithmic differentiation*.

Blackford, L. S., Petitet, A., Pozo, R., Remington, K., Whaley, R. C., Demmel, J., Dongarra, J., Duff, I., Hammarling, S., Henry, G., et al. (2002). An updated set of basic linear algebra subprograms (BLAS). *ACM Transactions on Mathematical Software*, 28(2):135–151.

Bolker, B. e R Development Core Team (2022). *bbmle: Tools for General Maximum Likelihood Estimation*. R package version 1.0.25.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., e White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3):127–135.

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.

Consul, P. C. e Famoye, F. (1992). Generalized poisson regression model. *Communications in Statistics-Theory and Methods*, 21(1):89–109.

Cribari-Neto, F. e Zeileis, A. (2010). Beta regression in r. *Journal of Statistical Software*, 34:1–24.

Csárdi, G. (2022). *cli: Helpers for Developing Command Line Interfaces*. R package version 3.4.0.

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.

Demétrio, D. G. B., Benedetti, E., Demétrio, C. G. B., Fonseca, J., Oliveira, M., Magalhaes, A., e Santos, R. M. d. (2020). How can we improve embryo production and pregnancy outcomes of holstein embryos produced in vitro? (12 years of practical results at a california dairy farm). *Animal Reproduction*, 17(3).

Dunn, P. K. e Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244.

Greene, W. (2008). Functional forms for the negative binomial model for count data. *Economics Letters*, 99(3):585–590.

Griswold, M. E. e Zeger, S. L. (2004). On marginalized multilevel models and their computation.

Guennebaud, G., Jacob, B., et al. (2010). *Eigen v3*.

Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.4.6.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, 55(3):688–698.

Hedeker, D., du Toit, S. H. C., Demirtas, H., e Gibbons, R. D. (2018). A note on marginalization of regression parameters from mixed models of binary outcomes. *Biometrics*, 74(1):354–361.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

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

Huang, A. (2017). Mean-parametrized conway–maxwell–poisson regression models for dispersed counts. *Statistical Modelling*, 17(6):359–380.

Iddi, S. e Molenberghs, G. (2012). A combined overdispersed and marginalized multilevel model. *Computational Statistics & Data Analysis*, 56(6):1944–1951.

Irwin, J. O. (1968). The generalized waring distribution applied to accident theory. *Journal of the Royal Statistical Society: Series A (General)*, 131(2):205–225.

Joe, H. (2008). Accuracy of laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis*, 52(12):5066–5074.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C., e Verbeke, G. (2012). Modeling overdispersed longitudinal binary data using a combined beta and normal random-effects model. *Archives of Public Health*, 70:1–13.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., e Bell, B. M. (2016). TMB: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, 70(5):1–21.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 209–225.

Lee, Y., Nelder, J. A., e Pawitan, Y. (2017). *Generalized linear models with random effects: unified analysis via H-likelihood*, volume 153. CRC Press.

Lenth, R. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.1-1.

McCullagh, P. e Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, Boca Raton London New York.

Molenberghs, G. e Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag GmbH.

Molenberghs, G., Verbeke, G., e Demétrio, C. G. B. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, 13(4):513–531.

Molenberghs, G., Verbeke, G., e Demétrio, C. G. B. (2017). Hierarchical models with normal and conjugate random effects: a review. *Statistics and Operations Research Transactions*, 41(2):191–253.

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

Molenberghs, G., Verbeke, G., Iddi, S., e Demétrio, C. G. B. (2012). A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. *Journal of Multivariate Analysis*, 111:94–109.

Monnahan, C. C. e Kristensen, K. (2018). No-u-turn sampling for fast bayesian inference in admb and tmb: Introducing the adnuts and tmbstan r packages. *PloS ONE*, 13(5):e0197954.

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

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A., Olmo-Jiménez, M. J., e Martínez-Rodríguez, A. M. (2009). A generalized waring regression model for count data. *Computational Statistics & Data Analysis*, 53(10):3717–3725.

SAS Institute (2014). *SAS 9.4 language reference concepts*. SAS Institute, Cary, NC.

Stroup, W. W. (2016). *Generalized linear mixed models : modern concepts, methods and applications*. Taylor & Francis Ltd.

Thygesen, U. H., Albertsen, C. M., Berg, C. W., Kristensen, K., e Nielsen, A. (2017). Validation of ecological state space models using the laplace approximation. *Environmental and Ecological Statistics*, 24(2):317–339.

Venables, W. N. e Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer-Verlag GmbH.

Vílchez-López, S., Sáez-Castillo, A. J., e Olmo-Jiménez, M. J. (2016). GWRM: An r package for identifying sources of variation in overdispersed count data. *PloS ONE*, 11(12):e0167570.

Zeger, S. L., Liang, K.-Y., e Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060.

Zhang, D. e Lin, X. (2008). Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In *Random Effect and Latent Variable Model Selection*, pages 19–36. Springer New York.

4 MODELAGEM DA HORMESIS POR REGRESSÃO NÃO LINEAR MULTIVARIADA

Resumo

Há mais de um século, os estudos ecotoxicológicos relatam a ocorrência de hormesis como um fenômeno significativo em muitas áreas da ciência. Na biologia vegetal, a pesquisa em hormesis concentra-se em mensurar alterações morfológicas, fisiológicas, bioquímicas e produtivas em plantas submetidas a baixas doses de herbicidas, ou seja, são estudos que envolvem múltiplas características, geralmente, correlacionadas. Todavia, o aspecto multivariado e as interdependências existentes entre os componentes de um sistema vegetal não são considerados na estrutura de modelagem adotada. Nesse sentido, propõe-se a modelagem não linear multivariada da hormesis, em que as informações concernentes às correlações das variáveis respostas são consideradas com uma matriz de variâncias e covariância obtida, a partir dos resíduos univariados. A metodologia proposta é avaliada por meio de um estudo de simulação Monte Carlo e uma aplicação em dados experimentais da cultura do cártamo (*Carthamus tinctorius* L.). No estudo de simulação, o modelo multivariado foi mais eficaz, apresentando-se, com maior precisão, menor viés e com maior acurácia das estimativas dos parâmetros, quando comparado aos ajustes dos modelos univariados. Esses resultados foram confirmados, também, para os dados experimentais. Por meio do método delta, doses médias das quantidades de interesse podem ser derivadas com seus erros-padrão associados. Este é o primeiro estudo que aborda a hormesis no contexto multivariado, possibilitando uma maior compreensão das relações bifásicas de dose-resposta, pois considera na estrutura de modelagem as inter-relações existentes entre as diversas características mensuradas no sistema vegetal, levando a estimativas mais precisas dos parâmetros .

Palavras-chave: Análise multivariada; Modelos não lineares; Curva de dose-resposta; Biologia vegetal; *Carthamus tinctorius* L.; Herbicidas.

4.1 Introdução

Há mais de um século, estudos ecotoxicológicos evidenciam o efeito estimulador de baixas doses de produtos químicos ou outros estressores, com efeitos benéficos no crescimento e desenvolvimento de organismos vivos como várias espécies de plantas, fungos e bactérias (Calabrese, 2005; Cedergreen, 2008). Esse fenômeno é conhecido pelo termo hormesis, estabelecido na década de 1940, por Southam e Erlich (Southam, 1943). Nas plantas, os herbicidas são os maiores responsáveis pela ocorrência da hormesis, com destaque para o herbicida mais utilizado no mundo, o glifosato (Cedergreen, 2008; Duke, 2019; Belz, 2020; Santos et al., 2021).

A maioria dos estudos em biologia vegetal que abordam hormesis, concentram-se em mensurar alterações morfológicas em plantas submetidas a baixas doses de glifosato e, em menor escala, mudanças fisiológicas e bioquímicas (Belz e Piepho, 2013; Pincelli-Souza et al., 2020; Santos et al., 2021; Belz e Duke, 2022), com reflexos na produção (Pincelli-Souza et al., 2020; Bortolheiro e Silva, 2021). Em geral, os estudos de hormesis são multicaracterística, e reconhecer os casos em que a hormesis realmente existe e, estabelecer, adequadamente, o significado desse fenômeno bifásico de dose-resposta (Belz e Piepho, 2012), constitui o principal desafio para os pesquisadores. Para isso, a modelagem estatística da curva completa de dose-resposta, é a melhor opção para prever quantitativamente os efeitos sub-NOAEL("No Observed Adverse Effect Level" ou NOAEC referente à concentração) (Belz e Duke, 2022). Segundo Belz e Duke (2022), a técnica de modelagem com maior destaque nos últimos dois anos para investigar hormesis foram os modelos de regressão não lineares.

O emprego dos modelos não lineares em estudos de tendências de dose-resposta após aplicações de herbicidas popularizaram-se na década de 80, principalmente, com os trabalhos de Streibig em plantas daninhas (Streibig, 1980, 1988). Nesse período, a hormesis era comentada apenas como discrepante em relação à curva sigmoide de dose-resposta (Streibig, 1980). Somente em 1989 surgiria o modelo de Brain e Cousens (1989) com a capacidade de modelar uma curva bifásica de dose-resposta, permitindo hormesis. Posteriormente, em 2005, surgiu o modelo de Cedergreen et al. (2005), sendo uma extensão do modelo de Brain e Cousens (1989). Desde então, os esforços concentram-se em obter funções reparametrizadas desses modelos (Schabenberger et al., 1999; Belz e Piepho, 2012, 2015; Belz e Duke, 2022), úteis para permitir inferência sobre quantidades de interesse. Um fato relevante é que todos os avanços estatísticos conseguidos nessa área, limitam-se ao contexto univariado.

Embora os projetos experimentais elaborados para estudar hormesis mensurem um conjunto de características morfológicas, fisiológicas e bioquímicas, geralmente, correlacionadas (Belz, 2018; Pincelli-Souza et al., 2020; Santos et al., 2022), o aspecto multivariado dos dados é negligenciado na maioria dos estudos. Ao avaliar o efeito de uma variável por vez, sem considerar a interdependência entre elas, pode resultar muitas vezes em conclusões parciais, por exemplo, pode-se detectar hormesis quando a altura das plantas é avaliada, e não se observar o mesmo para massa seca da parte aérea. Como Türkşen (2021) afirmou, se as respostas são correlacionadas, não devem ser investigadas individualmente e independentemente umas das outras. Assim, considerar que os componentes de uma mesma planta são independentes não reflete a realidade.

Na abordagem univariada não é possível inferir do ponto de vista global ou multivariado sobre a existência da hormesis, ou seja, quando todas as características são consideradas, pois não é possível realizar um teste estatístico formal confrontando as hipóteses de existência ou não existência de hormesis. Outra questão relevante é a correta

previsão da magnitude e a faixa de dosagem dos aumentos horméticos (Belz e Duke, 2014), por exemplo, obter a dosagem média das quantidades sub-NOAEL com base em ajustes univariados das diferentes variáveis resposta, pode resultar em erros-padrão muito grandes, reduzindo a credibilidade das estimativas. Essas limitações dificultam, por exemplo, o aproveitamento dos efeitos horméticos benéficos dos herbicidas nas culturas. O único exemplo de uso comercial eficiente da hormesis é na cana-de-açúcar (*Saccharum officinarum* L.), em que baixas doses do herbicida glifosato aumentam a produção de açúcar (Belz e Duke, 2014; Pincelli-Souza et al., 2020).

Diante das lacunas apresentadas, propõe-se a modelagem não linear multivariada da hormesis (MNMH), com base nos trabalhos de Zellner (1962) e Gallant (1987), em que as informações concernentes às correlações das respostas são levadas em consideração com uma matriz de variâncias e covariância obtida a partir dos resíduos univariados. Com isso, o conceito de hormesis pode ser ampliado permitindo recomendações de dosagens efetivas com base nas múltiplas características avaliadas, mas mantendo todas as informações dos ajustes univariados. O leque de aplicações possíveis se estende a qualquer estudo ecotoxicológico com estrutura de tratamento quantitativo.

O artigo está organizado da seguinte forma. Na Seção 4.2, as principais características da hormesis em biologia vegetal e principais modelos não lineares utilizados são apresentados. Os procedimentos de estimação e inferência da MNMH são abordados na Seção 4.3. O desempenho da metodologia proposta é avaliada por um estudo de simulação Monte Carlo, na Seção 4.4, enquanto que uma aplicação a dados reais é apresentada para ilustrar os benefícios da MNMH, na Seção 4.5. Na Seção 4.6, são apresentadas as principais razões para adoção da abordagem MNMH. Finalmente, a Seção 4.7 apresenta as conclusões.

4.2 Características importantes no estudo de hormesis

Na biologia vegetal, os modelos univariados não lineares mais utilizados para modelagem da curva bifásica de dose-resposta em forma de "U" invertido para quantificar hormesis, são os propostos por Brain e Cousens (1989) e Cedergreen et al. (2005), com cinco e seis parâmetros, respectivamente, dados por

$$\mathrm{E}(Y|x) \;=\; c + \frac{d - c + fx}{1 + \exp[b\log(x/e)]} \tag{4.1}$$

e

$$\mathrm{E}(Y|x) \;=\; c + \frac{d - c + f\exp(-1/x^a)}{1 + \exp[b\log(x/e)]} \tag{4.2}$$

em que $\mathrm{E}(Y|x)$ representa a média da variável resposta na dosagem x; c (assíntota) denota a resposta para doses tendendo a infinito; d é a resposta média do controle (não tratado, $x = 0$); f é o tamanho do efeito de hormesis: quanto maior seu valor, maior é o efeito de

hormesis ($f = 0$ corresponde a nenhum efeito de hormesis); a e b determinam a intensidade da inclinação da curva antes e depois do efeito máximo de hormesis, respectivamente. O parâmetro e não tem significado biológico evidente.

Ao longo dos anos, esses dois modelos se consolidaram para modelagem de hormesis, pois incluem parâmetros com interpretação biológica e permitem reparametrizações, de modo que todas as características quantitativas sub-NOAEL possam ser obtidas, contribuindo, efetivamente, para compreensão do fenômeno (Belz e Piepho, 2012, 2015; Belz e Duke, 2022). Nos últimos dois anos a maioria das publicações científicas que abordam o tema hormesis utilizaram esses modelos, sendo que o modelo de Brain e Cousens (1989) de equação (4.1), tem sido o preferido (Belz e Duke, 2022).

Dentre as principais características quantitativas sub-NOAEL que descrevem a expressão de uma resposta hormética destacam-se, a dose ED_{110} (igual a NOEL "No-Observed-Effect-Level") que causa 10% de estimulação sobre o nível de controle (não tratado), a dose M (ou MAX_x ou EC_{max}), y_{max} (ou MHSR "maximum hormetic stimulation response" ou MAX_y ou 'max') que é a resposta máxima da hormesis na dose M, frequentemente, expressa na forma de valor relativo y_{max} (% do y_{controle}), ou estimulação relativa acima do controle, a dose LDS ("Limited-Dose-for-Stimulation") ou NOAEL ("No Observed Adverse Effect Level") ou ZEP ("Zero Equivalent Point") em que o efeito hormético desaparece e ED_{50} a dose que causa a redução de 50% na resposta média do controle (não tratado).

No entanto, as características quantitativas sub-NOAEL mencionadas não podem ser obtidas diretamente a partir dos modelos originais de equações (4.1) e (4.2). Schabenberger et al. (1999) reparametrizaram o modelo de Brain e Cousens (1989), permitindo a obtenção das estimativas de M, LDS e doses arbitrárias ED_k, como funções dos parâmetros do modelo original, dadas por

$$\mathrm{E}(Y|x) = c + \frac{d - c + fx}{1 + \left(\frac{fM}{(d-c)b - fM(1-b)}\right)\exp[b\log(x/M)]}, \tag{4.3}$$

$$\mathrm{E}(Y|x) = c + \frac{d - c + fx}{1 + \left(\frac{fLDS}{d-c}\right)\exp[b\log(x/LDS)]}, \tag{4.4}$$

e

$$\mathrm{E}(Y|x) = c + \frac{d - c + fx}{1 + \left(\frac{k}{100-k} + \frac{100}{100-k} \times \frac{fED_k}{d-c}\right)\exp[b\log(x/ED_k)]}. \tag{4.5}$$

Funções similares foram obtidas por Belz e Piepho (2012) para o modelo de Cedergreen et al. (2005).

As funções reparametrizadas evidenciam que a obtenção das características quantitativas sub-NOAEL demanda vários ajustes individuais para cada característica mensurada em uma planta e/ou grupo de plantas. No entanto, esse procedimento não incorpora informações sobre a interdependência biológica existente entre as características men-

suradas. Mostra-se, a seguir, o procedimento geral para obter um modelo não linear multivariado.

4.3 Modelo não linear multivariado

Considere um modelo não linear univariado escrito como

$$Y_i = f(\boldsymbol{x}_i, \boldsymbol{\theta}) + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{4.6}$$

em que n é o número de observações, f é uma função não linear que depende do vetor de covariáveis $\boldsymbol{x}$ de dimensão $m \times 1$, $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^T$ é um vetor de parâmetros desconhecidos e ϵ_i é o componente aleatório independente e identicamente distribuído (i.i.d.). Uma extensão natural de (4.6) é assumir que existe um conjunto de w respostas cujas observações têm comportamentos não lineares, dependendo das variáveis regressoras $\boldsymbol{x}$, isto é,

$$Y_{ik} = f_k(\boldsymbol{x}_k, \boldsymbol{\theta}_k) + \epsilon_{ik}, \quad i = 1, 2, \ldots, n, \quad k = 1, 2, \ldots, w, \tag{4.7}$$

em que $\boldsymbol{\theta}_k$ é um vetor de parâmetros com dimensão $p_k \times 1$, $\boldsymbol{x}_k$ é um vetor de covariáveis e Y_{ik} é a resposta do elemento i da variável k e ϵ_{ik} é o erro. Portanto, o modelo para a k-ésima variável pode ser representado na forma vetorial como

$$\boldsymbol{Y}_k = f_k(\boldsymbol{\theta}_k) + \boldsymbol{\epsilon}_k, \quad k = 1, 2, \ldots, w. \tag{4.8}$$

O conjunto das w regressões pode ser arranjado de forma conveniente em uma única regressão na forma vetorial

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{\Theta}) + \boldsymbol{e}, \tag{4.9}$$

em que $\boldsymbol{Y} = [\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T, \ldots, \boldsymbol{Y}_w^T]^T$ é o vetor de respostas com dimensão $nw \times 1$, $\boldsymbol{f}(\boldsymbol{\Theta}) = [f_1(\boldsymbol{\theta}_1),\ f_2(\boldsymbol{\theta}_2),\ \ldots,\ f_w(\boldsymbol{\theta}_w)]^T$ é o vetor de funções de regressões não lineares, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \ldots, \boldsymbol{\theta}_w^T]^T$ de dimensão $\sum_{k=1}^{w} p_k \times 1$ é o vetor de parâmetros e $\boldsymbol{e} = [\boldsymbol{\epsilon}_1^T, \boldsymbol{\epsilon}_2^T, \ldots, \boldsymbol{\epsilon}_w^T]^T$ é o vetor de erros, respectivamente, com dimensão $nw \times 1$. A matriz de variâncias e covariância para o modelo multivariado de regressão não linear, dado em (4.9), é

$$\mathrm{E}[\boldsymbol{e}\boldsymbol{e}^T] = \boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes I_n, \tag{4.10}$$

em que $\boldsymbol{\Sigma} = [\sigma_{rk}]$ $r, k = 1, 2, \ldots, w$, é uma matriz simétrica positiva definida, sendo σ_{kk} a variância dos erros para a variável k e σ_{rk} a covariância entre os erros para as variáveis k e r, I_n é uma matriz identidade, $n \times n$, e $\otimes$ denota o produto de Kronecker.

O procedimento de estimação dos parâmetros do modelo não linear multivariado pode ser feito em três etapas.

Etapa I: Ajuste individual dos modelos de regressão não linear.

As estimativas dos parâmetros para cada regressão não linear, $\hat{\boldsymbol{\theta}}_k$, $k = 1, 2, \ldots, w$, dada na equação (4.8), são obtidas pelo método dos mínimos quadrados não lineares (NLS

"nonlinear least squares"), isto é, minimizando-se a soma do quadrado dos erros em relação a $\boldsymbol{\theta}_k$,

$$S(\boldsymbol{\theta}_k)^{NLS} = [\boldsymbol{Y}_k - f_k(\boldsymbol{\theta}_k)]^T[\boldsymbol{Y}_k - f_k(\boldsymbol{\theta}_k)]. \tag{4.11}$$

Entretanto, esse sistema de equações normais não lineares não tem solução explícita e processos iterativos, por exemplo, o método de Marquardt (Gallant, 1987; Bates e Watts, 1988), devem ser usados.

Etapa II: Obtenção da estimativa da matriz de variâncias e covariâncias dos resíduos.

Inicialmente, obtêm-se as estimativas dos erros para cada variável resposta ajustada, isto é, os resíduos

$$\hat{\boldsymbol{\epsilon}}_k = \boldsymbol{Y}_k - f_k(\hat{\boldsymbol{\theta}}_k), \quad k = 1, 2, \ldots, w. \tag{4.12}$$

A seguir, para a obtenção da estimativa da matriz de variâncias e covariâncias, $\hat{\boldsymbol{\Sigma}}$ (Zellner, 1962; Gallant, 1987; Seber e Wild, 2003), calculam-se os elementos

$$\hat{\sigma}_{rk} = \frac{\hat{\boldsymbol{\epsilon}}_r^T \hat{\boldsymbol{\epsilon}}_k}{n} \quad r, k = 1, 2, \ldots, w. \tag{4.13}$$

Etapa III: Composição do modelo não linear multivariado e obtenção da otimização única.

Essa etapa consiste em obter as estimativas de $\boldsymbol{\Theta}$, denotado por $\hat{\boldsymbol{\Theta}}$. Para isso, é necessário minimizar a função objetivo em relação a $\boldsymbol{\Theta}$

$$S(\boldsymbol{\Theta}, \boldsymbol{\Omega}) = [\boldsymbol{Y} - \boldsymbol{f}(\boldsymbol{\Theta})]^T \boldsymbol{\Omega}^{-1} [\boldsymbol{Y} - \boldsymbol{f}(\boldsymbol{\Theta})], \tag{4.14}$$

em que $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}^{-1} \otimes I_n$. Entretanto, a matriz $\boldsymbol{\Sigma}$ é desconhecida e Gallant (1987) sugere a sua substituição pelo estimador consistente, $\hat{\boldsymbol{\Sigma}} = [\hat{\sigma}_{rk}]$, dado em (4.13). Desse modo, pode-se reescrever a função objetivo (4.14) como

$$S(\boldsymbol{\Theta}, \boldsymbol{\Sigma}) = [\boldsymbol{Y} - \boldsymbol{f}(\boldsymbol{\Theta})]^T (\hat{\boldsymbol{\Sigma}}^{-1} \otimes I_n) [\boldsymbol{Y} - \boldsymbol{f}(\boldsymbol{\Theta})]. \tag{4.15}$$

Segundo Seber e Wild (2003), a minimização de (4.15) é conhecida como método de mínimos quadrados generalizados (GLS "Generalized least squares"). No entanto, usando-se a decomposição de Cholesky da inversa da estimativa da matriz de variâncias e covariâncias $\hat{\boldsymbol{\Sigma}}^{-1} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^{\mathbf{T}}$, em que $\hat{\boldsymbol{\Lambda}}$ é uma matriz triangular superior, chamada de fator de Cholesky, o modelo (4.9) pode ser rotacionado

$$(\hat{\boldsymbol{\Lambda}} \otimes I_n)\boldsymbol{Y} = (\hat{\boldsymbol{\Lambda}} \otimes I_n)\boldsymbol{f}(\boldsymbol{\Theta}) + (\hat{\boldsymbol{\Lambda}} \otimes I_n)\boldsymbol{e} \tag{4.16}$$

e simplificado para

$$\boldsymbol{Y}^* = \boldsymbol{f}(\boldsymbol{\Theta})^* + \boldsymbol{e}^*, \tag{4.17}$$

em que $\boldsymbol{e}^* = (\hat{\boldsymbol{\Lambda}} \otimes I_n)\boldsymbol{e}$.

Pode-se mostrar que $\mathrm{E}[\boldsymbol{e}^*\boldsymbol{e}^{*T}] = I_n w$, e que o modelo em (4.17) é um modelo não linear univariado, o que permite que o método NLS seja utilizado para o ajuste do modelo não linear multivariado. Portanto, o vetor de parâmetros $\boldsymbol{\Theta}$ pode ser estimado pela minimização de

$$S(\boldsymbol{\Theta}, \boldsymbol{\Sigma}) \;\; = \;\; [\boldsymbol{Y}^* - \boldsymbol{f}(\boldsymbol{\Theta})^*]^T [\boldsymbol{Y}^* - \boldsymbol{f}(\boldsymbol{\Theta})^*]. \tag{4.18}$$

Para ajustar os modelos, dados nas equações (4.11 e 4.18), o procedimento iterativo de Levenberg-Marquardt pode ser aplicado, utilizando-se a linguagem de computação estatística R (R Core Team, 2022), juntamente com a biblioteca `minpack.lm` (Elzhov et al., 2023) ou o procedimento NLIN do sistema SAS (SAS Institute Inc, 2015).

4.3.1 Inferência estatística e seleção de modelos

A normalidade multivariada dos dados ou pelo menos que cada uma das variáveis individuais tenha normalidade univariada é a principal suposição estabelecida para inferência sobre os parâmetros do modelo não linear multivariado. Nesse sentido, um dos principais testes utilizados, é o teste de Mardia, que avalia assimetria e curtose, e o teste de Shapiro-Wilk para normalidade univarida (Shapiro e Wilk, 1965; Mardia, 1974). Para testar, também, a hipótese H_0 de que a matriz de correlação é igual à matriz identidade, ou seja, não existe correlação entre as variáveis, pode-se usar o teste de esfericidade de Bartlett (1951), cuja estatística é dada por

$$\mathrm{T}_{obs} = -\log[\det(\boldsymbol{R})][n - 1 - (2w + 5)/6] \sim \chi^2_\nu,$$

em que $\boldsymbol{R}$ é a matriz de correlação estimada, obtida a partir da matriz $\hat{\boldsymbol{\Sigma}}$, w é o número de equações univariadas, $\nu = w(w-1)/2$ e n o número de observações. Rejeita-se H_0, a um nível de significância α, se $T_{obs} \geq \chi^2_\alpha(\nu)$.

Para testar se existe efeito de hormesis ($H_0 : f = 0$) nos modelos tradicionais dados nas equações (4.1) e (4.2), duas abordagens têm sido utilizadas. A primeira, envolve a construção do intervalo de confiança, geralmente, com 95% de confiança (IC_{95}) para f, isto é,

$$\hat{f} \pm z_\alpha/2\sqrt{V(\hat{f})},$$

em que $z_\alpha/2$ é o quantil da distribuição normal padrão para um nível de confiança $1-\alpha$ e $V(\hat{f})$ é a variância do estimador. Rejeita-se H_0, ao nível de 5% de significância, se o IC_{95} não contiver o zero ($f > 0$) (Schabenberger et al., 1999; Belz e Piepho, 2012) e nesse caso indicando a não ocorrência de hormesis.

A segunda abordagem, é por meio do teste da razão de verossimilhanças (LRT "likelihood-ratio test") para modelos encaixados, sob a suposição de normalidade dos dados (Gallant, 1987)

$$LRT = -2[\text{logLik(modelo reduzido) - logLik(modelo completo)}],$$

em que logLik é o logaritmo da função de verossimilhança. O modelo completo é um dos modelos de equações (4.1) e (4.2) e o modelo reduzido, chamado modelo de Streibig (1988), é o modelo sob H_0, isto é, com equação

$$\mathrm{E}(Y|x) = c + \frac{d-c}{1+\exp[b\log(x/e)]},$$

com o parâmetro e interpretado como ED_{50}.

Nesse caso, o teste LRT equivale a um teste F com estatística dada por

$$F_{obs} = \frac{\text{SQRes(modelo reduzido)} - \text{SQRes(modelo completo)}}{\text{QMRes(modelo completo)}} \sim F_{1,n-p}$$

em que SQ é a soma de quadrados dos resíduos e QM $= \frac{\text{SQ}}{n-p}$.

No caso do modelo não linear multivariado, pode-se testar a significância do efeito de hormesis para cada variável por meio do IC_{95} ou testar de forma simultânea o vetor $\boldsymbol{\theta}_f = [f_1,\ f_2, \ldots, f_w]^T$ por meio do teste LRT. Para o teste ajustam-se os modelos completos e reduzidos multivariados, seguindo os três passos mencionados na Seção 4.3 e calcula-se a estatística do teste é

$$F_{obs} = \frac{[S(\tilde{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}})_r - S(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}})_c]/\nu_d}{S(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}})_c/\nu_c} \sim F(\nu_d; \nu_c),$$

em que $S(\tilde{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}})_r$ e $S(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}})_c$ são as somas de quadrados dos modelos reduzido e completo, respectivamente, $\nu_d = p - q$ é a diferença em nos números de graus de liberdade dos dois modelos, sendo q e p os números de parâmetros dos modelos reduzido e completo e $\nu_c = nw - p$ é o número de graus de liberdade do modelo completo (w: número de equações univariadas; n: número de observações). Rejeita-se a hipótese H_0 de que o modelo reduzido é adequado, a um nível de significância α, se $F_{obs} \geq F_\alpha(\nu_d; \nu_c)$.

Outras estatísticas comumente usadas são a raiz do erro quadrático médio (REQM), que mede a acurácia das estimativas, e a média do viés absoluto (MVA), que fornece uma indicação geral da variância do modelo, isto é,

$$\begin{aligned} REQM &= \left(\frac{\sum_i^n (Y_i - \hat{Y}_i)^2}{n}\right)^{1/2} \\ MVA &= \frac{\sum_i^n |Y_i - \hat{Y}_i|}{n}, \end{aligned}$$

em que Y_i e $\hat{Y}_i$, $i = 1, 2, \ldots, n$, são os valores observados e preditos, respectivamente, e n o número de observações.

4.3.2 Características quantitativas médias sub-NOAEL

Do ponto de vista prático, além das estimativas M, LDS e ED_{50} individuais para cada característica, é interessante obter uma dosagem média com base em todas as

características avaliadas em um sistema vegetal. Desse modo, é possível realizar recomendações para o uso eficiente dos benefícios da hormesis. Para algumas culturas existem algumas recomendações de uso do herbicida glifosato, dentre elas, pode-se destacar: a cevada (*Hordeum vulgare* L.) com dose limite <63 g equivalente ácido (e.a.) ha^{-1} para promover o crescimento real (Cedergreen, 2008), na cana-de-açúcar a dose de 1,8 g e.a. ha^{-1} teve, efeitos benéficos ao longo do ciclo promovendo o crescimento e desenvolvimento (Pincelli-Souza et al., 2020), e no cártamo com base em variáveis morfológicas com 21 g e.a. ha^{-1} promoveu máximo rendimento (Santos et al., 2021).

Ajustado o MNMH, pode-se, então, obter as características quantitativas médias sub-NOAEL, por exemplo, a resposta máxima média da hormesis dada pela dose $\overline{M}$. Nesse contexto, para obter o erro padrão para dose $\overline{M}$ e, consequentemente, o intervalo de confiança, considere que os estimadores $\hat{M}_1, \hat{M}_2, \ldots, \hat{M}_w$ são variáveis aleatórias e $\hat{\varphi} = g(\hat{M}_1, \hat{M}_2, \ldots, \hat{M}_w)$ é a função de interesse. De forma aproximada, utilizando o método delta, o valor esperado e sua variância para $\hat{\varphi}$ são dados por

$$\begin{aligned} \mathrm{E}(\hat{\varphi}) &= g(\boldsymbol{M}), \\ \mathrm{Var}(\hat{\varphi}) &\approx \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^T, \quad \text{em que} \quad \boldsymbol{D} = \left.\frac{\partial g(\boldsymbol{M})}{\partial \boldsymbol{M}^T}\right|_{\boldsymbol{M}=\widehat{\boldsymbol{M}}}. \end{aligned}$$

Nas aplicações a matriz de variâncias e covariâncias estimada $\hat{\boldsymbol{\Sigma}}$ é usada no lugar $\boldsymbol{\Sigma}$.

Sob a suposição de normalidade multivariada, um intervalo de confiança, com cobertura nominal $1-\alpha$ para φ, é obtido por

$$\mathrm{E}(\hat{\varphi}) \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\varphi})},$$

em que $z_{1-\alpha/2}$ é o quantil $1-\alpha/2$ da distribuição normal padrão e $\mathrm{Var}(\hat{\varphi})$ é a variância do estimador.

De forma semelhante, podem ser obtidos erros-padrão para as estatísticas das características quantitativas médias sub-NOAEL, $\overline{LDS}$ e $\overline{ED}_{50}$, com os respectivos intervalos de confiança.

4.4 Estudo de simulação

4.4.1 Descrição

Um estudo de simulação Monte Carlo foi realizado para comparar a performance dos ajustes de modelos univariados e um modelo bivariado não linear bifásico de dose-reposta para o estudo de hormesis sob diferentes cenários, considerando a estrutura de um experimento em delineamento inteiramente casualizado.

Para isso, foram simulados dados de experimentos com diferentes número de repetições ($r = 4$, $r = 6$ e $r = 8$) e dez doses (0; 1,8; 3,6; 7,2; 18; 36; 72; 180; 360

e 720 g e.a. ha^{-1}) de um agente químico, por exemplo, o herbicida glifosato. Foram assumidos diferentes níveis de correlação ($\rho = 0,3; 0,5; 0,7; 0,9$) entre os resíduos das variáveis respostas. Para cada um dos 12 cenários, consideraram-se 1000 réplicas.

Os conjuntos de dados foram gerados de acordo com o modelo de Brain e Cousens (1989), com

$$\begin{aligned} f_k(x_{ij}, \boldsymbol{\theta}_k) &= c_k + (d_k - c_k + f_k x_{ij})/\{1 + \exp[b_k \log(x_{ij}/e_k)]\}, \\ & \quad i = 1; 2; \ldots, 10\ , \quad j = 1, 2, \ldots, r, \quad k = 1, 2, \end{aligned}$$

em que x_{ij} é a dose do agente químico na j-ésima repetição do i-ésimo tratamento, $\boldsymbol{\theta}_k = [c_k\ d_k\ f_k\ b_k\ e_k]^T$ são os vetores de parâmetros. Os parâmetros foram fixados com os valores $c_1 = 0$, $d_1 = 44,76$, $f_1 = 0,76$, $b_1 = 2,35$, $e_1 = 54,36$, $c_2 = 0$, $d_2 = 298,37$, $f_2 = 5,30$, $b_2 = 3,14$ e $e_2 = 60,94$.

O modelo de regressão não linear para as duas respostas pode ser escrito na forma vetorial como

$$\boldsymbol{Y}_k = f_k(\boldsymbol{\theta}_k) + \boldsymbol{\epsilon}_k.$$

Os valores das respostas foram obtidos adicionando os erros normais bivariados, com variâncias 14,0625 e 408,8484 e covariâncias $22,7475$ ($\rho = 0,3$), $37,9125$ ($\rho = 0,5$), $53,0775$ ($\rho = 0,7$) e $68,2425$ ($\rho = 0,9$) de acordo com o número de repetições.

Após a obtenção dos dados simulados, os vetores de parâmetros $\boldsymbol{\theta}_k, k = 1, 2$ foram estimados ($\widehat{\boldsymbol{\theta}}_k$) para cada cenário, seguindo os passos estabelecidos na Seção 4.3. Em seguida, foram obtidas as médias das estimativas dos parâmetros (média($\widehat{\boldsymbol{\theta}}_k$)) e desvio padrão, considerando os modelos univariados e bivariado, para cada conjunto de dados. O viés relativo (VR) e a raiz do erro quadrático médio relativo (REQMR) foram calculados.

4.4.2 Resultados

Os resultados, apresentados nas Tabelas suplementares S1 - S3, indicam que o ajuste bivariado do modelo Brain e Cousens (1989) superou os ajustes dos modelos univariados em, praticamente, todos os cenários, apresentando-se com maior precisão, menor viés e com maior acurácia para a estimação dos parâmetros. Quando isso não ocorreu, mostrou-se similar aos ajustes dos modelos univariados.

Geralmente, a superioridade do modelo bivariado é mais evidente com o aumento das correlações entre as características, independentemente do número de repetições adotadas em um experimento (Figuras 4.1 e 4.2). Avaliando os parâmetros com significado biológico direto d_1, d_2, f_1 e f_2, o VR se aproxima rapidamente de zero à medida que aumenta a correlação, enquanto que para os ajustes dos modelos univariados o VR se mantém aproximadamente constante (Figura 4.1).

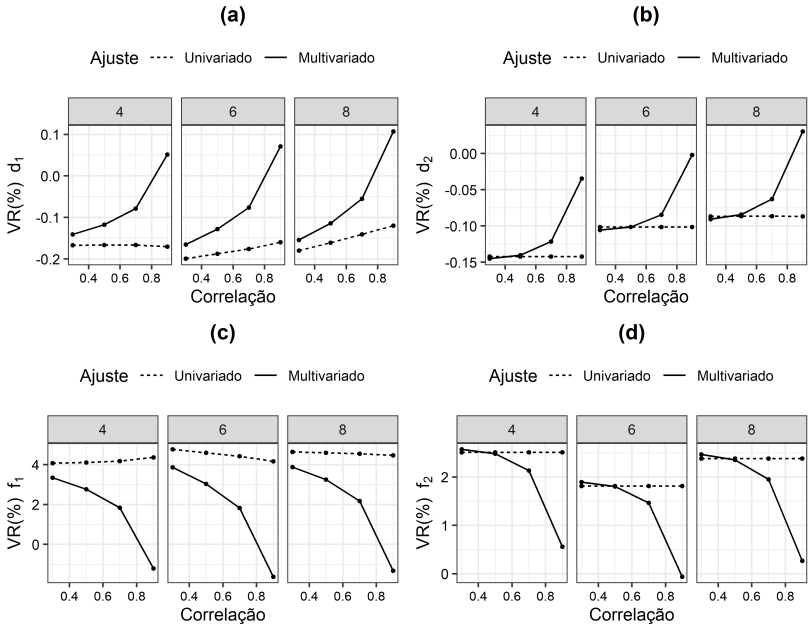

Figura 4.1. Estimativa do viés relativo (VR(%)), considerando 1000 simulações em experimentos com diferentes números de repetições ($r = 4, 6, 8$) por nível de dose do agente químico e diferentes níveis de correlação ($\rho = 0,3; 0,5; 0,7; 0,9$) entre os resíduos das variáveis resposta para os parâmetros: (a) d_1, (b) d_2, (c) f_1 e (d) f_2.

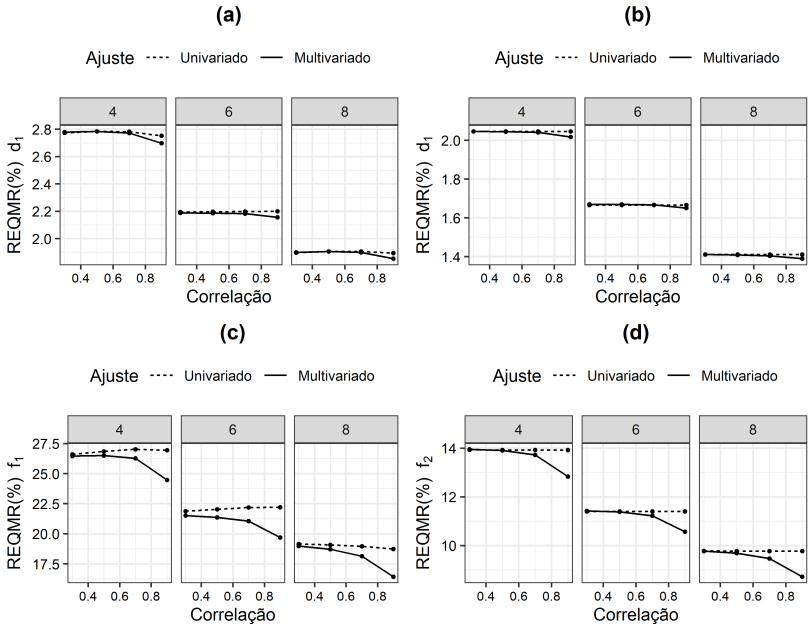

Figura 4.2. Estimativa da raiz do erro quadrático médio relativo (REQMR(%)), considerando 1000 simulações em experimentos com diferentes número de repetições ($r = 4, 6, 8$) por nível de dose do agente químico e diferentes níveis de correlação ($\rho = 0,3; 0,5; 0,7; 0,9$) entre os resíduos das variáveis resposta para os parâmetros: (a) d_1, (b) d_2, (c) f_1 e (d) f_2.

O REQMR evidencia resultados similares para as estimativas dos parâmetros d_1 e d_2 para os ajustes dos modelos bivariado e univariados. Em contrapartida, para os parâmetros f_1 e f_2, os resultados mostram a superioridade do modelo bivariado à medida

que a correlação entre as características aumenta (Figura 4.2). Além disso, fixando-se o valor da correlação e variando o número de repetições, o modelo bivariado, também, é superior para maioria dos parâmetros. Destaca-se que à medida que aumenta o número de repetições, o VR e o REQMR (Figuras suplementares S1 e S2) diminuem, consideravelmente.

4.5 Estudo de caso

4.5.1 Descrição

A metodologia proposta (MNMH) é ilustrada com a análise de dados de um estudo de dose-resposta bifásico na cultura do cártamo (*Carthamus tinctorius* L.) da linhagem IMA14, do Instituto Mato Grosso do Algodão (IMAmt, MT, Brasil). Em 2019, dois experimentos foram conduzidos em casa de vegetação, sendo que em um deles a cultura do cártamo foi submetida a um regime hídrico sem estresse (-10kPa) e no outro com estresse severo (-70kPa).

Após 28 dias da semeadura do cártamo, 10 doses do herbicida glifosato (0; 1,8; 3,6; 7,2; 18; 36; 72; 180; 360 e 720 g e.a. ha^{-1}), com quatro repetições, foram aplicadas em vasos com uma planta, de acordo com o delineamento completamente casualizado. Cada experimento foi repetido duas vezes.

Variáveis fisiológicas foram medidas aos 3, 7, 14, 21 e 28 dias após aplicação do herbicida enquanto que as variáveis de crescimento e desenvolvimento foram medidas aos 7, 14, 21 e 28 dias após aplicação do herbicida (Santos et al., 2021). Para as variáveis de crescimento e desenvolvimento, os resultados referentes às duas repetições de cada experimento foram combinados em um único experimento, com 10 tratmentos e oito repetições. Neste trabalho, utilizaram-se apenas os dados das variáveis de crescimento: altura de planta (AP), areá foliar (AF) e massa de matéria seca (MMS), mensuradas aos 28 após aplicação do herbicida, sob estresse severo (-70kPa).

4.5.2 Modelo

Os modelos de regressão não linear para as respostas AP, AF e MMS podem ser escritos como

$$Y_{ij,k} = f_k(x_{ij}, \boldsymbol{\theta}_k) + \epsilon_{ij}, \quad i = 1, 2, \ldots, 10\ , \quad j = 1, 2, \ldots, 8, \quad k = 1, 2, 3,$$

em que

$$f_k(x_{ij}, \boldsymbol{\theta}_k) \;\; = \;\; c_k + (d_k - c_k + f_k x_{ij})/\{1 + \exp[b_k \log(x_{ij}/e_k)]\}, \quad k = 1, 2, 3$$

com vetores de parâmetros $\boldsymbol{\theta}_k = [c_k\ d_k\ f_k\ b_k\ e_k]$.

O algoritmo para estimação dos parâmetros seguiram os passos conforme descrito na Seção 4.3. O mesmo procedimento foi realizado para obtenção das características quantitativas sub-NOAEL definidas nas equações (4.3), (4.4) e (4.5).

4.5.3 Resultados

Uma análise exploratória inicial mostra que as variáveis AP, AF e MMS apresentaram, correlações positivas acima de 0,9 de acordo com a Figura 4.3(a), dando indicação de forte relação biológica entre elas. Além disso, pelo teste de Mardia (assimetria $p = 0,056503$ e curtose $p = 0,53902$) (Mardia, 1974) há evidências, ao nível de 5% de significância, que essas variáveis possuem normalidade multivariada, possibilitando o emprego do modelo não linear multivariado para modelar a curva bifásica de hormesis.

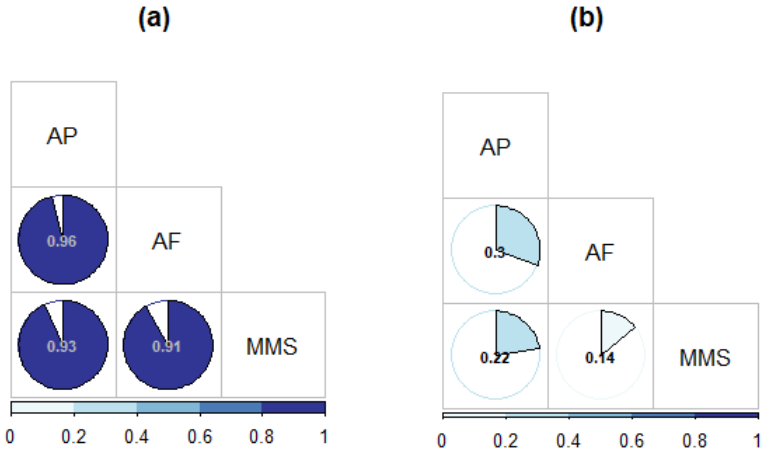

Figura 4.3. Gráficos da matriz de correlação (a) entre as variáveis AP, AF e MMS antes do ajuste e (b) entre os resíduos das variáveis AP, AF e MMS.

Na Tabela 4.1, são apresentadas as estimativas dos parâmetros dos modelos univariados e multivariados. As estimativas univariadas do parâmetro f indicam a existência de hormesis ($f > 0$). A partir dos resíduos para cada variável resposta, tem-se a estimativa da matriz de variâncias e covariâncias

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 21,460943 & 49,53397 & 0,47470 \\ 49,53397 & 1245,99621 & 2,29592 \\ 0,47470 & 2,29592 & 0,21489 \end{bmatrix}.$$

Além disso, pode-se observar que a correlação residual entre as três variáveis de crescimento é baixa, principalmente, entre AF e MMS (Figura 4.3(b)). Entretanto, as correlações não podem ser ignoradas pois o valor observado da estatística T do teste de esfericidade de Bartlett (1951) é 11,77 ($> \chi^2_{3;0,05} = 0,35$), levando à rejeição da hipótese nula (H_0: matriz de variâncias e covariâncias $\hat{\boldsymbol{\Sigma}}$ é uma matriz diagonal).

A maioria das estimativas dos parâmetros do modelo não linear multivariado (Tabela 4.1) foram maiores que as estimativas para os modelos univariados, exceto para

Tabela 4.1. Estimativas dos parâmetros dos modelos univariados e multivariados de Brain e Cousens (1989) com os respectivos erros-padrão (e.p.), para as variáveis AP, AF e MMS de *C. tinctorius* mensuradas aos 28 após o tratamento com baixas doses de glifosato, sob estresse severo (−70kPa).

Variável	Parâmetros	Univariado		Multivariado	
		Estimativa	e.p.	Estimativa	e.p.
AP [cm]	c_1	0,00000#	-	0,00000 #	-
	d_1	44,76400	1,08620	44,87469	1,07963
	f_1	0,76050	0,17510	0,72784	0,16918
	b_1	2,34750	0,14040	2,36761	0,14375
	e_1	54,35580	5,60390	55,50077	5,66525
AF [cm^2]	c_2	0,00000#	-	0,00000#	-
	d_2	298,36860	7,76210	299,61438	7,69406
	f_2	5,30010	0,91310	5,00613	0,86333
	b_2	3,14260	0,29940	3,23269	0,32378
	e_2	60,93790	4,13330	62,48938	4,05904
MMS [g]	c_3	0,57228	0,24988	0,69861	0,20548
	d_3	3,44596	0,12512	3,46450	0,12084
	f_3	0,10023	0,03444	0,08945	0,02938
	b_3	2,05042	0,25768	2,18887	0,27852
	e_3	38,46460	6,99381	40,40844	6,72735

#Parâmetro fixado.

os parâmetros $f's$ de hormesis, que foram menores, mostrando a influência das interdependências entre as variáveis. Observa-se, ainda que os erros-padrão foram menores para a maioria dos casos, quando se consideraram as correlações existentes entre as variáveis. As características quantitativas sub-NOAEL, também, apresentaram diferenças nas estimativas univariadas e multivariada (Tabela S4, Figura 4.4).

A presença de hormesis no contexto multivariado é confirmada pelas duas abordagens apresentadas na Seção 4.3.1. Os IC_{95} não contêm o zero ($f > 0$) (Figura 4.4) e o valor observado da estatística F do teste LRT para o vetor $\boldsymbol{\theta}_f = [f_1\ f_2\ f_3]^T$ é $33,33$ ($> F_{3,227;0,05} = 2,64$). Assim, rejeitou-se a hipótese nula de que não existe hormesis multivariada, ao nível de 5% de significância, ou seja, existem evidências que as três variáveis indicam de forma simultânea a presença do comportamento bifásico na cultura do *C. tinctorius* quando submetida ao estresse hídrico severo.

A resposta máxima média da hormesis é obtida com a dose $\overline{M} = 25,52(1,51)$ g e.a. ha^{-1} e desaparece com doses superiores a $\overline{LDS} = 55,24(2,29)$ g e.a. ha^{-1}. A dose média que causa a redução de 50% na resposta média do controle não tratado é $\overline{ED}_{50} = 101,09(5,96)$ g e.a. ha^{-1}. A resposta estimulatória relativa acima do controle ficou em média torno de ($Y_{max}[\%] = 129,41\%$), porém, não é possível calcular o erro padrão para essa quantidade por meio de combinações lineares.

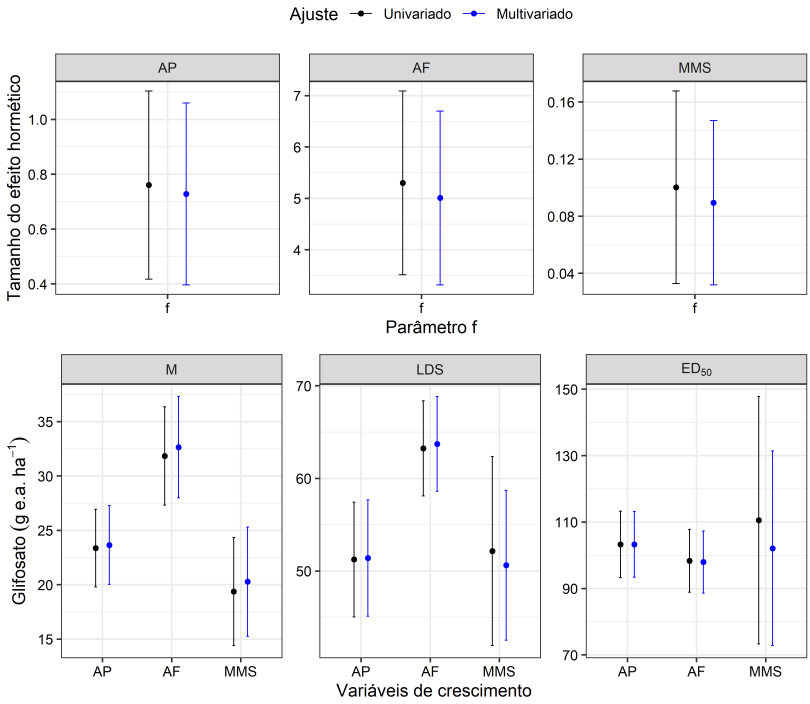

Figura 4.4. Intervalos, com 95% de confiança, para o parâmetro f e características quantitativas sub-NOAEL, considerando os modelos univariados (–) e multivariado (–): M [g e.a. ha^{-1}] = dose que causa estimulação máxima; LDS [g e.a. ha^{-1}] = dose limite para ocorrência de hormesis e ED_{50} [g e.a. ha^{-1}] = dose que causa a redução de 50% na resposta média do controle não tratado.

Os intervalos, com nível de confiança de 95%, para as características quantitativas médias sub-NOAEL, calculados a partir dos modelos univariados e multivariados são apresentados na Figura 4.5. Claramente, são mais estreitos para o modelo multivariado.

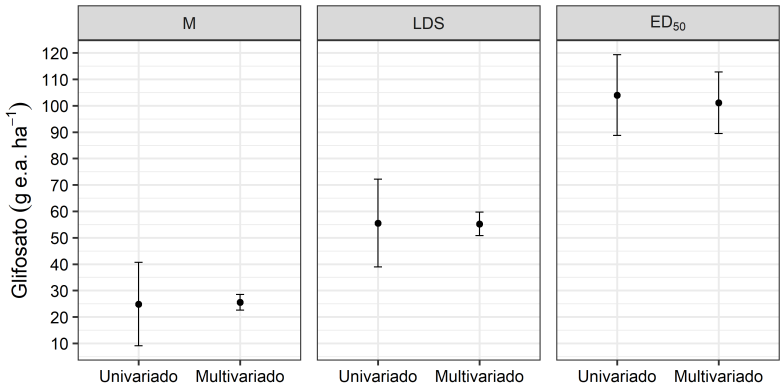

Figura 4.5. Intervalos, com 95% de confiança, para as médias das características quantitativas sub-NOAEL, considerando os modelos univariados e multivariado: M [g e.a. ha^{-1}] = dose que causa estimulação máxima; LDS [g e.a. ha^{-1}] = dose limite para ocorrência de hormesis e ED_{50} [g e.a. ha^{-1}] = dose que causa a redução de 50% na resposta média do controle não tratado.

4.6 Discussão

Dados correlacionados são comuns na maioria dos estudos de hormesis, presentes em características morfológicas, fisiológicas e bioquímicas, conforme destacado por Belz (2018), Pincelli-Souza et al. (2020) e Santos et al. (2022). Assim, não seria adequado do ponto de vista biológico analisar os diversos componentes de um sistema vegetal sem incorporar as correlações existentes entre variáveis mensuradas em uma mesma planta e/ou grupo de plantas.

Neste estudo, com base nos trabalhos de Zellner (1962) e Gallant (1987), para entender a dinâmica da hormesis do ponto de vista global (Belz e Duke, 2022), desenvolveu-se uma modelagem estatística não linear multivariada (MNMH). Usando um estudo de simulação e uma aplicação em dados reais, apontaram-se as vantagens da MNMH em relação às abordagens univariadas que ignoram as correlações.

Considerar as diversas correlações presentes em um sistema vegetal melhora o processo de modelagem da hormesis em dois principais aspectos: (a) o pedagógico, que permite ao pesquisador interpretar a influência da hormesis na totalidade sem perder de vista os detalhes individuais de cada característica mensurada, e (b) o estatístico pois as estimativas tendem a ser mais precisas, menos viesadas e com maior acurácia conforme apontado no estudo de simulação (Seção 4.4) e nos estudos iniciais sobre esse tema (Zellner, 1962; Gallant, 1987; Seber e Wild, 2003).

O processo de estimação dos parâmetros, no caso multivariado, ocorre em três etapas conforme a Seção 4.3, sendo que as etapas II e III constituem o principal diferencial em relação à modelagem tradicional univariada. Problemas de convergência ou de multicolinearidade parecem não ser um obstáculo. Uma boa prática para minimizar os possíveis problemas de convergência é utilizar, na etapa III, as estimativas dos parâmetros das análises univariadas como valores iniciais. A estimativa da matriz de variâncias e covariâncias, $\hat{\boldsymbol{\Sigma}}$, obtida na etapa II, é positiva definida e dessa forma, o problema de multicolinearidade é descartado (Gallant, 1987). Por fim, o uso da decomposição de Cholesky faz com que o método NLS possa ser utilizado ao invés do método GLS (Gallant, 1987; Seber e Wild, 2003).

4.6.1 Estudo de simulação

O uso de estudos de simulação é amplamente empregado como forma de validação de modelos estatísticos (Schabenberger e Birch, 2001; Seber e Wild, 2003; Cranmer et al., 2020; Türkşen, 2021). Considerou-se um delineamento inteiramente casualizado, comumente usado pela maioria dos pesquisadores em biologia vegetal que abordam o tema de hormesis. Os cenários simulados usaram experimentos com diferentes números de repetições para cada nível de dose do agente químico e diferentes graus de correlação residual entre as variáveis respostas.

Comprovou-se que a abordagem multivariada foi superior na maioria dos cenários consoante os indicadores de precisão (SE), viés (VR) e acurácia (REQMR), apontando ganho de eficiência com o aumento da correlação entre os resíduos dos diferentes modelos não lineares confirmando as observações feitas, inicialmente, por Zellner (1962). Evidenciou-se (Figuras 4.1 e 4.2) que para a modelagem univariada, por não capturar as correlações existentes entre os diversos componentes de uma planta, as estimativas de VR e REQMR se mantêm, praticamente, constantes à medida que aumenta a correlação residual. Os parâmetros mais afetados pelo grau de correlação entre os resíduos das variáveis respostas foram d e f no modelo de Brain e Cousens (1989). Belz e Piepho (2013), em um estudo com *Lactuca sativa* L., mostraram que parâmetros d e f do modelo de Cedergreen et al. (2005) influenciam, diretamente, a resposta estimulatória máxima (Y_{max}) e que o efeito hormético aumenta com o aumento dos valores de f. Logo, a influência das correlações entre as multicaracterísticas ou das inter-relações dos parâmetros que governam a forma de uma curva de dose-resposta hormética não podem ser negligenciados.

4.6.2 Estudo de caso

As variáveis morfológicas AP, AF e MMS, mensuradas aos 28 dias após a semeadura, sob estresse severo (-70kPa) apresentaram alta correlação. Além disso, essas variáveis apresentaram normalidade univariada e multivariada pelos testes de Shapiro-Wilk e Mardia, respectivamente (Shapiro e Wilk, 1965; Mardia, 1974).

As correlações residuais estimadas entre as variáveis morfológicas AP, AF e MMS, obtidas após o ajuste dos modelos univariados, são baixas mas não podem ser ignoradas, segundo o teste de esfericidade Bartlett (Bartlett, 1951). Nessa mesma linha, os resultados do estudo de simulação indicam que em situações de baixa correlação ($\rho = 0,3$), VR diminui quando a modelagem multivariada é considerada (Figura S1). Assim, as diferenças observadas nas estimativas dos parâmetros entre os ajustes univariado e multivariado são consequência direta das correlações entre as variáveis envolvidas (Tabela 4.1).

A presença de hormesis é confirmada sob os aspectos univariado e multivariado, contudo, sendo que a hipótese de hormesis multivariada pode ser verificada pelo teste LRT. Esse teste é amplamente utilizado no contexto univariado (Schabenberger et al., 1999; Santos et al., 2021, 2022), por outro lado, não existem relatos na literatura de hormesis multivariada, consistindo, portanto, um avanço na modelagem da hormesis.

As características quantitativas sub-NOAEL (M, LDS e ED_{50}) estimadas pelas funções reparametrizadas de equações (4.3), (4.4) e (4.5), apresentaram poucas diferenças entre os ajustes univariado e multivariado nas variáveis morfológicas AP, AF e MMS, devido à baixa correlação residual (Figura 4.3), mas em situações de maior correlação residual, essas diferenças podem ser acentuadas, conforme evidenciado no estudo de simulação (Tabelas suplementares S1 - S3). O $Y_{max}[\%]$ médio ficou em conformidade com estudo anterior apresentado por Santos et al. (2021). Na Figura 4.5, as doses M, LDS

e ED_{50} são as recomendações médias para as plantas de *C. tinctorius*, sob deficiência hídrica severa.

4.6.3 Por que modelar hormesis com uma estrutura não linear multivariada?

A metodologia apresentada neste trabalho por ser estendida para os diversos campos das ciências biológicas e toxicológicas, permitindo que variáveis medidas em escalas diferentes (cm, cm^2, g, entre outras unidades) possam ser agrupadas na mesma estrutura de modelagem sem perda de informação. Com referência a MNMH, três aspectos parecem de extrema importância para considerá-la. *Primeiro*, as estimativas dos parâmetros são mais precisas, pois, as interdependências entre as variáveis mensuradas são consideradas. *Em segundo lugar*, a hipótese de hormesis multivariada pode ser testada de forma simultânea para o vetor $\boldsymbol{\theta}_f = [f_1\ f_2\ , \ldots, f_w]^T$ por meio do teste LRT. Portanto, o teste de significância para o vetor $\boldsymbol{\theta}_f$, é mais promissor para confirmar a existência da hormesis do que os ajustes individuais, assim fornecer maior confiabilidade para os uso dos efeitos benéficos da hormesis comercialmente. *Em terceiro lugar*, o pesquisador necessita recomendar uma dosagem média da dose M e/ou ED_k. Na maioria das vezes, essa recomendação é feita simplesmente pela média aritmética dessas quantidades, nos caracteres morfológicas, fisiológicos e bioquímicos. Nesse sentido, uma característica importante do método de ajuste não linear multivariado é a possibilidade de construir intervalos de confiança mais precisos para essas características quantitativas sub-NOAEL (Figura 4.5). Finalmente, a única desvantagem da modelagem multivariada pode ser que o esforço de modelagem seja maior, mas as possibilidades de modelagem aprimorada e a robustez da técnica devem compensar isso.

A metodologia apresentada pode ser facilmente estendida para outros modelos, como por exemplo, o modelo de Cedergreen et al. (2005), e o modelo bilogístico apresentado por Nweke et al. (2022), para avaliação da ocorrência de hormesis. As métricas REQM e MVA (Seção 4.3.1) podem ser utilizadas para seleção do modelo mais adequado para cada conjunto de dados.

A versão atual do algoritmo de ajuste (Apêndice B) é uma implementação preliminar da modelagem não linear multivariada. Planejamos desenvolver um pacote `R` com uma interface amigável que contemple as principais características quantitativas sub-NOAEL.

4.7 Conclusão

Este é o primeiro estudo que aborda a hormesis no contexto multivariado, possibilitando uma maior compreensão desse fenômeno bifásico, pois considera na estrutura de modelagem as inter-relações entre as diversas características que podem ser mensuradas em um sistema vegetal. Com isso, as estimativas dos parâmetros tendem a ser mais precisas. Portanto, metodologia proposta neste trabalho é uma ferramenta valiosa para

quantificar as características da resposta hormética e também para avaliadores de risco que estão preocupados com a aplicação de doses perigosas de produtos químicos.

Referências

Bartlett, M. S. (1951). The effect of standardization on a χ^2 approximation in factor analysis. *Biometrika*, 38(3/4):337–344.

Bates, D. M. e Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. Wiley.

Belz, R. G. (2018). Herbicide hormesis can act as a driver of resistance evolution in weeds–psii-target site resistance in *Chenopodium album* L. as a case study. *Pest Management Science*, 74(12):2874–2883.

Belz, R. G. (2020). Low herbicide doses can change the responses of weeds to subsequent treatments in the next generation: metamitron exposed psii-target-site resistant chenopodium album as a case study. *Pest Management Science*, 76(9):3056–3065.

Belz, R. G. e Duke, S. O. (2014). Herbicides and plant hormesis. *Pest Management Science*, 70(5):698–707.

Belz, R. G. e Duke, S. O. (2022). Modelling biphasic hormetic dose responses to predict sub-noael effects using plant biology as an example. *Current Opinion in Toxicology*.

Belz, R. G. e Piepho, H.-P. (2012). Modeling effective dosages in hormetic dose-response studies. *PloS ONE*, 7(3):e33432.

Belz, R. G. e Piepho, H. P. (2013). Variability of hormetic dose responses of the antiauxin PCIB on *Lactuca sativa* in a plant bioassay. *Weed Research*, 53(6):418–428.

Belz, R. G. e Piepho, H.-P. (2015). Statistical modeling of the hormetic dose zone and the toxic potency completes the quantitative description of hormetic dose responses. *Environmental Toxicology and Chemistry*, 34(5):1169–1177.

Bortolheiro, F. P. A. P. e Silva, M. A. (2021). Low doses of glyphosate can affect the nutrient composition of common beans depending on the sowing season. *Science of the Total Environment*, 794:148733.

Brain, P. e Cousens, R. (1989). An equation to describe dose responses where there is stimulation of growth at low doses. *Weed Research*, 29(2):93–96.

Calabrese, E. J. (2005). Paradigm lost, paradigm found: the re-emergence of hormesis as a fundamental dose response model in the toxicological sciences. *Environmental Pollution*, 138(3):378–411.

Cedergreen, N. (2008). Is the growth stimulation by low doses of glyphosate sustained over time? *Environmental Pollution*, 156(3):1099–1104.

Cedergreen, N., Ritz, C., e Streibig, J. C. (2005). Improved empirical models describing hormesis. *Environmental Toxicology and Chemistry: An International Journal*, 24(12):3166–3172.

Cranmer, K., Brehmer, J., e Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062.

Duke, S. O. (2019). Enhanced Metabolic Degradation: The Last Evolved Glyphosate Resistance Mechanism of Weeds? *Plant Physiology*, 181(4):1401–1403.

Elzhov, T. V., Mullen, K. M., Spiess, A.-N., e Bolker, B. (2023). *minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds*. R package version 1.2-3.

Gallant, A. R. (1987). *Nonlinear Statistical Models*. John Wiley and Sons, New York.

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 115–128.

Nweke, C. O., Nwangwu, O. R., Okechi, R. N., Araka, N. N., e Ogbonna, C. J. (2022). Statistical modeling of hormesis quantities in inverted u-shaped dose-response relationships by reparameterization of a bilogistic model. *Journal of Environmental Science and Health, Part A*, 57(12):1003–1023.

Pincelli-Souza, R. P., Bortolheiro, F. P. A. P., Carbonari, C. A., Velini, E. D., e Silva, M. A. (2020). Hormetic effect of glyphosate persists during the entire growth period and increases sugarcane yield. *Pest Management Science*, 76(7):2388–2394.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Santos, J. C. C., da Silva, D. M. R., Amorim, D. J., Sab, M. P. V., e de Silva, M. A. (2021). Glyphosate hormesis mitigates the effect of water deficit in safflower (*Carthamus tinctorius* L.). *Pest Management Science*, 77(4):2029–2044.

Santos, J. C. C., Silva, D. M. R., Amorim, D. J., Rosa, V. R., Santos, A. L. F., Veline, E. D., Carbonari, C. A., e Silva, M. A. (2022). Glyphosate hormesis attenuates water deficit stress in safflower (*Carthamus tinctorius* L.) by modulating physiological and biochemical mediators. *Science of The Total Environment*, 810:152204.

SAS Institute Inc (2015). *SAS/STAT® 14.1 user's guide*. Cary, NC: SAS Institute Inc.

Schabenberger, O. e Birch, J. B. (2001). Statistical dose-response models with hormetic effects. *Human and Ecological Risk Assessment*, 7(4):891–908.

Schabenberger, O., Tharp, B. E., Kells, J. J., e Penner, D. (1999). Statistical tests for hormesis and effective dosages in herbicide dose response. *Agronomy Journal*, 91(4):713–721.

Seber, G. A. F. e Wild, C. J. (2003). *Nonlinear Regression*. Wiley Series in Probability and Statistics. Wiley.

Shapiro, S. S. e Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Southam, C. M. (1943). Effects of extract of western red-cedar heartwood on certain wood-decaying fungi in culture. *Phytopathology*, 33:517–524.

Streibig, J. C. (1980). Models for curve-fitting herbicide dose response data. *Acta Agriculturae Scandinavica*, 30(1):59–64.

Streibig, J. C. (1988). Herbicide bioassay. *Weed Research*, 28(6):479–484.

Türkşen, Ö. (2021). A novel perspective for parameter estimation of seemingly unrelated nonlinear regression. *Journal of Applied Statistics*, 48(13-15):2326–2347.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368.

5 CONSIDERAÇÕES FINAIS

Neste trabalho, foram desenvolvidos modelos estatísticos e ferramentas computacionais, com aplicações em dados de produção de embriões e biologia vegetal. O primeiro artigo apresentado no Capítulo 2, explorou os desenvolvimentos estatísticos existentes para análise de dados na forma de contagens fornecidos pela RuAnn Genetics Laboratory (Riverdale, Califórnia, EUA), referentes ao ano de 2020 para raça bovina holandesa. Foram discutidos métodos para selecionar o melhor modelo e as implementações foram apresentadas no software `R` ao longo do texto. Foi observada a presença do fenômeno de superdispersão e a presença de correlação resultante das medidas repetidas. Dentre os modelos ajustados, os que apresentaram melhor desempenho foram o Poisson-normal e Poisson-gama-normal. Entretanto, o modelo Poisson-normal foi adotado para as inferências, pois foi considerado mais parcimonioso. Do ponto de vista prático, o intervalo entre aspirações e o uso do protocolo hormonal com FSH não alterou o números totais de oócitos.

No Capítulo 3, foi abordada a implementação do pacote `R` **combTMB**, sendo uma ferramenta útil para ajustar e avaliar uma gama de diferentes modelos, incluindo os modelos lineares generalizados, modelos lineares generalizados mistos e modelos combinados. Esse pacote utiliza a aproximação de Laplace para fazer a integração numérica em relação aos efeitos aleatórios e diferenciação automática para o cálculo das derivadas (usando o pacote **TMB**), para obter as estimativas dos parâmetros por máxima verossimilhança ou, opcionalmente, por máxima verossimilhança restrita. O pacote **combTMB** ainda está em desenvolvimento, e o código-fonte pode ser encontrado no GitHub do autor[1], sugestões e cometários são bem vindos.

O conceito de hormesis em plantas é abordado no Capítulo 4. Neste trabalho, propõe-se a modelagem não linear multivariada da hormesis, em que as informações concernentes às correlações das respostas são consideradas com uma matriz de covariância obtida, a partir dos resíduos univariados. Assim, o pesquisador pode obter uma maior compreensão das relações bifásicas de dose-resposta, pois considera na estrutura de modelagem as inter-relações existentes entre as diversas características mensuradas no sistema vegetal.

O pacote **combTMB** e a modelagem não linear multivariada da hormesis são as principais contribuições deste trabalho. Como sugestão para trabalhos futuros, podem-se desenvolver estudos relacionados às estruturas de variância e covariâncias, a fim de entender melhor as possíveis dependências entre as covariáveis do modelo, no contexto dos Capítulos 2 e 3. No contexto do Capítulo 4, pode-se estender a metodologia apresentada para outros modelos de regressão não linear, bem como outros fenômenos biológicos.

[1] `https://github.com/deoclecioamorim/combTMB`

APÊNDICES

Apêndice A: material suplementar do Capítulo 3

Neste material suplementar, mostra-se de forma simples os recursos do pacote **combTMB** para o ajuste de alguns modelos a dados na forma de contagens e proporções. Nessas análises, foi utilizado o conjunto de dados "embryos" disponibilizado em conjunto com o pacote.

Preliminares

Chamando os pacotes e o conjunto de dados

```
library(combTMB)
library(bbmle)
library(emmeans)
library(tidyverse)
library(lme4)
library(microbenchmark)
data(embryos, package = "combTMB")
```

A1: Dados de contagem

Para os dados de contagem a variável resposta é o número de oócitos totais (OT), e o interesse está em verificar se é influenciada pelos fatores fixos período do ano (`Period`) e status da doadora (`Status`). Além disso, considera-se efeito aleatório em nível de doadora (`Donor`) nos preditores lineares.

Seleção de modelos - inferência sobre efeitos aleatórios

Para testar a hipótese $H_0 : \sigma^2_O = 0$ versus $H_0 : \sigma^2_O > 0$, o teste da razão de verossimilhanças (LRT) é feito pela comparação dos modelos Poisson (M1) versus Poisson-normal (M2) e, posteriormente, pela comparação dos modelos binomial negativo (M3) versus Poisson-gama-normal (M4), com preditor linear de equação (3.26). O ajuste desses modelos no pacote **combTMB** é feito usando os seguintes comandos:

```
##--Modelo Poisson (MLG)
M1 <- combTMB(OT ~ Period * Status, embryos, family=poisson)
##--Modelo Poisson-normal (MLGM)
M2 <- combTMB(OT ~ Period * Status + (1|Donor), embryos, family=poisson)
##--Modelo Binomial negativo (MLG)
M3 <- combTMB(OT ~ Period * Status, embryos, family=poigamma)
##--Modelo Poisson-gama-normal (MC)
```

```
M4 <- combTMB(OT ~ Period * Status + (1|Donor), embryos, family=poigamma)
```

Para construir a Tabela 3.2, o usuário pode utilizar as funções e/ou métodos `logLik()`, `df.residual()` e `AIC()` disponíveis no pacote **combTMB**. Dessa forma, executam-se os comandos:

```
R> -2*logLik(M1); -2*logLik(M2); -2*logLik(M3); -2*logLik(M4) #-2log(L)
'log Lik.' 10883.23 (df=6)
'log Lik.' 7780.387 (df=7)
'log Lik.' 8172.57 (df=7)
'log Lik.' 7597.504 (df=8)
R> df.residual(M1); df.residual(M2);df.residual(M3); df.residual(M4) #GL
[1] 1142
[1] 1141
[1] 1141
[1] 1140
R> AIC(M1); AIC(M2); AIC(M3); AIC(M4) #AIC
[1] 10895.23
[1] 7794.387
[1] 8186.57
[1] 7613.504
```

A lista completa dos métodos disponíveis no **combTMB** é obtida com o seguinte comando:

```
 R> methods(class = "combTMB")
 [1] anova        coefDisp     df.residual  dispersion   emm_basis
 [6] family       fitted       fixef        formula      getME
[11] logLik       model.frame  model.matrix nobs         partvar
[16] predict      print        ranef        recover_data residuals
[21] simulate     summary      terms        vcov
see '?methods' for accessing help and source code
```

O teste LRT pode ser executado da seguinte forma:

```
R> ##--Teste LRT
R> ##--Poisson versus Poisson-normal
R> (dif <- as.numeric(-2*(logLik(M1)-logLik(M2))))
[1] 3102.84
R> valor_p <- 0.5*pchisq(dif,df=1,lower=FALSE)
R> data.frame(dif, valor_p)
      dif valor_p
```

```
1 3102.84       0

R> ##--Binomial negativo versus Poisson-gama-normal
R> (dif <- as.numeric(-2*(logLik(M3)-logLik(M4))))
[1] 575.0658
R> valor_p <- 0.5*pchisq(dif,df=1,lower=FALSE)
R> data.frame(dif, valor_p)
       dif       valor_p
1 575.0658 2.220023e-127
```

Os valores p obtidos sugerem que existem fortes evidências para se rejeitar H_0. Portanto, há necessidade do efeito aleatório de doadora nos preditores lineares dos modelos Poisson-normal e Poisson-gama-normal.

Seleção de modelos - inferência sobre efeitos fixos

Os códigos para o ajuste dos modelos com os preditores lineares (3.24) e (3.25) são dados por:

```
##--Seleção parte fixa do modelo
##--MCs
M5 <- combTMB(OT ~ Period + (1|Donor), embryos, family=poigamma)
M6 <- combTMB(OT ~ Period + Status + (1|Donor), embryos, family=poigamma)
```

O critério AIC foi usado para comparar os modelos M4, M5 e M6. Com auxílio da função `AICtab()` do pacote **bbmle** esse processo pode ser otimizado.

```
R> ##--Função bbmle::AICtab
R> AICtab(M4,M5,M6)
   dAIC df
M4  0   8
M6  5   6
M5 45   4
```

O modelo M4 é o modelo mais parcimonioso com efeitos de `Period`, `Staus` e interação `Period * Status`.

Avaliação da qualidade de ajuste

A função `sanitycombTMB()` pode ser utilizada para verificação inicial da qualidade de ajuste. Essa função aplicada ao modelo M4 não sinalizou problemas conforme a saída demonstrada a seguir:

```
R> sanitycombTMB(M4)
 Suggests successful convergence!
 Hessian matrix is positive definite!
 No extreme or very small eigen values detected!
 No fixed-effect standard errors are NA
 No fixed-effect standard errors look unreasonably large
```

Outra etapa importante para o estabelecimento do modelo é a detecção de possíveis falhas, que pode ser feita por meio de análises gráficas dos resíduos. No **combTMB** a análise de resíduos pode executada com a função `dharma_residuals()`. Essa função, por padrão, retorna um gráfico `QQ plot`, mas o usuário pode desabilitar a geração automática de gráficos com `plot = FALSE` e armazenar a saída em um objeto para posterior elaboração de gráficos conforme o interesse do usuário (por exemplo, a Figura 3.1).

```
R> ##---Resíduos com a função dharma_residuals
R> dharma_residuals(M4, nsim=250, plot=TRUE)
R> dharma_residuals(M4, nsim=500, plot=TRUE)
```

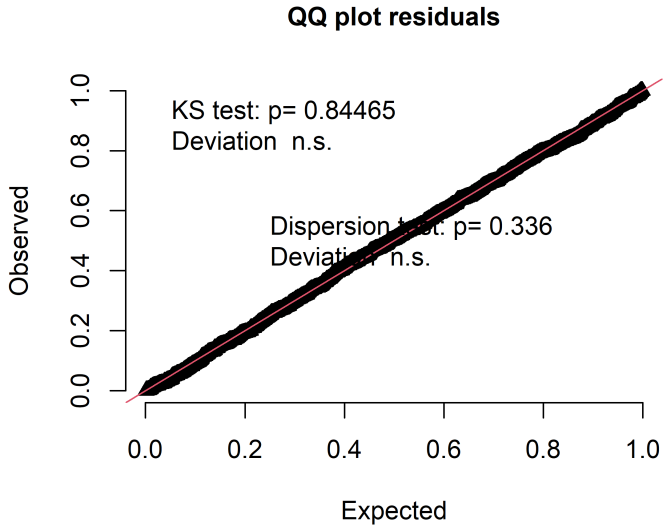

Figura S1. Resíduos quantílicos aleatorizados obtidos por meio da função `dharma_residuals()` para o modelo M4 obtidos a partir de 250 simulações

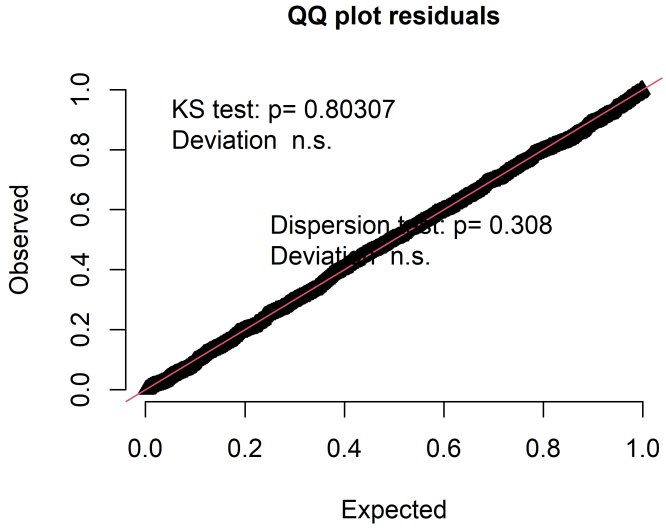

Figura S2. Resíduos quantílicos aleatorizados obtidos por meio da função `dharma_residuals()` para o modelo M4 obtidos a partir de 500 simulações

Interpretação condicional e marginal

O modelo M4 possui apenas o efeito aleatório de doadora e, neste caso, o pacote **combTMB** permite que a interpretação do vetor de parâmetros $\boldsymbol{\beta}$, também, sob o contexto marginal. A obtenção dos $\boldsymbol{\beta^m}$ é feita habilitando o argumento `doMarginal=TRUE`, assim o modelo é reestimado e obtêm-se as estimativas dos $\boldsymbol{\beta^m}$.

```
M4MCM <- combTMB(OT ~ Period * Status + (1|Donor), embryos,
                 family=poigamma, doMarginal = TRUE)
```

```
R> summary(M4MCM)
Family: poigamma
Link function: log
Formula: OT ~ Period * Status + (1 | Donor)
Dformula: ~1
Number of obs: 1148
-2 x logLik         AIC         BIC    df.resid
     7597.5      7613.5      7653.9        1140

                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.120123   0.055973  55.743  < 2e-16 ***
PeriodP2         -0.002989   0.035812  -0.083  0.93349
StatusH          -0.499996   0.071490  -6.994 2.67e-12 ***
StatusM          -0.371633   0.088879  -4.181 2.90e-05 ***
PeriodP2:StatusH  0.100132   0.053750   1.863  0.06248 .
```

```
PeriodP2:StatusM  0.219405   0.077670   2.825  0.00473 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Random effects:
                  Estimate Std. Error
Donor (Intercept)   0.2069      0.021
Overdisp.(theta)   22.2978      2.569

Number of subjects: 318
```

A interpretação marginal pode ser complementada com uso da função `emmeans` do pacote **emmeans** para o calculo das médias marginais, por exemplo, para o estudo da interação `Period*Status`:

```
R> ##---Médias marginais obtidas com o pacote emmeans
R> mm_M4MCM <- emmeans(M4MCM, ~ Status * Period, type="response")
R> mm_M4MCM <- data.frame(mm_M4MCM)
R> mm_M4MCM
  Status Period response        SE   df lower.CL upper.CL
1      D     P1 22.64918 1.2677488 1140 20.29351 25.27828
2      H     P1 13.73748 0.6541010 1140 12.51222 15.08271
3      M     P1 15.61903 1.1498064 1140 13.51841 18.04607
4      D     P2 22.58159 1.1561102 1140 20.42345 24.96777
5      H     P2 15.13895 0.6624756 1140 13.89337 16.49619
6      M     P2 19.39290 1.1878002 1140 17.19696 21.86923
```

Os códigos utilizados para obter a Figura 3.2 podem ser encontrados em `https://github.com/deoclecioamorim/cap2_SM/blob/master/emm_cont.R`.

A2: Dados na forma de proporções

Nesse caso, a resposta de interesse é a taxa de clivagem, e como é influenciada pelos fatores fixos período do ano (`Period`) e status da doadora (`Status`). Além disso, consideram-se os efeitos aleatórios em nível de doadora (`Donor`) e touro (`Sire`) nos preditores lineares.

Seleção de modelos - inferência sobre efeitos aleatórios

De forma semelhante ao caso de contagens, podem-se realizar testes sobre os efeitos aleatórios de doadora ($H_0 : \sigma_d^2 = 0$) e de touro ($H_0 : \sigma_s^2 = 0$). Para isso, diferentes

modelos (M1 a M8) foram ajustados usando o preditor linear (3.26). Os códigos para ajustar os diferentes modelos no pacote **combTMB** são dados a seguir:

```
##--Seleção parte aleatória do modelo
##--Modelo binomial (MLG)
M1 <- combTMB(cbind(D3, IVC-D3) ~ Period*Status, embryos,
                    family=binomial)
##--MLGMs
##--Modelo binomial-normal
M2 <- combTMB(cbind(D3, IVC-D3) ~ Period*Status + (1|Donor), embryos,
                   family=binomial)
##--Modelo binomial-normal
M3 <- combTMB(cbind(D3, IVC-D3) ~ Period*Status + (1|Sire), embryos,
                  family=binomial)
##--Modelo binomial-normal
M4 <- combTMB(cbind(D3, IVC-D3) ~ Period*Status + (1|Donor) + (1|Sire),
              embryos, family=binomial)
##--Modelo beta-binomial (MLG)
M5 <- combTMB(cbind(D3, IVC-D3) ~ Period*Status, embryos,
              family =  betabinomial)
##--MCs
##--Modelo beta-binomial-normal
M6 <- combTMB(cbind(D3, IVC-D3) ~ Period*Status+ (1|Donor), embryos,
              family =  betabinomial)
##--Modelo beta-binomial-normal
M7 <- combTMB(cbind(D3, IVC-D3) ~ Period*Status+ (1|Sire), embryos,
              family =  betabinomial)
##--Modelo beta-binomial-normal
M8 <- combTMB(cbind(D3, IVC-D3) ~ Period*Status+ (1|Donor)+(1|Sire), embryos,
              family =  betabinomial)
```

As informações presentes na Tabela 3.5 foram obtidas com auxílio das funções `logLik()`, `df.residual()` e `AIC()` aplicadas a cada um dos modelos.

Seleção de modelos - inferência sobre efeitos fixos

Após realizada a seleção e inferência sobre os efeitos aleatórios de doadora e touro, o próximo passo consiste em realizar testes sobre os efeitos fixos. A seguir, mostram-se os códigos para ajustar o modelo beta-binomial-normal com os preditores lineares dados

pelas equações (3.27) e (3.28) a fim de verificar se o modelo M8 com preditor linear da equação (3.29) pode ser simplificado.

```
##--MCs
M9 <- combTMB(cbind(D3, IVC-D3) ~ Period + (1|Donor) + (1|Sire),
              embryos, family=betabinomial)

M10 <- combTMB(cbind(D3, IVC-D3) ~ Period + Status + (1|Donor) + (1|Sire),
               embryos, family=betabinomial)
```

Inicialmente, para selecionar o modelo mais parcimonioso, utilizou-se o critério AIC. Posteriormente, empregou-se o teste LRT por meio da função `anova()`

```
R> ##--Função bbmle::AICtab
R> AICtab(M8, M9, M10)
    dAIC df
M8   0.0 9
M10  1.4 7
M9  16.5 5

##--Teste LRT
R> anova(M10,M8)
Likelihood ratio test for combTMB regression models

 Model 1: cbind(D3, IVC - D3) ~ Period + Status + (1 | Donor) + (1 | Sire)
 Model 2: cbind(D3, IVC - D3) ~ Period * Status + (1 | Donor) + (1 | Sire)

           AIC    BIC Resid.df  logLik Chisq.df  Chisq Pr(>Chisq)
Model 1 4582.1 4617.4     1141 -2284.1
Model 2 4580.7 4626.1     1139 -2281.3        2 5.4078    0.06695 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De acordo com o teste LRT, a interação `Period * Status` pode ser descartada do modelo. Portanto, o modelo M10 deve ser o escolhido para analisar a taxa de clivagem.

Avaliação da qualidade de ajuste

A verificação inicial do ajuste do modelo M10 pode ser feita utilizando a função `sanitycombTMB()`, complementada pelo uso da função `dharma_residuals()`. Observa-se que não houve sinalização de problemas, como pode ser observado na saída a seguir:

```
R> sanitycombTMB(M10)
 Suggests successful convergence!
 Hessian matrix is positive definite!
 No extreme or very small eigen values detected!
 No fixed-effect standard errors are NA
 No fixed-effect standard errors look unreasonably large
```

Ao utilizar a função `dharma_residuals()`, o usuário pode desativar a plotagem automática com o argumento `plot=FALSE` e armazenar a saída em um objeto, permitindo a elaboração de gráficos com outros layouts. Por exemplo, o seguinte código foi utilizado para elaborar a Figura 3.3:

```
R> resi1 <- dharma_residuals(M10, nsim=250,plot=FALSE)
R> head(resi1$out_1)
  Observed    Expected
1        0 0.000870322
2        0 0.001740644
3        0 0.002610966
4        0 0.003481288
5        0 0.004351610
6        0 0.005221932
R> resi2 <- dharma_residuals(M10, nsim=500, plot=FALSE)
R> g1 <- ggplot(data=resi1$out_1, aes(x=Expected,y=Observed))+
+    geom_point()+geom_abline(color="red")+ggtitle("(a)")+
+   theme_bw()+theme(
+     axis.text.x = element_text(color="black"),
+     axis.text.y = element_text(color="black"),
+     axis.ticks = element_line(color = "black")
+   )+ggtitle("(a)")+personal_title
R> g2 <- ggplot(data=resi2$out_1, aes(x=Expected,y=Observed))+
+    geom_point()+geom_abline(color="red")+ggtitle("(b)")+
+    theme_bw()+theme(
+      axis.text.x = element_text(color="black"),
+      axis.text.y = element_text(color="black"),
+      axis.ticks = element_line(color = "black")
+    )+ggtitle("(b)")+personal_title+ylab("")
R> gridExtra::grid.arrange(g1, g2, ncol=2)
```

Ajuste dos modelos binomial, beta-binomial e binomial normal com o preditor linar 3.28

```
##--Modelo binomial (MLG)
M11 <- combTMB(cbind(D3, IVC-D3) ~ Period + Status, embryos, family=binomial)
##--Modelo beta-binomial (MLG)
M12 <- combTMB(cbind(D3, IVC-D3) ~ Period + Status,  embryos,
               family=betabinomial)
##--Modelo binomial-normal (MLGM)
M13 <- combTMB(cbind(D3, IVC-D3) ~ Period + Status+(1|Donor)+(1|Sire),
               embryos, family=binomial)
```

O resultado dos ajustes desses modelos estão sumarizados na Tabela 3.7.

Valores preditos

Outro recurso importante no pacote **combTMB** é a função `predict()`. Os valores preditos podem ser obtidos na escala da variável resposta ou na escala da função de ligação. Além disso, na presença de efeitos aleatórios, as previsões podem ser feitas em nível individual (`re_form=NULL`) ou em nível de população (`re_form=~0` ou `re_form=NA`). Para obter os valores previstos na escala da variável resposta, juntamente com os seus respectivos erros-padrão, para o modelo beta-binomial-normal com preditor linear de equação (3.28), execute os seguintes comandos.

```
R> ##--Valores preditos
R> data_pred<-predict(M10,re_form=~0,type="response",se_fit=TRUE)
R> head(data_pred$fit) #valores preditos
mu_predict mu_predict mu_predict mu_predict mu_predict mu_predict
 0.7281765  0.7781390  0.7781390  0.7281765  0.7281765  0.7281765
R> head(data_pred$se_fit) #erros-padrão
mu_predict mu_predict mu_predict mu_predict mu_predict mu_predict
0.01521233 0.01030458 0.01030458 0.01521233 0.01521233 0.01521233
```

Os códigos utilizados para obter a Figura 3.4 podem ser encontrados em `https://github.com/deoclecioamorim/cap2_SM/blob/master/figprop.R`.

Apêndice B: material suplementar do Capítulo 4

Figuras suplementares S1 e S2

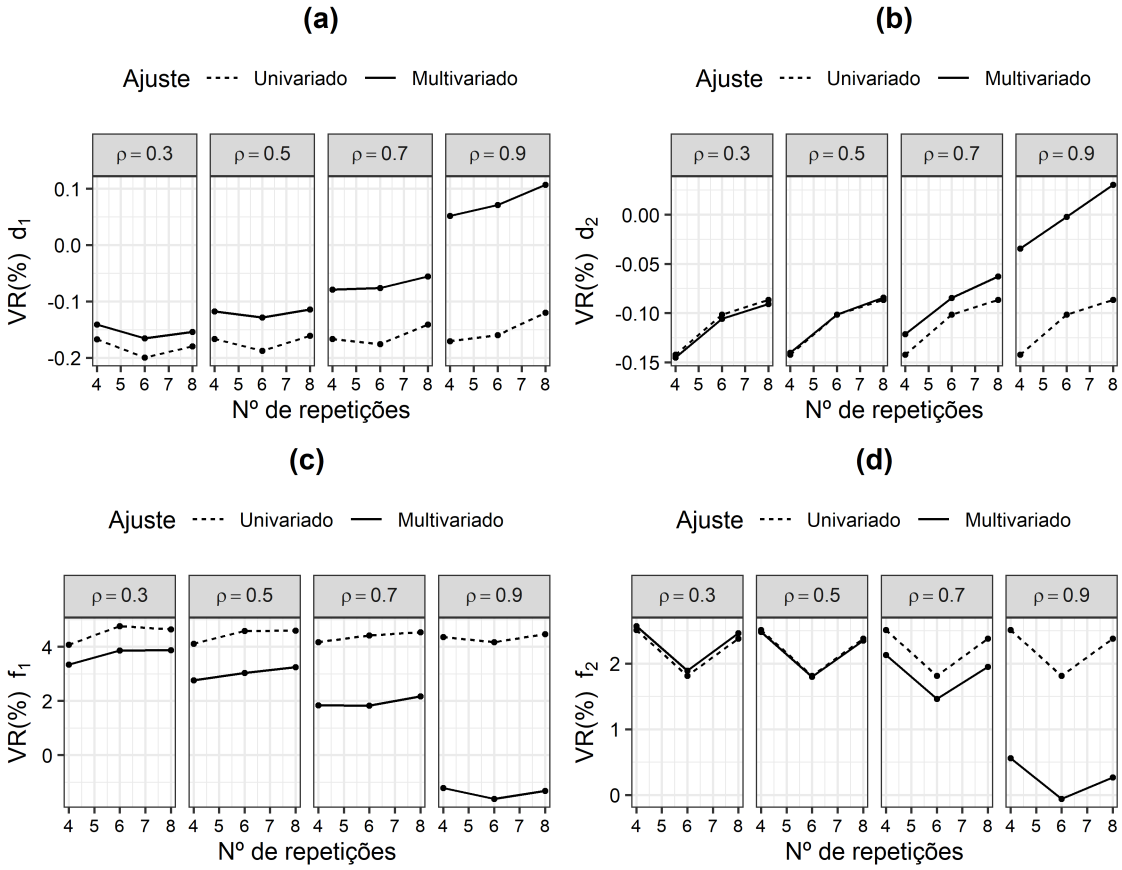

Figura S1. Estimativa do viés relativo (VR(%)), considerando 1000 simulações em experimentos com diferentes número de repetições ($r = 4, 6, 8$) por nível de dose do agente químico e diferentes níveis de correlação ($\rho = 0,3; 0,5; 0,7; 0,9$) entre os resíduos das variáveis resposta: (a) parâmetro d_1, (b) parâmetro d_2, (c) parâmetro f_1 e (d) parâmetro f_2.

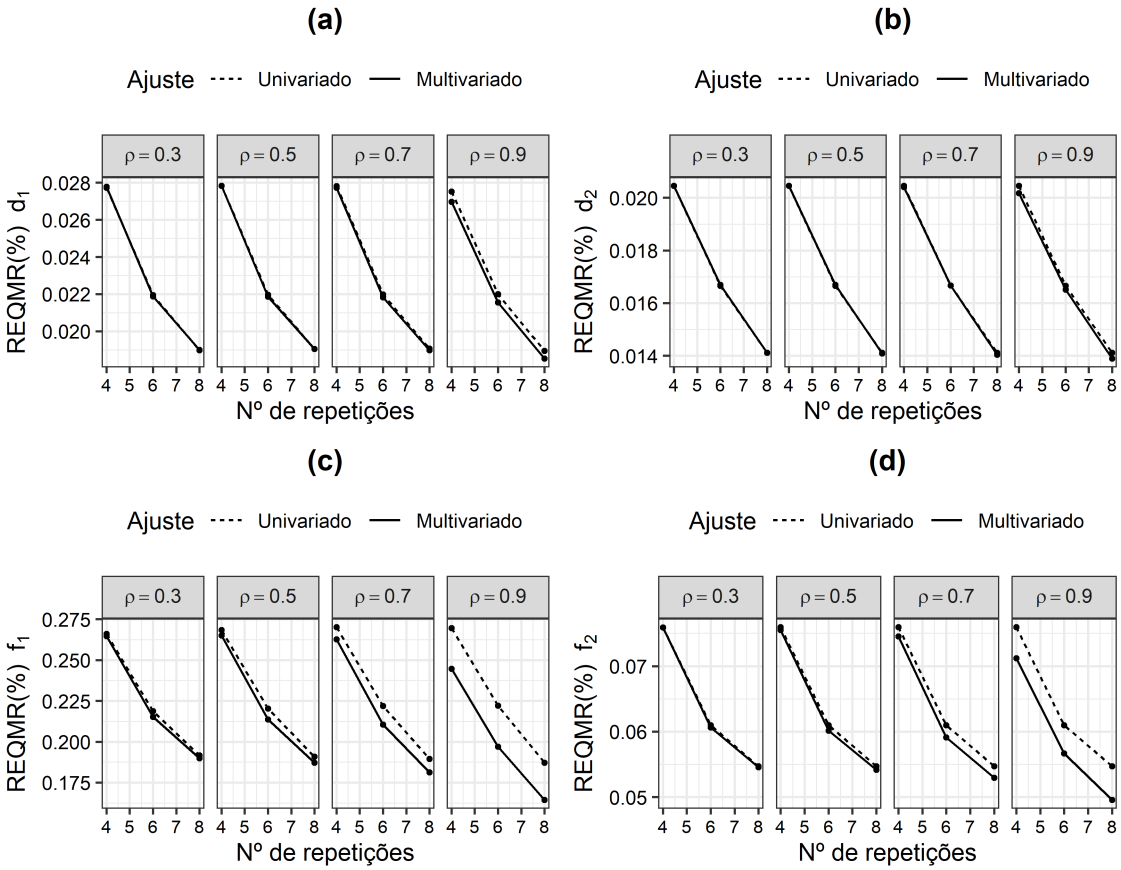

Figura S2. Estimativa da raiz erro quadrático médio relativo (REQMR(%)), considerando 1000 simulações em experimentos com diferentes número de repetições ($r = 4, 6, 8$) por nível de dose do agente químico e diferentes níveis de correlação ($\rho = 0,3; 0,5; 0,7; 0,9$) entre os resíduos das variáveis resposta: (a) parâmetro d_1, (b) parâmetro d_2, (c) parâmetro f_1 e (d) parâmetro f_2.

Tabelas suplementares S1-S4

Tabela S1. Resultados para os modelos não lineares univariados e bivariados de dose-resposta, baseado em 1000 simulações considerando um experimento com quatro repetições, sob diferentes níveis de correlação ($\rho = 0,3; 0,5; 0,7; 0,9$) entre os resíduos das variáveis resposta.

Ajuste	Correlação	Mensurações	d_1	e_1	b_1	f_1	d_2	e_2	b_2	f_2
Univariado	0,3	Estimativa	44,68520	54,36100	2,34843	0,79095	297,94572	60,55576	3,11746	5,43297
Univariado	0,3	e.p.	1,23984	6,33132	0,15387	0,19985	6,09304	3,31832	0,23741	0,72653
Univariado	0,3	VR	-0,16712	0,00184	-0,06684	4,07261	-0,14220	-0,63052	-0,71773	2,50882
Univariado	0,3	REQMR	2,77363	11,64120	6,54487	26,59651	2,04603	5,47890	7,59105	13,92909
Multivariado	0,3	Estimativa	44,69696	54,55337	2,35764	0,78534	297,93679	60,53959	3,11551	5,43618
Multivariado	0,3	e.p.	1,24266	6,35245	0,15431	0,19960	6,09213	3,31834	0,23712	0,72700
Multivariado	0,3	VR	-0,14085	0,35572	0,32511	3,33481	-0,14519	-0,65706	-0,78003	2,56936
Multivariado	0,3	REQMR	2,77845	11,68546	6,57094	26,46106	2,04594	5,48205	7,58799	13,94888
Univariado	0,5	Estimativa	44,68550	54,36541	2,34691	0,79119	297,94552	60,55485	3,11740	5,43304
Univariado	0,5	e.p.	1,24457	6,35213	0,15217	0,20169	6,09244	3,31815	0,23748	0,72635
Univariado	0,5	VR	-0,16645	0,00996	-0,13156	4,10434	-0,14227	-0,63202	-0,71983	2,51028
Univariado	0,5	REQMR	2,78413	11,67946	6,47347	26,84024	2,04584	5,47881	7,59338	13,92600
Multivariado	0,5	Estimativa	44,70728	54,71040	2,36328	0,78099	297,95098	60,56026	3,11783	5,43156
Multivariado	0,5	e.p.	1,24575	6,37477	0,15283	0,20039	6,09037	3,30709	0,23624	0,72541
Multivariado	0,5	VR	-0,11777	0,64460	0,56499	2,76222	-0,14044	-0,62313	-0,70614	2,48222
Multivariado	0,5	REQMR	2,78428	11,73879	6,52464	26,49840	2,04502	5,45976	7,55274	13,90350
Univariado	0,7	Estimativa	44,68550	54,35054	2,34515	0,79168	297,94543	60,55447	3,11737	5,43308
Univariado	0,7	e.p.	1,24365	6,37078	0,15125	0,20293	6,09220	3,31809	0,23750	0,72628
Univariado	0,7	VR	-0,16643	-0,01740	-0,20646	4,16802	-0,14230	-0,63264	-0,72069	2,51088
Univariado	0,7	REQMR	2,78208	11,71377	6,43628	27,01164	2,04576	5,47877	7,59433	13,92475
Multivariado	0,7	Estimativa	44,72464	54,93830	2,37263	0,77394	298,00765	60,64792	3,12746	5,41287
Multivariado	0,7	e.p.	1,24111	6,36589	0,15215	0,19930	6,08105	3,27055	0,23388	0,71876
Multivariado	0,7	VR	-0,07900	1,06384	0,96315	1,83420	-0,12144	-0,47929	-0,39937	2,12961

Continua na próxima página...

Tabela S1 – *Continuação da Tabela*

Ajuste	Correlação	Mensurações	d_1	e_1	b_1	f_1	d_2	e_2	b_2	f_2
Multivariado	0,7	REQMR	2,77254	11,75301	6,54242	26,27431	2,04069	5,38553	7,45524	13,72105
Univariado	0,9	Estimativa	44,68367	54,27950	2,34228	0,79308	297,94559	60,55520	3,11742	5,43301
Univariado	0,9	e.p.	1,23009	6,38259	0,15149	0,20221	6,09267	3,31822	0,23745	0,72642
Univariado	0,9	VR	-0,17053	-0,14808	-0,32845	4,35301	-0,14224	-0,63144	-0,71902	2,50971
Univariado	0,9	REQMR	2,75210	11,73640	6,45136	26,94688	2,04591	5,47884	7,59248	13,92719
Multivariado	0,9	Estimativa	44,78311	55,61288	2,40267	0,75077	298,26669	61,03832	3,16946	5,32961
Multivariado	0,9	e.p.	1,20789	6,07985	0,14901	0,18579	6,02215	3,07193	0,22180	0,67976
Multivariado	0,9	VR	0,05162	2,30477	2,24109	-1,21489	-0,03463	0,16134	0,93836	0,55877
Multivariado	0,9	REQMR	2,69774	11,41395	6,72206	24,46413	2,01764	5,04098	7,12234	12,83146

e.p. = erro padrão; REQMR(%) = estimativa da raiz erro quadrático médio relativo; VR(%) = estimativa do viés relativo.

Tabela S2. Resultados para os modelos não lineares univariados e bivariados, baseado em 1000 simulações considerando um experimento com seis repetições, sob diferentes níveis de correlação ($\rho = 0,3; 0,5; 0,7; 0,9$) entre os resíduos das variáveis resposta.

Ajuste	Correlação	Mensurações	d_1	e_1	b_1	f_1	d_2	e_2	b_2	f_2
Univariado	0,3	Estimativa	44,67072	53,73174	2,32576	0,79617	298,06675	60,57290	3,10577	5,39616
Univariado	0,3	e.p.	0,97853	5,05141	0,11794	0,16237	4,96426	2,73358	0,18842	0,59676
Univariado	0,3	VR	-0,19947	-1,15574	-1,03145	4,75942	-0,10164	-0,60240	-1,09015	1,81427
Univariado	0,3	REQMR	2,19416	9,35950	5,12106	21,87781	1,66607	4,52374	6,09598	11,39927
Multivariado	0,3	Estimativa	44,68598	53,95687	2,33507	0,78937	298,05442	60,55254	3,10324	5,40027
Multivariado	0,3	e.p.	0,97684	5,05567	0,11742	0,16092	4,97226	2,73508	0,18683	0,59735
Multivariado	0,3	VR	-0,16538	-0,74159	-0,63525	3,86487	-0,10577	-0,63580	-1,17061	1,89194
Multivariado	0,3	REQMR	2,18757	9,32524	5,03449	21,51337	1,66900	4,53075	6,06106	11,42280

Continua na próxima página...

Tabela S2 – *Continuação da Tabela*

Ajuste	Correlação	Mensurações	d_1	e_1	b_1	f_1	d_2	e_2	b_2	f_2
Univariado	0,5	Estimativa	44,67590	53,76368	2,32576	0,79484	298,06700	60,57320	3,10579	5,39610
Univariado	0,5	e.p.	0,98015	5,07765	0,11849	0,16391	4,96469	2,73397	0,18847	0,59673
Univariado	0,5	VR	-0,18790	-1,09698	-1,03151	4,58402	-0,10155	-0,60191	-1,08940	1,81315
Univariado	0,5	REQMR	2,19674	9,40035	5,14412	22,03807	1,66620	4,52431	6,09746	11,39863
Multivariado	0,5	Estimativa	44,70255	54,15111	2,34199	0,78303	298,06722	60,57514	3,10625	5,39547
Multivariado	0,5	e.p.	0,97727	5,06965	0,11777	0,16079	4,97391	2,72684	0,18593	0,59571
Multivariado	0,5	VR	-0,12835	-0,38427	-0,34103	3,03061	-0,10148	-0,59872	-1,07490	1,80132
Multivariado	0,5	REQMR	2,18603	9,32932	5,02054	21,36226	1,66928	4,51229	6,01509	11,37768
Univariado	0,7	Estimativa	44,68126	53,79512	2,32575	0,79353	298,06710	60,57332	3,10580	5,39607
Univariado	0,7	e.p.	0,98149	5,11454	0,11932	0,16529	4,96486	2,73413	0,18850	0,59672
Univariado	0,7	VR	-0,17591	-1,03915	-1,03207	4,41185	-0,10152	-0,60171	-1,08909	1,81270
Univariado	0,7	REQMR	2,19873	9,46118	5,17868	22,18041	1,66626	4,52454	6,09807	11,39837
Multivariado	0,7	Estimativa	44,72587	54,43491	2,35319	0,77388	298,11785	60,66030	3,11680	5,37761
Multivariado	0,7	e.p.	0,97684	5,07382	0,11842	0,15940	4,96937	2,69777	0,18422	0,59006
Multivariado	0,7	VR	-0,07624	0,13780	0,13563	1,82632	-0,08451	-0,45897	-0,73871	1,46431
Multivariado	0,7	REQMR	2,18264	9,33008	5,03853	21,04224	1,66682	4,44846	5,91036	11,22362
Univariado	0,9	Estimativa	44,68850	53,84587	2,32634	0,79165	298,06690	60,57308	3,10578	5,39612
Univariado	0,9	e.p.	0,98256	5,18283	0,12088	0,16586	4,96453	2,73382	0,18845	0,59674
Univariado	0,9	VR	-0,15975	-0,94579	-1,00699	4,16484	-0,10159	-0,60210	-1,08970	1,81359
Univariado	0,9	REQMR	2,19988	9,57632	5,23901	22,20676	1,66615	4,52409	6,09689	11,39887
Multivariado	0,9	Estimativa	44,79171	55,25414	2,38821	0,74766	298,36345	61,04609	3,16184	5,29677
Multivariado	0,9	e.p.	0,96490	4,90326	0,11784	0,14925	4,92730	2,54077	0,17672	0,56024
Multivariado	0,9	VR	0,07085	1,64485	1,62587	-1,62407	-0,00220	0,17408	0,69559	-0,06092
Multivariado	0,9	REQMR	2,15580	9,16429	5,26927	19,69516	1,65058	4,17085	5,66798	10,56550

e.p. = erro padrão; REQMR(%) = estimativa da raiz erro quadrático médio relativo; VR(%) = estimativa do viés relativo.

Tabela S3. Resultados para os modelos não lineares univariados e bivariados de dose-resposta, baseado em 1000 simulações considerando um experimento com oito repetições, sob diferentes níveis de correlação ($\rho = 0,3$, $\rho = 0,5$, $\rho = 0,7$ e $\rho = 0,9$) entre os resíduos das variáveis resposta.

Ajuste	Correlação	Mensurações	d_1	e_1	b_1	f_1	d_2	e_2	b_2	f_2
Univariado	0,3	Estimativa	44,67947	53,66947	2,32389	0,79524	298,11116	60,40560	3,09642	5,42622
Univariado	0,3	e.p.	0,84607	4,32194	0,10694	0,14134	4,20617	2,23355	0,16621	0,50268
Univariado	0,3	VR	-0,17992	-1,27029	-1,11112	4,63733	-0,08675	-0,87692	-1,38781	2,38146
Univariado	0,3	REQMR	1,89783	8,04750	4,68215	19,15830	1,41168	3,76682	5,46955	9,77433
Multivariado	0,3	Estimativa	44,69104	53,86840	2,33340	0,78945	298,09922	60,38319	3,09346	5,43058
Multivariado	0,3	e.p.	0,84779	4,34756	0,10739	0,14130	4,20335	2,22396	0,16503	0,50137
Multivariado	0,3	VR	-0,15406	-0,90434	-0,70636	3,87457	-0,09075	-0,91370	-1,48215	2,46381
Multivariado	0,3	REQMR	1,89940	8,04471	4,62197	18,98193	1,41099	3,76030	5,45831	9,77082
Univariado	0,5	Estimativa	44,68804	53,66225	2,32317	0,79489	298,11145	60,40567	3,09643	5,42615
Univariado	0,5	e.p.	0,85020	4,29375	0,10723	0,14083	4,20521	2,23310	0,16620	0,50254
Univariado	0,5	VR	-0,16076	-1,28358	-1,14182	4,59113	-0,08666	-0,87682	-1,38763	2,38023
Univariado	0,5	REQMR	1,90531	7,99845	4,70161	19,08191	1,41135	3,76608	5,46928	9,77139
Multivariado	0,5	Estimativa	44,70882	54,00908	2,33971	0,78468	298,11823	60,41178	3,09643	5,42449
Multivariado	0,5	e.p.	0,85170	4,32218	0,10796	0,14017	4,19650	2,20708	0,16449	0,49828
Multivariado	0,5	VR	-0,11434	-0,64554	-0,43794	3,24696	-0,08438	-0,86679	-1,38767	2,34884
Multivariado	0,5	REQMR	1,90529	7,97324	4,61255	18,71748	1,40830	3,72225	5,41682	9,68586
Univariado	0,7	Estimativa	44,69690	53,65237	2,32247	0,79448	298,11156	60,40570	3,09643	5,42613
Univariado	0,7	e.p.	0,85144	4,26495	0,10778	0,13995	4,20481	2,23291	0,16620	0,50248
Univariado	0,7	VR	-0,14097	-1,30176	-1,17141	4,53743	-0,08662	-0,87677	-1,38757	2,37973
Univariado	0,7	REQMR	1,90651	7,94914	4,73126	18,95668	1,41122	3,76577	5,46916	9,77018
Multivariado	0,7	Estimativa	44,73513	54,25243	2,35048	0,77646	298,18230	60,51342	3,10752	5,40339
Multivariado	0,7	e.p.	0,84970	4,25493	0,10823	0,13692	4,18558	2,17048	0,16314	0,49114
Multivariado	0,7	VR	-0,05556	-0,19788	0,02028	2,16523	-0,06291	-0,69999	-1,03452	1,95084

Continua na próxima página...

Tabela S3 – *Continuação da Tabela*

Ajuste	Correlação	Mensurações	d_1	e_1	b_1	f_1	d_2	e_2	b_2	f_2
Multivariado	0,7	REQMR	1,89820	7,82591	4,60344	18,13706	1,40353	3,62805	5,29486	9,46540
Univariado	0,9	Estimativa	44,70642	53,63813	2,32176	0,79393	298,11134	60,40564	3,09643	5,42618
Univariado	0,9	e.p.	0,84640	4,23323	0,10831	0,13825	4,20558	2,23327	0,16620	0,50259
Univariado	0,9	VR	-0,11971	-1,32794	-1,20164	4,46414	-0,08669	-0,87686	-1,38770	2,38071
Univariado	0,9	REQMR	1,89381	7,89598	4,76085	18,72159	1,41148	3,76636	5,46939	9,77252
Multivariado	0,9	Estimativa	44,80794	55,07079	2,38561	0,74991	298,46095	60,94624	3,15509	5,31405
Multivariado	0,9	e.p.	0,82848	3,97924	0,10498	0,12456	4,14690	2,02212	0,15497	0,46212
Multivariado	0,9	VR	0,10710	1,30756	1,51551	-1,32817	0,03048	0,01024	0,48043	0,26502
Multivariado	0,9	REQMR	1,85311	7,43242	4,71530	16,43466	1,38949	3,31656	4,95634	8,71891

e.p. = erro padrão; REQMR(%) = estimativa da raiz erro quadrático médio relativo; VR(%) = estimativa do viés relativo.

Tabela S4. Estimativas das características quantitativas sub-NOAEL considerando os modelos univariados e multivariados de Brain e Cousens (1989) com os respectivos erros-padrão (e.p.), para as variáveis AP, AF e MMS de *C. tinctorius* mensuradas aos 28 após o tratamento com baixas doses de glifosato, sob estresse severo (−70kPa).

Variável	Parâmetros	Univariado		Multivariado	
		Estimativa	e.p.	Estimativa	e.p.
AP [cm]	M_1 [g e.a. ha^{-1}]	23,35610	1,82260	23,64154	1,85054
AF [cm^2]	M_2 [g e.a. ha^{-1}]	31,82540	2,30730	32,64291	2,37495
MMS [g]	M_3 [g e.a. ha^{-1}]	19,35833	2,53481	20,26787	2,56223
AP [cm]	$Ymax[\%]_1$	122,77510	4,77220	122,14900	4,72100
AF [cm^2]	$Ymax[\%]_2$	138,54440	5,60320	137,66970	5,54550
MMS [g]	$Ymax[\%]_3$	128,84350	6,8439	128,42300	6,67280
AP [cm]	LDS_1 [g e.a. ha^{-1}]	51,23430	3,15790	51,39338	3,20000
AF [cm^2]	LDS_2 [g e.a. ha^{-1}]	63,23430	2,61780	63,70921	2,60942
MMS [g]	LDS_3 [g e.a. ha^{-1}]	52,14845	5,20799	50,60980	4,12722
AP [cm]	$ED_{50;1}$ [g e.a. ha^{-1}]	103,23060	5,10390	103,26977	5,06278
AF [cm^2]	$ED_{50;2}$ [g e.a. ha^{-1}]	98,28670	4,82160	97,92625	4,77620
MMS [g]	$ED_{50;3}$ [g e.a. ha^{-1}]	110,54029	18,99287	102,08892	14,95253

M [g e.a. ha^{-1}] = dose que causa estimulação máxima; $Ymax[\%]$ = resposta estimulatória relativa acima do controle; LDS [g e.a. ha^{-1}] = dose limite para ocorrência de hormesis; ED_{50} [g e.a. ha^{-1}] = dose que causa a redução de 50% na resposta média do controle não tratado.

Códigos utilizados no Capítulo 4

A seguir, são apresentados os códigos em `R` utilizados para a modelagem da hormese por meio de regressão não-linear multivariada, considerando o estudo de caso apresentado na Seção 4.5.

```
#########################################################################
##--Modelagem da hormesis por regressão não linear multivariada
#########################################################################
##--Chamando os pacotes e o conjunto de dados
library(readxl)
library(MVN)
library(minpack.lm)
library(MASS)
library(car)
Dados<-read_excel("Dados/Dados.xlsx",sheet =1)
##--Ajustando o modelo de Modelo de Brain-Cousens (1989)
##--Etapa I: Ajuste individual dos modelos de regressão não linear
```

```
##--Altura de planta (AP)
c=0 #Parâmetros fixado
mode1.AP <-nlsLM(plant_height~
                 (c+(((d-c)+(f*Dose))/(1+exp(b*log(Dose/E))))),
                 start =list(d=54,E=100,b=2,f=11),
                 control = nls.lm.control(maxiter = 600),
                 algorithm="LM",trace= T,data =Dados)

##--Área folia (AF)
c=0 #Parâmetros fixado
mode1.AF <- nlsLM(leaf_area~
                      (c+(((d-c)+(f*Dose))/(1+exp(b*log(Dose/E))))),
                       start =list(d=416,E=100,b=2,f=100),
                       control = nls.lm.control(maxiter = 600),
                       algorithm="LM",trace= T,data =Dados)

##--Massa de matéria seca (MMS)
mode1.MMS <- nlsLM(dry_matter_mass~
                     (c+(((d-c)+(f*Dose))/(1+exp(b*log(Dose/E))))),
                      start =list(d=6,E=40,c=1.5,b=2,f=2),
                      control = nls.lm.control(maxiter = 600),
                      algorithm="LM",trace= T,data =Dados)

##--Etapa II: Obtenção da estimativa da matriz de variâncias
##--e covariâncias dos resíduos

N<-length(Dados$plant_height) #comprimento dos dados
M<-3 #Três modelos univariados

#Construindo a matriz Sigma para as variáveis AP, AF e MMS
resid_m1<-as.matrix(residuals(mode1.AP))
(resid_m1<-t(resid_m1))
resid_m2<-as.matrix(residuals(mode1.AF))
(resid_m2<-t(resid_m2))
resid_m3<-as.matrix(residuals(mode1.MMS))
(resid_m3<-t(resid_m3))
Matrix.erros_m<-matrix(c(resid_m1,resid_m2,resid_m3),
                       nrow = M, ncol = N,byrow = T)
##--Matriz Sigma
```

```
SIGMA<-((Matrix.erros_m)%*%t(Matrix.erros_m))/N

##--Etapa III: Composição do modelo não linear multivariado
##--e obtenção da otimização única
##--Inversa de Moore Penrose
inv_Sigma<-ginv(SIGMA)
cf <- chol(inv_Sigma) #Fator de Cholesky

j<-matrix(1,N,1) #Vetor de 1's
In<-diag(1,N,N)  #Matriz identidade
p1<-kronecker(diag(1, M), j) %*% cf[,1]
p2<-kronecker(diag(1, M), j) %*% cf[,2]
p3<-kronecker(diag(1, M), j) %*% cf[,3]

#Y empilhado
y_emp<-c(Dados$plant_height,Dados$leaf_area,Dados$dry_matter_mass)
Yr<-kronecker(cf, In)%*%y_emp
#########################################################################
# Modelo não linear multivariado: Resultado apresentado na Tabela 4.1
#########################################################################
##--Parâmetros fixados
c1=0;
c2=0;
mult1<-nlsLM(Yr~p1*(c1+(((d1-c1)+(f1*Dose))/(1+exp(b1*log(Dose/E1)))))+
               p2*(c2+(((d2-c2)+(f2*Dose))/(1+exp(b2*log(Dose/E2)))))+
               p3*(c3+(((d3-c3)+(f3*Dose))/(1+exp(b3*log(Dose/E3))))),
             start =list(d1=44, f1=0.7, b1=2, E1=55,
                         d2=300,f2=5, b2=2, E2=60,
                         c3=3, d3=4,f3=0.08, b3=2, E3=40),
             algorithm="LM",trace= T,data =Dados)
summary(mult1)

##--Teste de hormesis no contexto multivariado empregando
##--o teste da razão de verossimilhanças

#--Modelo reduzido
mult2<-nlsLM(Yr~p1*(c1+(((d1-c1))/(1+exp(b1*log(Dose/E1)))))+
               p2*(c2+(((d2-c2))/(1+exp(b2*log(Dose/E2)))))+
               p3*(c3+(((d3-c3))/(1+exp(b3*log(Dose/E3))))),
```

```
              start =list(d1=44, b1=2, E1=55,
                          d2=300, b2=2, E2=60,
                          c3=3,   d3=4, b3=2, E3=40),
              algorithm="LM",trace= T,data =Dados)

summary(mult2)

##--Teste LRT
anova(mult2,mult1)
qf(0.95,3,227)
```

Exemplo do script utilizado para obter a estimativa da dose M (dose que causa estimulação máxima). O restante do código utilizado pode ser encontrado em `https://github.com/deoclecioamorim/MNMH.git`.

```
#############################################################################
# Ajustes univariados: Dose-M
#############################################################################
##--Dose M
##--Altura de planta (AP)
c1=0; #Parâmetro fixado
M_AP<-nlsLM(plant_height~
               (c1+((d1-c1)+(f1*Dose))/(1+(f1*M1/(((d1-c1)*b1)
               -f1*M1*(1-b1)))*exp(b1*log(Dose/M1)))),
               start =list(d1=44, M1=20,b1=2,f1=0.7),
               algorithm="LM",trace= T,
               control = nls.lm.control(maxiter = 600),data =Dados)

##--Área foliar (AF)
c2=0  #Parâmetro fixado
M_AF<-nlsLM(leaf_area~
             (c2+((d2-c2)+(f2*Dose))/(1+(f2*M2/(((d2-c2)*b2)
             -f2*M2*(1-b2)))*exp(b2*log(Dose/M2)))),
             start =list(d2=300,M2=30,b2=2,f2=5),
             algorithm="LM",trace= T,
             control = nls.lm.control(maxiter = 600),data =Dados)

##--Massa de matéria seca (MSS)
M_MSS <-nlsLM(dry_matter_mass~(c3+((d3-c3)+(f3*Dose))
               /(1+(f3*M3/(((d3-c3)*b3)-f3*M3*(1-b3)))*
```

```
              exp(b3*log(Dose/M3)))),
              start =list(d3=4,  M3=20,b3=2,f3=0.08,c3=2),
              algorithm="LM",trace= T,
              control = nls.lm.control(maxiter = 600),data =Dados)
##--Extraindo informações para construção da Tabela S4
summary(M_AP)
summary(M_AF)
summary(M_MSS)
#########################################################################
#         Dose-M: Modelo não linear multivariado
#########################################################################
doseM<-nlsLM(Yr~p1*(c1+((d1-c1)+(f1*Dose))/(1+(f1*M1/
            (((d1-c1)*b1)-f1*M1*(1-b1)))*exp(b1*log(Dose/M1))))+
            p2*(c2+((d2-c2)+(f2*Dose))/(1+(f2*M2/
            (((d2-c2)*b2)-f2*M2*(1-b2)))*exp(b2*log(Dose/M2))))+
            p3*(c3+((d3-c3)+(f3*Dose))/(1+(f3*M3/
            (((d3-c3)*b3)-f3*M3*(1-b3)))*exp(b3*log(Dose/M3)))),
             start =list(d1=44, M1=20,b1=2,f1=0.7,
                         d2=300,M2=30,b2=2,f2=5,
                         d3=4,  M3=20,b3=2,f3=0.08,c3=2),
                         algorithm="LM",trace= T,
          control = nls.lm.control(maxiter = 600),data =Dados)
summary(doseM)
##--Dose M média
car::deltaMethod(doseM,"(M1+M2+M3)/3",vcov. = vcov(doseM))
```