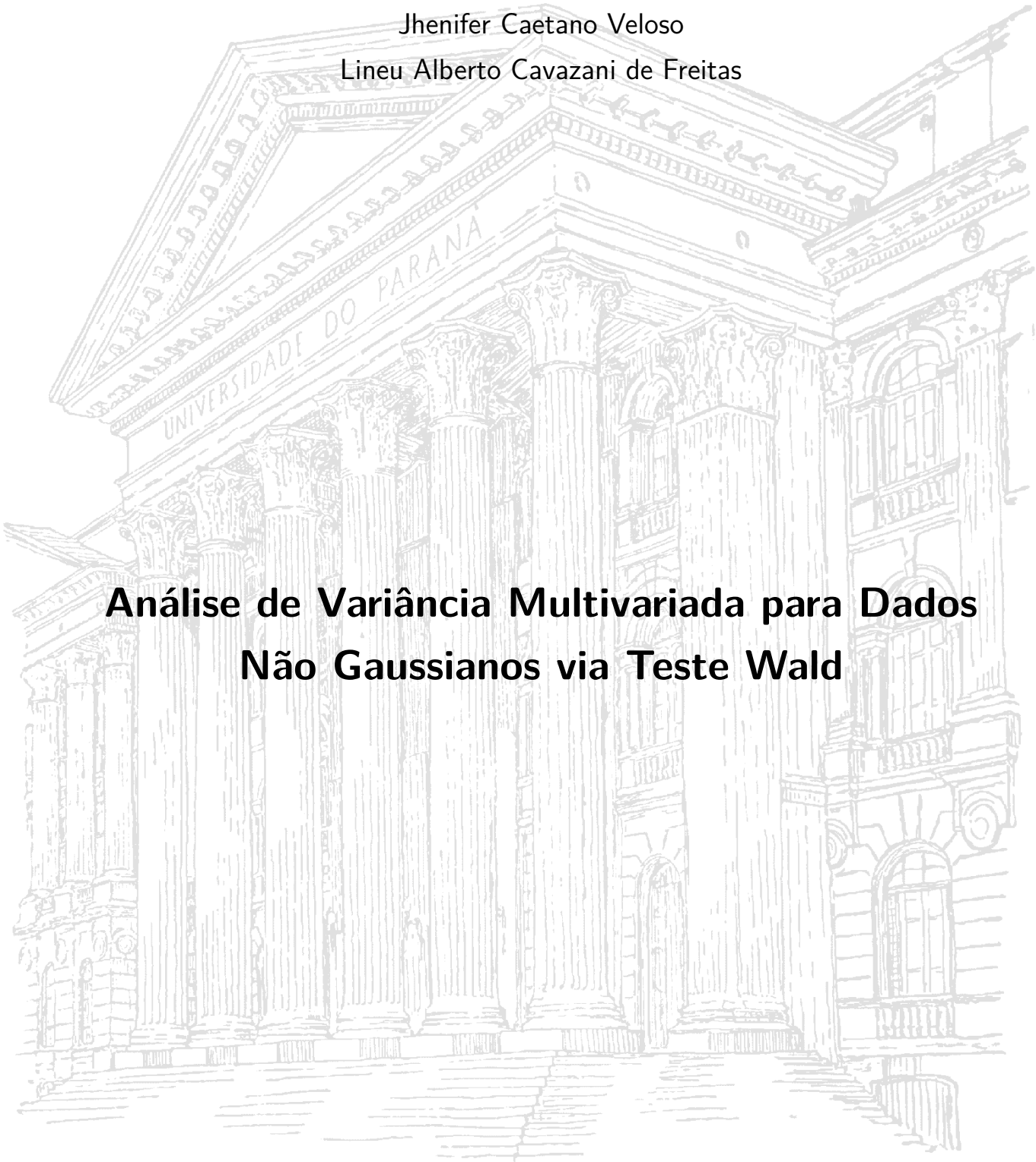
Universidade Federal do Paraná

Jhenifer Caetano Veloso

Lineu Alberto Cavazani de Freitas



Análise de Variância Multivariada para Dados Não Gaussianos via Teste Wald

Curitiba

2019

Jhenifer Caetano Veloso
Lineu Alberto Cavazani de Freitas

Análise de Variância Multivariada para Dados Não Gaussianos via Teste Wald

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

Universidade Federal do Paraná

Setor de Ciências Exatas

Departamento de Estatística

Orientador: Prof. Dr. Wagner Hugo Bonat

Curitiba

2019

Resumo

No âmbito da Estatística aplicada, modelos de regressão são uma das principais e mais difundidas ferramentas utilizadas em diversas áreas do conhecimento, sendo que o caso mais conhecido é o modelo linear normal. Todavia, há casos em que são coletadas mais de uma resposta por unidade experimental e há o interesse de modelá-las em função de um conjunto de variáveis explicativas. Para tal, uma alternativa são os modelos lineares multivariados. Porém, por maior que seja seu potencial de aplicação, essa classe apresenta limitações como a necessidade de normalidade multivariada, homogeneidade das matrizes de variâncias e covariâncias, além de independência entre as observações. Isto posto, uma alternativa para solucionar tais limitações são os Modelos Multivariados de Covariância Linear Generalizada que permitem lidar com múltiplas respostas e de diferentes naturezas e, de alguma forma, correlacionadas. Para essa classe de modelos, neste trabalho é proposto e implementado o teste de Wald para análise de variância multivariada para dados não gaussianos. Seu comportamento foi discutido através de estudos de simulação, a fim de verificar o poder do teste sob diferentes distribuições da variável resposta, diferentes tamanhos amostrais e diferentes valores para os parâmetros de regressão, apresentando, de forma geral, resultados satisfatórios para tamanhos amostrais maiores que 250 em todas as distribuições de variável resposta simuladas. O emprego do teste proposto é ilustrado através de dois conjuntos de dados, um com respostas binárias e outro com respostas de contagem e proporção.

Palavras-chave: Modelos multivariados de covariância linear generalizada; teste Wald; MANOVA; não gaussianos; correlação.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

No âmbito da Estatística aplicada, modelos de regressão são uma das principais e mais difundidas ferramentas utilizadas em diversas áreas do conhecimento, sendo comum o interesse em i) explicar a associação entre uma variável resposta e um conjunto de variáveis explicativas e ii) utilizar o modelo para realizar predições para uma população.

Os modelos de regressão, nos casos univariados mais gerais, associam uma única variável resposta, também chamada de variável dependente, a uma ou mais variáveis explicativas, conhecidas como variáveis independentes. De forma geral, um modelo de regressão é uma expressão matemática que relaciona a média da variável resposta às variáveis preditoras (covariáveis), em que a variável resposta segue uma distribuição de probabilidade condicional às covariáveis e a média é descrita por um preditor linear. O caso mais conhecido é o modelo linear normal, o qual estabelece que a variável resposta condicional às variáveis explicativas segue distribuição Normal.

Todavia, não são raras as situações em que o fenômeno sob estudo não apresenta uma variável resposta em que a suposição de normalidade seja atendida. Uma alternativa adotada é buscar uma transformação da variável resposta a fim de atender os pressupostos do modelo, tal como a transformação Box-Cox (BOX; COX, 1964). Contudo, este tipo de solução leva a dificuldades na interpretação dos resultados.

Neste contexto, a proposta de maior renome foi apresentada por Nelder e Wedderburn (1972), que introduz os Modelos Lineares Generalizados (MLGs). Essa classe de modelos flexibilizou a distribuição da variável resposta, permitindo que esta pertença à família exponencial de distribuições. Em meio aos casos especiais de distribuições possíveis nesta classe de modelos estão a Bernoulli, Binomial, Poisson, Normal, Gama, Normal inversa, entre outras. Trata-se portanto, de uma classe de modelos de regressão univariados para dados de diferentes naturezas, tais como dados contínuos simétricos e assimétricos, contagens, proporções, assim por diante. Tais características tornam esta classe uma flexível ferramenta de modelagem aplicável a diversos tipos de problema.

Embora as técnicas citadas sejam úteis, há casos em que são coletadas mais de uma resposta por unidade experimental e há o interesse de modelá-las em função de um conjunto de variáveis explicativas. Para problemas com essa estrutura, uma alternativa são os modelos lineares multivariados, nos quais associa-se um conjunto de respostas a uma ou mais covariáveis e, além disso, as variáveis respostas seguem distribuição Normal multivariada. Porém, por maior que seja seu potencial de aplicação, essa classe apresenta limitações como a necessidade de normalidade multivariada, homogeneidade

das matrizes de variâncias e covariâncias, além de independência entre as observações.

Uma alternativa para solucionar tais limitações são os Modelos Multivariados de Covariância Linear Generalizada (MCGLMs), prospostos por Bonat e Jørgensen (2016). Essa classe permite lidar com múltiplas respostas de diferentes naturezas e, de alguma forma, correlacionadas, em que essa correlação é estruturada em termos de uma função de ligação de covariância com um preditor linear matricial que envolve matrizes conhecidas. De forma geral, o MCGLM é uma estrutura para modelagem de múltiplas respostas, de diferentes naturezas e não necessariamente independentes, com extensões multivariadas para medidas repetidas, séries temporais, dados longitudinais, espaciais e espaço-temporais.

Para testar os efeitos das covariáveis sobre a variável resposta, no modelo linear univariado, faz-se uso da análise de variância (ANOVA), a qual utiliza o teste F para testar a hipótese de nulidade dos parâmetros do modelo ajustado, verificando assim, se o modelo tem poder de explicação. Quando se está na classe de modelos multivariados para dados gaussianos, extende-se a análise de variância para a análise de variância multivariada (MANOVA) (SMITH; GNANADESIKAN; HUGHES, 1962) e dentre os testes de hipótese multivariados já discutidos na literatura, destacam-se o λ de Wilk's (WILKS, 1932), traço de Hotelling-Lawley (LAWLEY, 1938; HOTELLING, 1951), traço de Pillai (PILLAI et al., 1955) e maior raiz de Roy (ROY, 1953).

No entanto, considerando um cenário com múltiplas respostas não gaussianas, são escassas as discussões na literatura a respeito de testes de hipóteses para dados multivariados. Portanto, o presente trabalho visa colaborar com a literatura i) propondo e implementando o teste de Wald para análise de variância multivariada para dados não gaussianos e ii) discutindo as propriedades e comportamento do teste proposto com base em estudos de simulação e aplicação a conjuntos de dados reais.

Este trabalho está organizado em seis capítulos: na atual seção foi exposto o tema de forma a enfatizar as características dos modelos lineares e testes de hipóteses. No Capítulo 2 são apresentados os conjuntos de dados utilizados para análise. O Capítulo 3 é dedicado à revisão bibliográfica da estrutura dos MCGLM e à apresentação do teste proposto. O Capítulo 4 apresenta os resultados dos estudos de simulação para verificar as principais propriedades do teste proposto. O Capítulo 5 apresenta a aplicação do método aos conjuntos de dados apresentados no Capítulo 2. E, por fim, no Capítulo 6 são apresentados os comentários finais e conclusões a respeito do estudo.

2 Conjuntos de Dados

Este capítulo é destinado à apresentação dos conjuntos de dados utilizados para avaliar o desempenho do teste de hipóteses proposto. O primeiro conjunto de dados, refere-se a processos movidos contra grandes litigantes e o segundo trata-se de um estudo sobre comportamento de ovelhas. Para ambos os conjuntos de dados, estão expostos neste capítulo 1) descrição do experimento ou estudo, 2) definição das variáveis respostas e covariáveis e 3) apresentação de características que podem ser relevantes para análise.

2.1 Processos Movidos Contra Grandes Litigantes

Conjunto de dados proveniente da Consulta de Processos de Primeiro Grau (CJPG) do Tribunal de Justiça de São Paulo composto de sentenças em primeira instância de processos homologados em 2014 e movidos contra um número reduzido de réus, conhecidos como grandes litigantes. Os dados foram obtidos por Trecenti (2015) que, na ocasião, propôs modelos preditivos para os resultados dos processos utilizando diagramas de influência, porém fora adaptado para aplicação neste trabalho. Para fins de análise, o conjunto de dados contém 8578 observações e as seguintes 13 variáveis:

- `foro`: Nome do foro onde o processo foi executado.
- `tipo_vara`: Fator de 3 níveis que indica o tipo de vara em que transcorre o processo: vara cível, Juizados Especiais Cíveis (JEC) com advogado ou JEC sem advogado.
- `adv_reu`: Fator de 2 níveis que indica presença de advogado do autor da ação.
- `empresa`: Empresa envolvida como ré da ação.
- `tipo_dano`: Fator de 3 níveis que indica o tipo de dano: material, moral ou ambos.
- `valor_acao`: Valor da ação, em salários mínimos.
- `serasa`: Fator de 2 níveis que indica a presença de discussões a respeito dos órgãos de proteção ao crédito.
- `terceiro`: Fator de 2 npiveis que indica a presença de discussões sobre terceiros.
- `consumo`: Fator de 2 níveis que indica a presença de discussões a respeito de relação de consumo.
- `gratuidade`: Fator de 2 níveis que indica a presença de discussões a respeito de gratuidade judiciária.

- `procedente`: Fator de 2 níveis que indica sentença procedente, isto é, se houve ganho de causa para o autor da ação.
- `improcedente`: Fator de 2 níveis que indica sentença improcedente, ou seja, se não houve ganho de causa para o autor da ação.
- `acordo`: Fator de 2 níveis que indica sentença acordo, isto é, se houve acordo entre as partes do processo.

As variáveis `procedente`, `improcedente` e `acordo` são as variáveis respostas desse estudo, pois conjuntamente, determinam o resultado do processo. A Figura 1 apresenta a frequência observada para cada uma das respostas. Nota-se que o resultado mais frequente é a procedência da ação, ou seja, quando houve ganho de causa para o autor da ação. Apesar das diferentes proporções de ocorrência das respostas, todos os resultados são razoavelmente frequentes, permitindo o ajuste do modelo para todas as respostas.

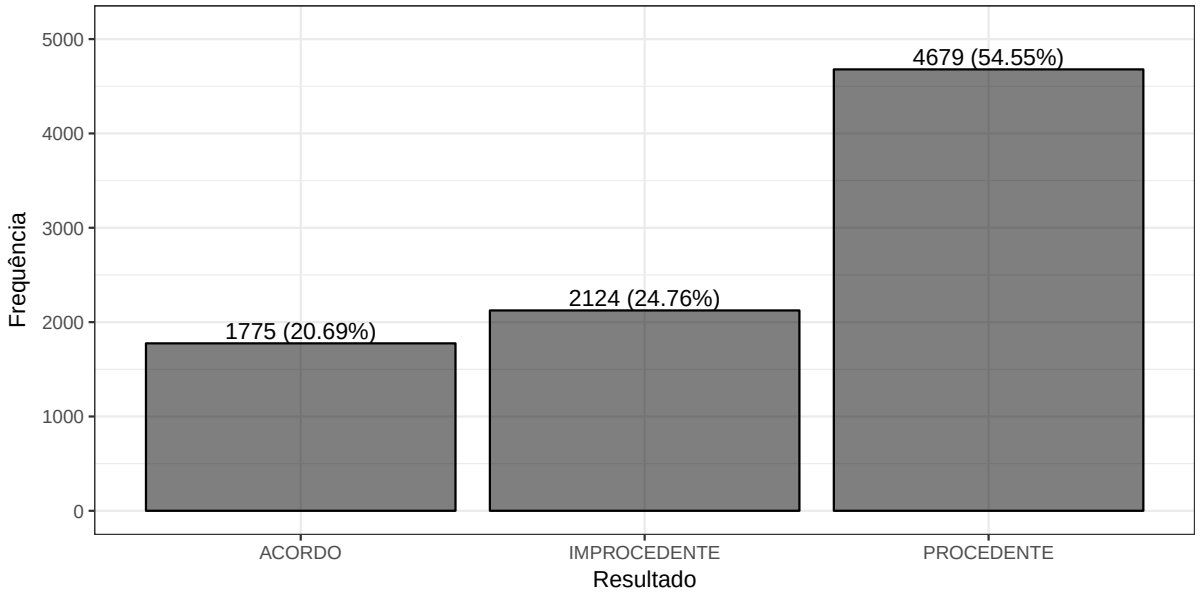


Figura 1 – Frequência observada para cada tipo de resultado dos processos.

Com exceção do valor da ação, todas as covariáveis são categóricas. As proporções de processos nas categorias de cada covariável são suficientemente altas, o que permite o ajuste do modelo. O valor da ação ou valor da causa é um valor usado como base para cálculo das custas judiciais, cada ação possui um valor diferente baseado em critérios diferentes, em alguns casos esse valor é estipulado por lei. Na Figura 2 são apresentadas as distribuições dos valores das ações, à esquerda na escala original e à direita na escala logarítmica de base 10. Devido à forte assimetria à esquerda, para prosseguir com as análises, utilizou-se os valores das ações na escala $\log_{10}$.

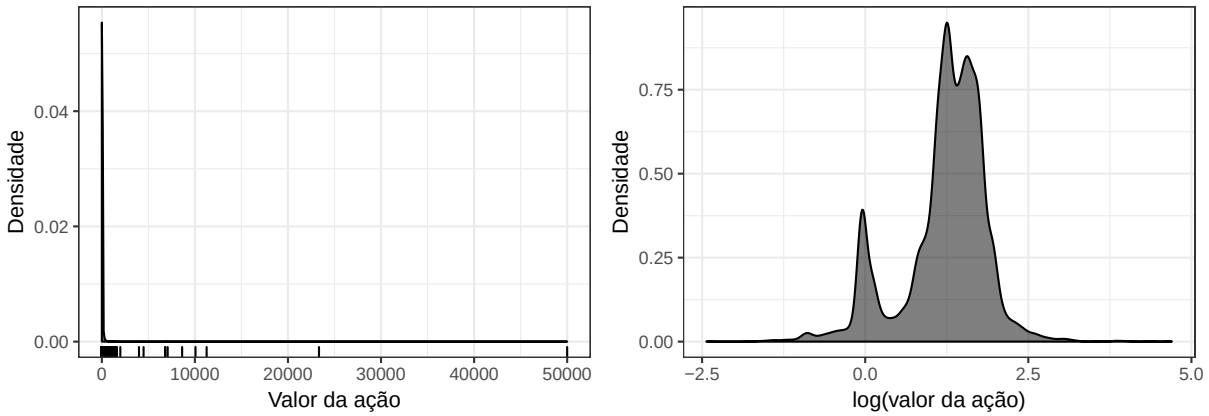


Figura 2 – Densidade do valor da ação (esquerda) e densidade do logarítmo na base 10 do valor da ação (direita).

2.2 Comportamento de Ovelhas Submetidas à Intervenção Humana

Conjunto de dados proveniente de um experimento sobre o comportamento de ovelhas, conduzido na fazenda experimental INRA La Fage, Roqueford, França, em setembro de 2015 com o objetivo de verificar o efeito de linhagem genética, escovação e isolamento nas respostas comportamentais dos animais (TAMIOSO et al., 2017). Na ocasião, vinte ovelhas classificadas como reativas ou não reativas ao isolamento social temporário foram submetidas à escovação por um humano familiar. As ovelhas tinham 15 meses de idade, eram não gestantes e não amamentavam quando foram observadas.

O experimento foi conduzido em três sessões experimentais: na primeira tinha-se uma grade de metal separando o animal testado dos demais animais, sem distância entre eles. Na segunda havia duas grades de metal separando os animais a uma distância de 1,7 metros, ou seja, foi imposta a condição de isolamento social. E na terceira sessão, os animais voltaram a ser separados por apenas uma grade.

As sessões de testes ocorreram dois dias após a fase de adaptação dos animais ao equipamento e aos humanos e, em cada sessão, as ovelhas foram observadas em 3 momentos distintos: fase de pré escovação, com duração de 2 minutos e 30 segundos; fase de escovação, com duração de 3 minutos; e pós escovação, com duração de 2 minutos e 30 segundos.

Os dados coletados dizem respeito ao número de mudanças de postura dos animais e à proporção do tempo em que os animais permaneceram em determinadas posturas, tratando-se então, de um conjunto de dados com múltiplas respostas em que não há observações independentes, já que cada animal contribui com nove medidas. Portanto, há a necessidade de incorporar as correlações entre as medidas num mesmo animal e do animal dentro de cada sessão experimental, além da correlação entre as respostas. Para fins de análise, 8 respostas foram consideradas:

- `ncorpo`: Número de mudanças de postura de corpo.
- `ncabeca`: Número de mudanças de postura de cabeça.
- `norelha`: Número de mudanças de postura de orelha.
- `nolho`: Número de mudanças de postura de olho.
- `resporelha2`: Proporção do tempo com as orelhas levantadas ou assimétricas.
- `respcauda`: Proporção do tempo com a cauda em movimento.
- `respolho`: Proporção do tempo com os olhos fechados ou semi-cerrados.
- `resporlev`: Proporção do tempo com as orelhas levantadas.

Foram avaliados os efeitos de:

- `sessão`: Fator de 3 níveis que indica a sessão experimental (Se1, Se2, Se3).
- `momento`: Fator de 3 níveis em que indica o momento experimental (antes, durante, depois).
- `linhagem`: Fator de 2 níveis que classifica os animais como reativos ou não reativos ao isolamento social temporário.

A Tabela 1 apresenta a matriz de correlações duas a duas das variáveis respostas avaliadas no estudo. Nota-se que todas as respostas que configuram contagens apresentam correlações positivas entre si e menos correlacionadas com as respostas que configuram proporções.

Tabela 1 – Matriz de correlações de Pearson para as respostas em estudo.

	ncorpo	ncabeca	norelha	nolho	resporelha2	respcauda	respolho	resporlev
ncorpo	1,00	0,63	0,49	0,39	-0,13	-0,01	-0,10	-0,01
ncabeca	0,63	1,00	0,63	0,47	-0,24	-0,03	-0,09	-0,12
norelha	0,49	0,63	1,00	0,36	-0,22	0,07	-0,08	-0,09
nolho	0,39	0,47	0,36	1,00	-0,13	-0,07	-0,08	0,02
resporelha2	-0,13	-0,24	-0,22	-0,13	1,00	0,01	-0,32	0,56
respcauda	-0,01	-0,03	0,07	-0,07	0,01	1,00	0,27	-0,12
respolho	-0,10	-0,09	-0,08	-0,08	-0,32	0,27	1,00	-0,53
resporlev	-0,01	-0,12	-0,09	0,02	0,56	-0,12	-0,53	1,00

Na Figura 3 é apresentada a análise exploratória das variáveis respostas de contagem, tendo em (A) a distribuição das respostas e em (B) o comportamento das respostas em função das variáveis explicativas. É possível observar, em (A), que o número de mudanças de postura de corpo é a variável de contagem com maior frequência de valores iguais a 0, isto é, animais que não movimentaram o corpo durante o período sob observação, seguida pelo número de mudanças de postura de cabeça. O número de mudanças de postura de orelha é menos inflacionada em zero que as já mencionadas,

enquanto que para o número de mudanças de postura de olho nenhuma resposta igual a 0 foi observada. Nota-se em (B) que, marginalmente, há diferença entre os diferentes momentos experimentais para o número de mudanças de postura de cabeça, corpo e orelha. Para as mudanças de postura de olho a figura não aponta para efeito marginal das variáveis explicativas sob análise.

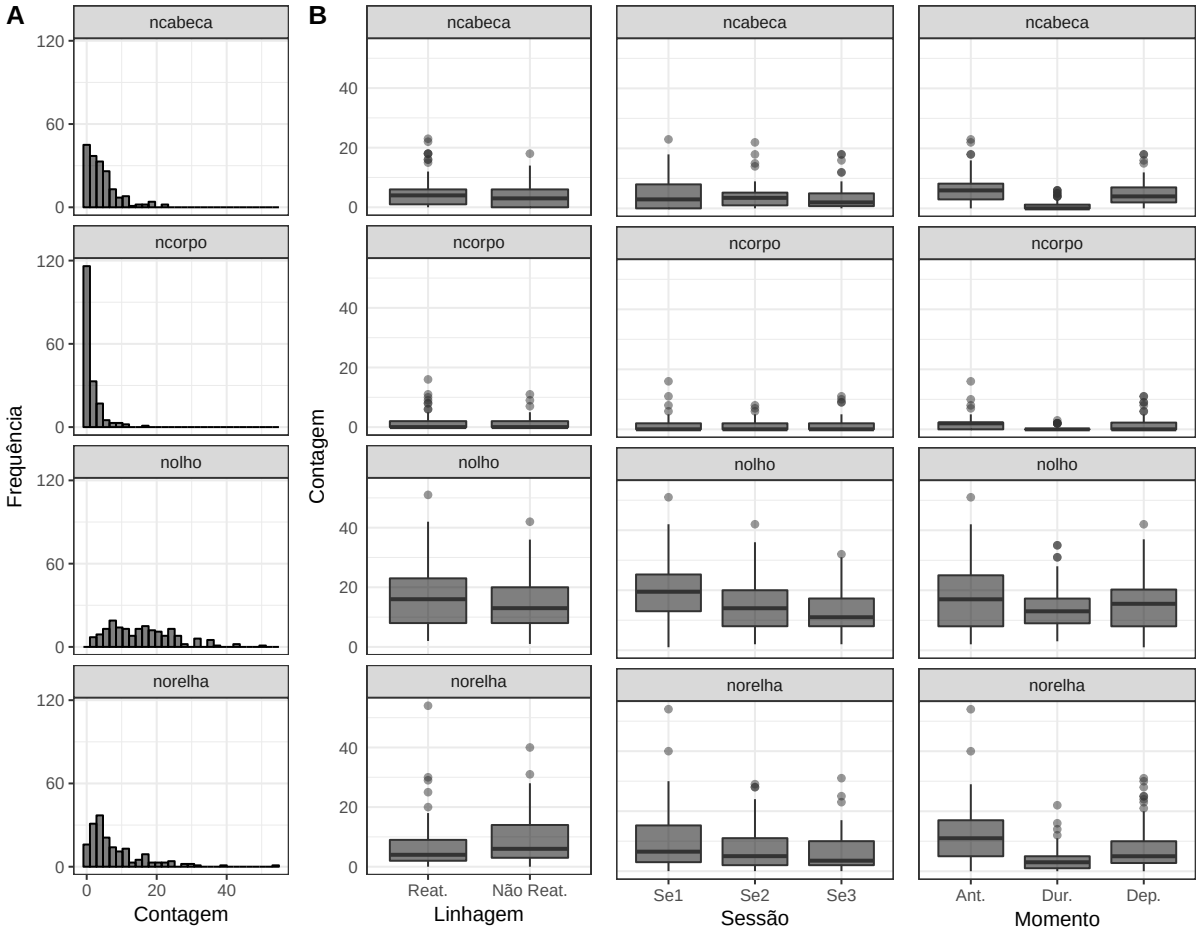


Figura 3 – Distribuição das variáveis respostas (A) e distribuição das respostas em função das covariáveis (B).

Na Figura 4 é apresentada a distribuição das variáveis resposta de proporção, tendo em (A) a distribuição das respostas e em (B) o comportamento das respostas em função das variáveis explicativas. Nota-se, em (A), que a proporção do tempo que os animais permaneceram com as orelhas levantadas ou assimétricas é uma variável inflacionada em 1, isto é, em boa parte das observações os animais permaneceram todo o tempo com as orelhas levantadas ou assimétricas. As variáveis que representam a proporção do tempo com a cauda em movimento e a proporção do tempo com os olhos fechados ou semi-cerrados são variáveis com frequência elevada de zeros, enquanto que o tempo com as orelhas levantadas apresenta frequências elevadas em 0 e em 1 e frequências menores entre estes dois valores. O gráfico aponta ainda para o efeito marginal (B) de momento para as duas variáveis respostas referentes às posturas de

orelha. Para a proporção do tempo com a cauda em movimento nota-se que há um efeito marginal de linhagem e para a proporção em que o animal permaneceu com os olhos fechados ou semi-cerrados é possível verificar diferenças entre os níveis de todas as covariáveis.

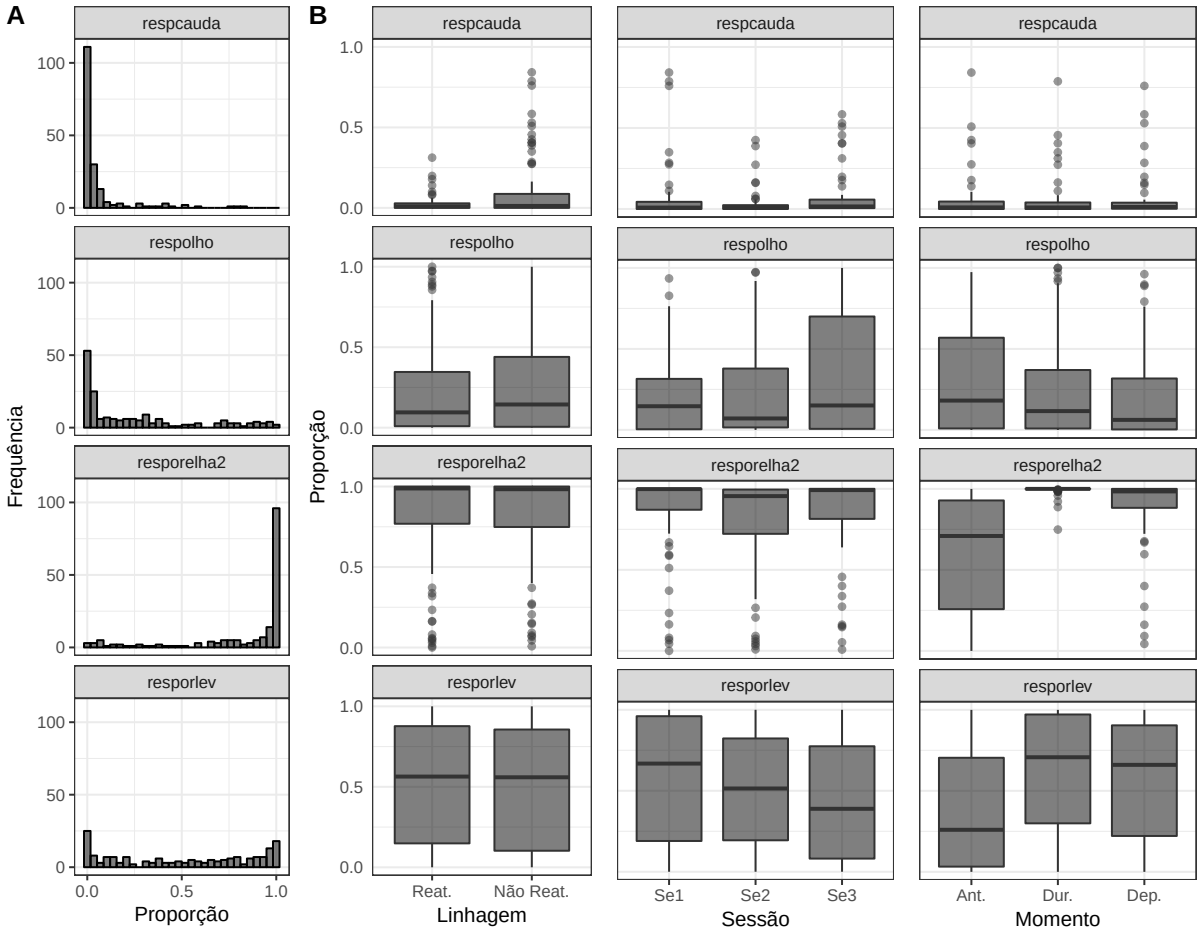


Figura 4 – Comportamento das variáveis respostas (A) e comportamento das respostas em função das covariáveis (B).

3 Metodologia

Este capítulo é destinado à revisão bibliográfica e apresentação do método proposto. Na seção 3.1 tem-se como revisão de literatura a construção e características dos Modelos Multivariados de Covariância Linear Generalizada, na subseção 3.1.1 é apresentada uma visão geral do processo de estimação e da distribuição assintótica dos estimadores e na seção 3.2 é apresentado o teste de Wald proposto para lidar com múltiplas respostas e de diferentes naturezas.

3.1 Modelos Multivariados de Covariância Linear Generalizada

Os MLGs são uma forma de modelagem univariada para dados de diferentes naturezas, tais como dados contínuos simétricos e assimétricos, contagens, dentre outras. Tais características tornam essa classe de modelos uma flexível ferramenta de modelagem aplicável a diversos tipos de problemas. Contudo, por mais flexível e discutida na literatura, essa classe apresenta duas principais restrições: 1) a incapacidade de lidar com obsrvações dependentes e 2) múltiplas respostas simultaneamente.

Com o objetivo de solucionar esses problemas, foi proposta por Bonat e Jørgensen (2016), uma estrutura geral para análise de dados não gaussianos com múltiplas respostas em que não se faz suposições quanto à independência das observações, denominada Modelos Multivariados de Covariância Linear Generalizada (MCGLM). Isto posto, é apresentado o MCGLM como extensão do MLG.

Seja $\boldsymbol{Y}$ um vetor $N \times 1$ de valores observados da variável resposta, $\boldsymbol{X}$ uma matriz de delineamento $N \times k$ e $\boldsymbol{\beta}$ um vetor de parâmetros de regressão $k \times 1$, um MLG pode ser descrito da forma

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}) &= \boldsymbol{\mu} = g^{-1}(\boldsymbol{X}\boldsymbol{\beta}), \\ \mathrm{Var}(\boldsymbol{Y}) &= \Sigma = \mathrm{V}\left(\boldsymbol{\mu}; p\right)^{1/2} \left(\tau_0 \boldsymbol{I}\right) \mathrm{V}\left(\boldsymbol{\mu}; p\right)^{1/2}, \end{aligned} \tag{3.1}$$

em que $g(.)$ é a função de ligação, $\mathrm{V}\left(\boldsymbol{\mu}; p\right)$ é uma matriz diagonal em que as entradas principais são dadas pela função de variância aplicada ao vetor $\boldsymbol{\mu}$, p é o parâmetro de potência, τ_0 o parâmetro de dispersão e $\boldsymbol{I}$ é a matriz identidade de ordem $N \times N$.

Os MCGLMs fazem uso de apenas duas funções, a função de variância e de ligação. Diferentes escolhas de funções de variância implicam em diferentes suposições a respeito da distribuição da variável resposta. Dentre as funções de variância conhecidas, podemos citar

(i) a função de variância potência, que caracteriza a família Tweedie de distribuições, em que a função de variância é dada por $\vartheta(\boldsymbol{\mu}; p) = \mu^p$, na qual destacam-se a distribuições: normal (p = 0), Poisson (p = 1), gama (p = 2) e inversa gaussiana (p = 3) (JØRGENSEN, 1987; JØRGENSEN, 1997).

(ii) a função de dispersão Poisson–Tweedie, a qual caracteriza a família Poisson-Tweedie de distribuições, que visa contornar a inflexibilidade da utilização da função de variância potência para respostas discretas. A família Poisson-Tweedie tem função de dispersão dada por $\vartheta(\boldsymbol{\mu}; p) = \mu + \mu^p$ e tem como casos particulares os mais famosos modelos para dados de contagem: Hermite (p = 0), Neyman tipo A (p = 1), binomial negativa (p = 2) e Poisson–inversa gaussiana (p = 3) (JØRGENSEN; KOKONENDJI, 2015).

(iii) a função de variância binomial, dada por $\vartheta(\boldsymbol{\mu}) = \mu(1-\mu)$, utilizada quando a variável resposta é binária, restrita a um intervalo ou quando tem-se o número de sucessos em um número de tentativas.

A proposta inicial do MCGLM é uma alternativa para problemas em que a suposição de independência entre as observações não é atendida. Neste caso, a solução proposta é substituir a matriz identidade $\boldsymbol{I}$ da equação que descreve a matriz de variância e covariância por uma matriz não diagonal $\boldsymbol{\Omega}(\boldsymbol{\tau})$ que descreva adequadamente a estrutura de correlação entre as observações. Trata-se de uma ideia similar à proposta de Liang e Zeger (1986) nos modelos GEE (*Generalized Estimation Equation*) quando utiliza-se uma matriz de correlação de trabalho para considerar a dependência entre as observações. A matriz $\boldsymbol{\Omega}(\boldsymbol{\tau})$ é descrita como uma combinação de matrizes conhecidas tal como nas propostas de Anderson et al. (1973) e Pourahmadi (2000), podendo ser escrita da forma

$$h\{\boldsymbol{\Omega}(\boldsymbol{\tau})\} = \tau_0 Z_0 + \ldots + \tau_D Z_D, \tag{3.2}$$

em que $h(.)$ é a função de ligação de covariância, Z_d com d = 0,..., D são matrizes que representam a estrutura de covariância presente nos dados e $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_D)$ é um vetor $(D+1) \times 1$ de parâmetros de dispersão. Tal estrutura pode ser vista como um análogo ao preditor linear para a média e foi nomeado como preditor linear matricial. A especificação da função de ligação de covariância é discutida por Pinheiro e Bates (1996) e é possível selecionar combinações de matrizes para se obter os mais conhecidos na literatura modelos para dados longitudinais, séries temporais, dados espaciais e espaço-temporais, maiores detalhes são discutidos por Demidenko (2013).

Com isso, substituindo a matriz identidade pela equação do preditor linear matricial tem-se o chamado *covariance generalized linear model*, uma classe com toda a flexibilidade dos MLGs, porém contornando a restrição da independência entre as observações desde que o preditor linear matricial seja adequadamente especificado.

Além de permitir a modelagem de dados com estrutura de covariância, os MCGLMs permitem modelar múltiplas respostas. Seja $\boldsymbol{Y}_{N\times R} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_R\}$ uma matriz de variáveis respostas, $\boldsymbol{M}_{N\times R} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_R\}$ uma matriz de valores esperados. Cada uma das variáveis respostas tem sua própria matriz de variância e covariância, responsável por modelar a covariância dentro de cada resposta, sendo expressa por

$$\Sigma_r = \mathrm{V}_r\left(\boldsymbol{\mu}_r; p\right)^{1/2} \boldsymbol{\Omega}_r\left(\boldsymbol{\tau}\right) \mathrm{V}_r\left(\boldsymbol{\mu}_r; p\right)^{1/2}. \tag{3.3}$$

Além disso, é necessária uma matriz de correlação Σ_b, de ordem $R \times R$, que descreve a correlação entre as variáveis respostas. Para a especificação da matriz de variância e covariância conjunta é utilizado o produto Kronecker generalizado, proposto por Martinez-Beneito (2013).

Finalmente, um MCGLM é descrito como

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}) &= \boldsymbol{M} = \{g_1^{-1}(\boldsymbol{X}_1\boldsymbol{\beta}_1), \ldots, g_R^{-1}(\boldsymbol{X}_R\boldsymbol{\beta}_R)\} \\ \mathrm{Var}(\boldsymbol{Y}) &= \boldsymbol{C} = \boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b, \end{aligned} \tag{3.4}$$

em que $\boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b = \mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \ldots, \tilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b \otimes \boldsymbol{I})\mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^\top, \ldots, \tilde{\boldsymbol{\Sigma}}_R^\top)$ é o produto generalizado de Kronecker, a matriz $\tilde{\boldsymbol{\Sigma}}_r$ denota a matriz triangular inferior da decomposição de Cholesky da matriz $\boldsymbol{\Sigma}_r$, o operador Bdiag denota a matriz bloco-diagonal e $\boldsymbol{I}$ uma matriz identidade $N \times N$.

Toda metodologia do MCGLM está implementada no pacote `mcglm` (BONAT, 2018) do *software* estatístico R (R Core Team, 2018).

3.1.1 Estimação e Inferência

Os MCGLMs são ajustados baseados no método de funções de estimação descritos em detalhes por Bonat e Jørgensen (2016) e Jørgensen e Knudsen (2004). Nesta seção é apresentada uma visão geral do algoritmo e da distribuição assintótica dos estimadores baseados em funções de estimação.

As suposições de segundo momento dos MCGLMs permitem a divisão dos parâmetros em dois conjuntos: $\boldsymbol{\theta} = (\beta^\top, \boldsymbol{\lambda}^\top)^\top$. Desta forma, $\beta = (\beta_1^\top, \ldots, \beta_R^\top)^\top$ é um vetor $K \times 1$ de parâmetros de regressão e $\boldsymbol{\lambda} = (\rho_1, \ldots, \rho_{R(R-1)/2}, p_1, \ldots, p_R, \boldsymbol{\tau}_1^\top, \ldots, \boldsymbol{\tau}_R^\top)^\top$ é um vetor $Q \times 1$ de parâmetros de dispersão. Além disso, $\mathcal{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_R^\top)^\top$ denota o vetor empilhado de ordem $NR \times 1$ da matriz de variáveis respostas $\boldsymbol{Y}_{N\times R}$ e $\mathcal{M} = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_R^\top)^\top$ denota o vetor empilhado de ordem $NR \times 1$ da matriz de valores esperados $\boldsymbol{M}_{N\times R}$.

Para estimação dos parâmetros de regressão é utilizada a função quasi-score (LIANG; ZEGER, 1986), representada por

$$\psi_\beta(\beta, \boldsymbol{\lambda}) = \boldsymbol{D}^\top \boldsymbol{C}^{-1}(\mathcal{Y} - \mathcal{M}), \tag{3.5}$$

em que $\boldsymbol{D} = \nabla_\beta \mathcal{M}$ é uma matriz $NR \times K$, e ∇_β denota o operador gradiente. Utilizando a função quasi-score a matriz $K \times K$ de sensitividade de ψ_β é dada por

$$S_\beta = E(\nabla_{\beta\psi\beta}) = -\boldsymbol{D}^\top \boldsymbol{C}^{-1} \boldsymbol{D}, \tag{3.6}$$

enquanto que a matriz $K \times K$ de variabilidade de ψ_β é escrita como

$$V_\beta = VAR(\psi\beta) = \boldsymbol{D}^\top \boldsymbol{C}^{-1} \boldsymbol{D}. \tag{3.7}$$

Para os parâmetros de dispersão é utilizada a função de estimação de Pearson, definida da forma

$$\psi_{\lambda_i}(\beta, \boldsymbol{\lambda}) = \text{tr}(W_{\lambda i}(\boldsymbol{r}^\top \boldsymbol{r} - \boldsymbol{C})), i = 1, .., Q, \tag{3.8}$$

em que $W_{\lambda i} = -\frac{\partial C^{-1}}{\partial \lambda_i}$ e $\boldsymbol{r} = (\mathcal{Y} - \mathcal{M})$. A entrada (i, j) da matriz de sensitividade $Q \times Q$ de $\psi_{\boldsymbol{\lambda}}$ é dada por

$$S_{\lambda_{ij}} = E\left(\frac{\partial}{\partial \boldsymbol{\lambda_i}} \psi \boldsymbol{\lambda}_j\right) = -tr(W_{\lambda_i} C W_{\lambda_J} C). \tag{3.9}$$

Já a entrada (i, j) da matriz de variabilidade $Q \times Q$ de $\psi_{\boldsymbol{\lambda}}$ é definida por

$$V_{\lambda_{ij}} = Cov\left(\psi_{\lambda_i}, \psi_{\lambda_j}\right) = 2tr(W_{\lambda_i} C W_{\lambda_J} C) + \sum_{l=1}^{NR} k_l^{(4)} (W_{\lambda_i})_{ll} (W_{\lambda_j})_{ll}, \tag{3.10}$$

em que $k_l^{(4)}$ denota a quarta cumulante de $\mathcal{Y}_l$. No processo de estimação dos MCGLMs são usadas as versões empíricas.

Para se levar em conta a covariância entre os vetores β e $\boldsymbol{\lambda}$, Bonat e Jørgensen (2016) obtiveram as matrizes de sensitividade e variabilidade cruzadas, denotadas por $S_{\boldsymbol{\lambda}\beta}$, $S_{\beta\boldsymbol{\lambda}}$ e $V_{\boldsymbol{\lambda}\beta}$, mais detalhes em Bonat e Jørgensen (2016). As matrizes de sensitividade e variabilidade conjuntas de ψ_β e $\psi_{\boldsymbol{\lambda}}$ são denotados por

$$S_{\boldsymbol{\theta}} = \begin{bmatrix} S_\beta & S_{\beta\boldsymbol{\lambda}} \\ S_{\boldsymbol{\lambda}\beta} & S_{\boldsymbol{\lambda}} \end{bmatrix} \text{ e } V_{\boldsymbol{\theta}} = \begin{bmatrix} V_\beta & V_{\boldsymbol{\lambda}\beta}^\top \\ V_{\boldsymbol{\lambda}\beta} & V_{\boldsymbol{\lambda}} \end{bmatrix}. \tag{3.11}$$

Seja $\hat{\boldsymbol{\theta}} = (\hat{\beta}^\top, \hat{\boldsymbol{\lambda}}^\top)^\top$ o estimador baseado em funções de estimação de $\boldsymbol{\theta}$. Então, a distribuição assintótica de $\hat{\boldsymbol{\theta}}$ é

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, J_{\boldsymbol{\theta}}^{-1}), \tag{3.12}$$

em que $J_{\boldsymbol{\theta}}^{-1}$ é a inversa da matriz de informação de Godambe, dada por $J_{\boldsymbol{\theta}}^{-1} = S_{\boldsymbol{\theta}}^{-1} V_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}^{-\top}$, em que $S_{\boldsymbol{\theta}}^{-\top} = (S_{\boldsymbol{\theta}}^{-1})^\top$.

Para resolver o sistema de equações $\psi_\beta = 0$ e $\psi_{\boldsymbol{\lambda}} = 0$ faz-se uso do algoritmo Chaser modificado, proposto por Jørgensen e Knudsen (2004), que fica definido como

$$\begin{aligned} \beta^{(i+1)} &= \beta^{(i)} - S_\beta^{-1} \psi\beta(\beta^{(i)}, \boldsymbol{\lambda}^{(i)}), \\ \boldsymbol{\lambda}^{(i+1)} &= \boldsymbol{\lambda}^{(i)} \alpha S_{\boldsymbol{\lambda}}^{-1} \psi\boldsymbol{\lambda}(\beta^{(i+1)}, \boldsymbol{\lambda}^{(i)}). \end{aligned} \tag{3.13}$$

3.2 Teste Wald para Análise de Variância Multivariada para Dados Não Gaussianos

A MANOVA clássica (SMITH; GNANADESIKAN; HUGHES, 1962) é um assunto com vasta discussão na literatura e possui diversas propostas com o objetivo de verificar a nulidade conjunta dos parâmetros de um modelo de regressão multivariado, como os já citados, lambda de Wilk's (WILKS, 1932), traço de Hotelling-Lawley (LAWLEY, 1938; HOTELLING, 1951), traço de Pillai (PILLAI et al., 1955) e maior raiz de Roy (ROY, 1953).

Nesta seção, será apresentado uma extensão do clássico teste de Wald para lidar com múltiplas respostas não gaussianas e de diferentes naturezas. Sendo assim, seguindo a Equação 3.4 que descreve a média e a matriz de variância e covariância conjunta de um MCGLM, pode-se enunciar as hipóteses de interesse como

$$\begin{aligned} H_0 &= \boldsymbol{L}\boldsymbol{\beta} = \mathbf{0} \\ H_1 &= \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{c}, \end{aligned} \tag{3.14}$$

em que, na hipótese nula, $\boldsymbol{L} = \boldsymbol{G} \otimes \boldsymbol{F}$. A matriz $\boldsymbol{G}$ tem dimensão $R \times R$ e especifica as hipóteses referentes às respostas, enquanto que a matriz $\boldsymbol{F}$, a qual especifica as hipóteses entre tratamentos, tem dimensão $s \times p$, onde s é o número de restrições lineares e p é o número de coeficientes de regressão para cada variável resposta. Sendo assim, a matriz $\boldsymbol{L}$ tem dimensão $(sR \times p)$. E na hipótese alternativa, $\boldsymbol{c}$ é um vetor qualquer não nulo.

Para avaliar a hipótese linear especificada na Equação 3.14, fez-se uso da estatística de Wald representada por

$$W_s = (\boldsymbol{L}\boldsymbol{\beta})^T \left(\boldsymbol{L} J_\beta^{-1} \boldsymbol{L}^T\right)^{-1} (\boldsymbol{L}\boldsymbol{\beta}), \tag{3.15}$$

em que J_β^{-1} é a parte da inversa da matriz de informação de Godambe que considera apenas os parâmetros de regressão. Sob a hipótese nula, a estatística segue assintoticamente distribuição Qui-quadrado com sR graus de liberdade. É possível notar que essa construção permite realizar o teste de hipóteses para todas as variáveis respostas, bem como entre combinações de variáveis e todos os possíveis contrastes entre os níveis de tratamento.

Para exemplificar a flexibilidade do método, considere um caso com duas variáveis respostas e uma variável explicativa X categórica com três níveis (A, B e C). Neste caso, o preditor linear é dado por

$$\mu_r = \beta_{r0} + \beta_{r1}\,[X = B] + \beta_{r2}\,[X = C],$$

em que β_{r0} é o intercepto e β_{r1} e β_{r2} representam a diferença do tratamento A para B e A para C, respectivamente. Nesta parametrização, em que A é a categoria de referência,

a matriz de delineamento tem a forma

$$D = \begin{pmatrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & 1 & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & 1 \end{pmatrix}.$$

A hipótese nula do teste pode ser escrita como

$$H_0 = \beta_{r1} = \beta_{r2} = 0, \text{ para } r = 1, 2,$$

o que equivale, em notação matricial a

$$H_0 : \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

de forma que

$$\boldsymbol{G} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ e } \boldsymbol{F} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Com isso, calcula-se as estatísticas de Wald seguindo a Equação 3.15 e realiza-se a interpretação do teste. É importante ressaltar que, embora a matriz de variância e covariância $\boldsymbol{\Sigma}$ não apareça explicitamente nas estatísticas de Wald, ela tem um impacto crucial, pois a matriz de informação de Godambe ($J_{\boldsymbol{\theta}}$) depende dela. Consequentemente, o teste apresentado tem a MANOVA clássica como um caso particular.

4 Estudo de Simulação

Com o objetivo de verificar o poder do teste de hipóteses proposto para testar a nulidade simultânea dos parâmetros, foram simulados conjuntos de dados com 3 variáveis respostas seguindo uma mesma distribuição de probabilidade: Poisson, Binomial (n = 10) ou Beta. Considerou-se ainda uma variável explicativa categórica de 5 níveis. Além disso, a correlação entre as variáveis respostas foi fixada em

$$\boldsymbol{\Sigma}_b = \begin{bmatrix} 1 & 0,75 & 0,5 \\ 0,75 & 1 & 0,25 \\ 0,5 & 0,25 & 1 \end{bmatrix}. \tag{4.1}$$

No ajuste dos modelos, os interceptos foram fixados em 10 para as respostas seguindo distribuição de Poisson e em 0,5 para as respostas seguindo as distribuições Binomial com n=10 e Beta. Os demais parâmetros de regressão foram fixados em um mesmo valor entre -0,5 até 0,5, variando de 0,1 em 0,1 totalizando 11 cenários para os valores dos parâmetros. Para cada cenário foram gerados 1000 conjuntos de dados com diferentes tamanhos amostrais: 50, 100, 250, 500 e 1000. As amostras foram geradas utilizando o método NORTA (CARIO; NELSON, 1997) para geração de respostas multivariadas correlacionadas. Considerando todas as possibilidades, foram simulados, para cada uma das 3 distribuições de variável resposta, 55 cenários, contendo cada um 1000 conjuntos de dados com diferentes valores para os coeficientes de regressão e diferentes tamanhos amostrais.

Para cada amostra gerada foi ajustado um MCGLM em que, para os conjuntos com variáveis respostas seguindo distribuição Poisson a função de ligação utilizada foi a logarítmica e para função de variância utilizou-se a Tweedie. Já para as respostas seguindo distribuição Binomial e Beta foi utilizada a função de ligação logito com função de variância Binomial. Em todos os casos o preditor matricial para a matriz de variância e covariância foi especificado de forma a explicitar que as observações são independentes dentro de cada resposta.

A especificação dos modelos é dada por

$$g_r(\mu_{ijr}) = \beta_{0r} + \beta_{1r}X_{ij}, \tag{4.2}$$

em que o índice r varia de 1 até 3 e refere-se às variáveis respostas das amostras simuladas, j varia de 1 até 5 e diz respeito aos diferentes níveis da variável categórica simulada X e i diz respeito ao número de observações. O preditor matricial foi especificado como $h\{\boldsymbol{\Omega}(\boldsymbol{\tau})\} = \tau_0 Z_0$, em que Z_0 é uma matriz Identidade de ordem dada pelo tamanho amostral do conjunto de dados simulado.

Ajustou-se então, o modelo para as amostras geradas e efetuou-se a MANOVA. O comportamento esperado do teste, considerando um nível de significância de 5% era de que para cenários com valores dos efeitos próximos de 0, o número de não rejeições fosse maior. Além disso espera-se que, quando os valores dos efeitos sejam iguais a 0, o teste não rejeita em 95% dos casos, evidenciando assim, o poder do teste.

Na Tabela 2 são mostrados os valores esperados sob os diferentes valores de parâmetros de regressão fixados. Os valores esperados mostram que para cenários simulados considerando distribuição de Poisson para a resposta variam de 6,07 até 16,49 enquanto que para as respostas simuladas com as distribuições Binomial e Beta esses valores variam de 0,38 até 0,62 mostrando que o estudo foi delineado de forma que as diferenças nos valores esperados sob diferentes valores dos parâmetros de regressão não fossem pequenas demais. Vale lembrar que em cada cenário todos os parâmetros de todas as respostas foram fixados no mesmo valor.

Tabela 2 – Valores esperados na escala da resposta para os parâmetros de regressão fixados.

Parâmetro	Valor esperado	
	Poisson	Binomial/Beta
Intercepto (fixado)	10	0,5
-0,5	6,07	0,38
-0,4	6,70	0,40
-0,3	7,41	0,43
-0,2	8,19	0,45
-0,1	9,05	0,48
0	10	0,50
0,1	11,05	0,52
0,2	12,21	0,55
0,3	13,50	0,57
0,4	14,92	0,60
0,5	16,49	0,62

A Figura 5 mostra o resultado do estudo de simulação. Para cada cenário são apresentadas as proporções de rejeição e não rejeição juntamente com uma linha pontilhada em vermelho no nível de 5% de significância. O único cenário não apresentado é para o caso de resposta com distribuição beta com $n = 1000$, neste caso obteve-se apenas resultados parciais pois não conseguiu-se gerar todas as amostras com a metodologia NORTA.

Nota-se, pela figura, um comportamento similar para as diferentes distribuições das variáveis respostas. Conforme o tamanho amostral aumenta, maior é a proporção de rejeições da hipótese nula para valores de beta diferentes de 0. Este comportamento é mais evidente nas distribuições Poisson e Beta, em que desde os tamanhos amostrais pequenos observam-se altas proporções de rejeições. Além disso, o teste proposto teve

o pior desempenho observado para as amostras com distribuição Binomial, em que no menor tamanho amostral observou-se proporção considerável de não rejeições nos casos com beta iguais a $\pm$ 0,5. Porém, em todos os casos, com n maior ou igual a 250, constatou-se que, sob hipótese nula, o nível de confiança atingiu os esperados 95% de confiança.

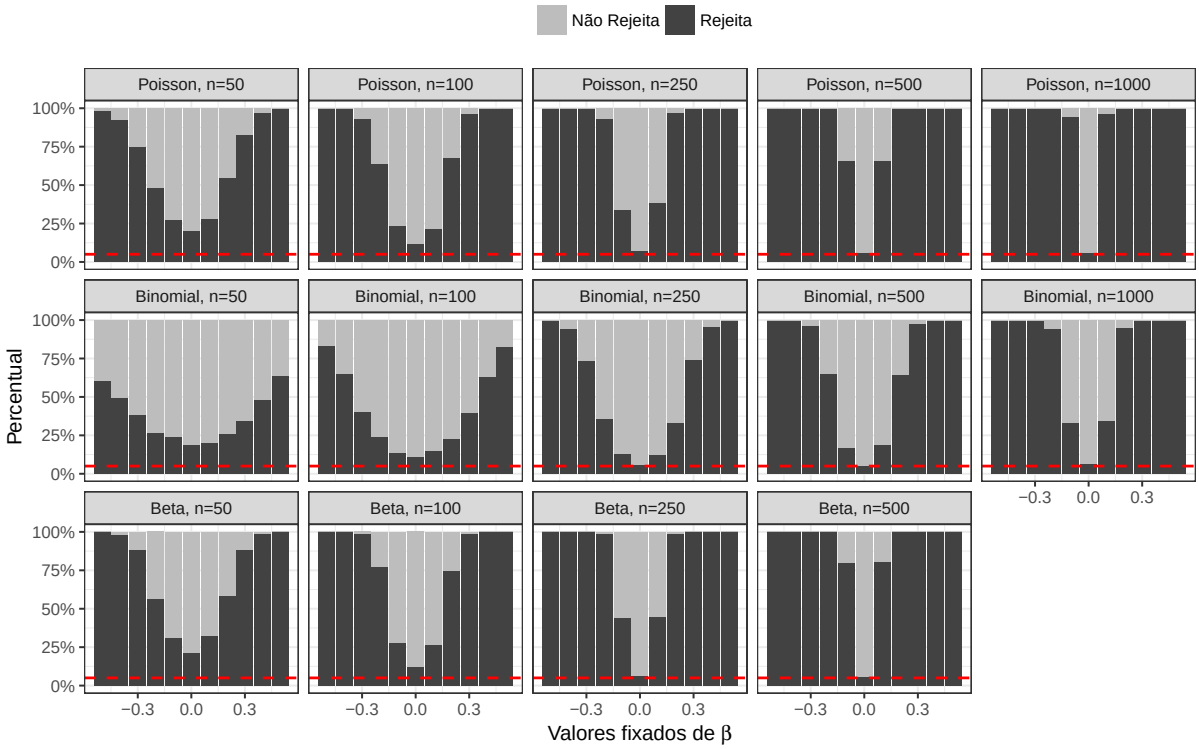


Figura 5 – Percentual de rejeição e não rejeição da hipótese de nulidade conjunta dos parâmetros dos modelos para as amostras simuladas.

5 Resultados e Discussão

Nesse capítulo são apresentados os resultados e discussões da aplicação da MANOVA aos dados apresentados no Capítulo 2. Para ambos os conjuntos de dados, são expostos neste capítulo, a especificação do modelo ajustado, a análise de resíduo e principais discussões.

5.1 Análise dos Processos Movidos contra Grandes Litigantes

Para ajuste do modelo foram consideradas as respostas `procedente` e `improcedente`, ambas do tipo binário. A resposta `acordo` não foi considerada no ajuste do modelo, pois ela é complementar às outras duas respostas, isto é, caso o resultado do processo não seja procedente nem improcedente, obrigatóriamente será acordo. Além disso, considerou-se um modelo com efeitos fixos dispostos de forma aditiva, além de considerar independência entre as observações. Portanto, os preditores lineares são descritos como

$$\begin{aligned} g_r(\mu_{ir}) =& \beta_{0r} + \beta 1_{ar}\text{foro}_i + \beta 2_{br}\text{tipo\_vara}_i + \beta 3_r\text{adv\_reu}_i + \\ & \beta 4_{cr}\text{empresa}_i + \beta 5_{dr}\text{tipo\_dano}_i + \beta 6_r\log_{10}(\text{valor\_acao}_i) \\ & \beta 7_r\text{serasa}_i + \beta 8_r\text{terceiro}_i + \beta 9_r\text{consumo}_i + \beta 10_r\text{gratuidade}_i, \end{aligned} \tag{5.1}$$

em que o índice r refere-se às variáveis respostas `procedente` e `improcedente`, $g_r(.)$ é a função de ligação logito, i varia de 1 a 8578 e diz respeito as observações, $\beta 1_a$, $\beta 2_b$, $\beta 3$, $\beta 4_c$, $\beta 5_d$, $\beta 6$, $\beta 7$, $\beta 8$, $\beta 9$ e $\beta 10$ representam os efeitos de foro, tipo de vara, réu com advogado, empresa, tipo de dano, $\log_{10}$(valor da acao), serasa, terceiro, consumo e gratuidade respectivamente, a indexa os níveis do foro: Central (referência), Ipiranga, Itaquera, Jabaquara, Lapa, Penha de França, Pinheiros, Santana , Santo Amaro, São Miguel Paulista, Tatuapé e Vila Prudente, b os níveis do tipo de vara: justiça comum com advogado (referência), JEC com advogado e JEC sem advogado, réu com advogado possui os níveis: não (referência) e sim, c indexa os níveis da empresa: Banco do Brasil (referência), Bradesco, Claro, Eletropaulo, Itaú, NET, Nextel, Santander, TIM e Vivo, d os níveis do tipo de dano: dano material (referência), moral e dano material e moral, a presença de discussões a respeito dos órgãos de proteção ao crédito possui os níveis: não (referência) e sim, presença de discussões sobre terceiros: não (referência) e sim, discussões sobre consumo: não (referência) e sim, discussões sobre gratuidade judiciária: não (referência) e sim. Sendo assim, para cada tipo de resposta estimou-se 31 parâmetros.

O preditor matricial dado por $h\{\boldsymbol{\Omega}(\boldsymbol{\tau})\} = \tau_0 Z_0$, em que a função $h(.)$ utilizada foi a identidade, τ_0 são os parâmetros de dispersão e Z_0 uma matriz identidade de

ordem 8578 × 8578, é comum a todas as respostas e foi especificado de forma a explicitar que as observações são independentes.

A Tabela 3 mostra as estimativas do parâmetro de dispersão. Para ambos os tipos de resposta os valores estão próximos de 1, indicando que a distribuição binomial é adequada para o ajuste.

Tabela 3 – Estimativas dos parâmetros de dispersão e erro-padrão (ep).

	τ_0
Resposta	Est. (ep)
improcedente	1.059 (0.024)
procedente	1.146 (0.034)

Com o objetivo de verificar a qualidade de ajuste do modelo, realizou-se a análise de resíduos de Pearson apresentada na Figura 6. O comportamento dos resíduos mostra uma distribuição com dois picos em torno do zero para a resposta improcedente e uma distribuição aproximadamente simétrica em torno do zero para a resposta procedente, porém ambas possuem variação entre -3 e 3 com poucos valores atípicos.

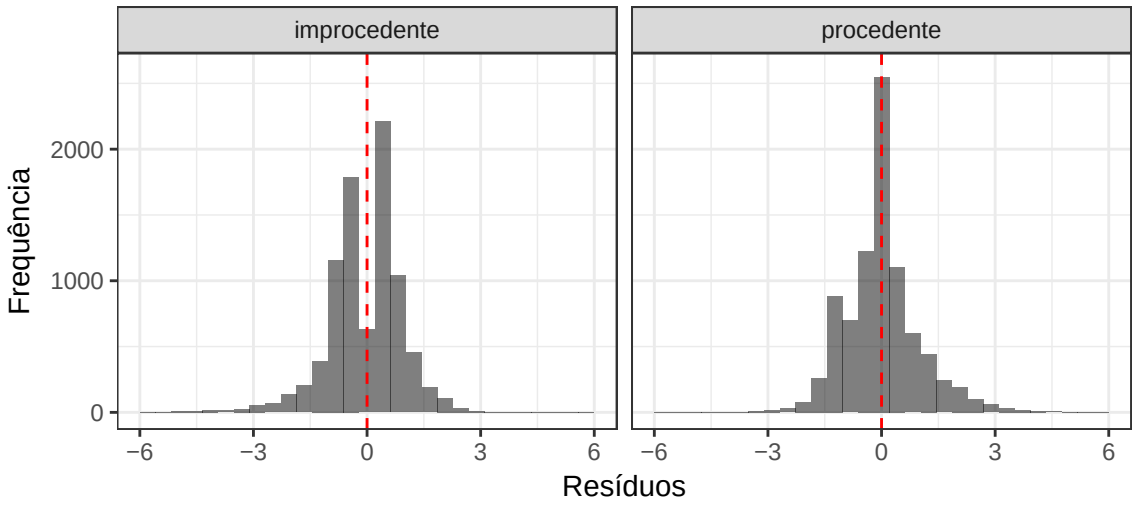


Figura 6 – Histogramas dos resíduos de Pearson por resposta.

Devido ao comportamento atípico dos resíduos observados na Figura 6, investigou-se o comportamento dos erros das reamostragens via bootstrap. Para tal, realizou-se 200 reamostragens de tamanho 8578 com reposição e para cada uma das reamostras ajustou-se o modelo apresentado pela Equação 5.1. Na Figura 7 são apresentadas as razões entre os erros bootstrap com os errões padrões originais. Nota-se que as proporções estão próximas de 1, apontando para pouca diferença entre os erros estimados pelo modelo e pelos erros estimados via bootstrap, indicando bom ajuste do modelo.

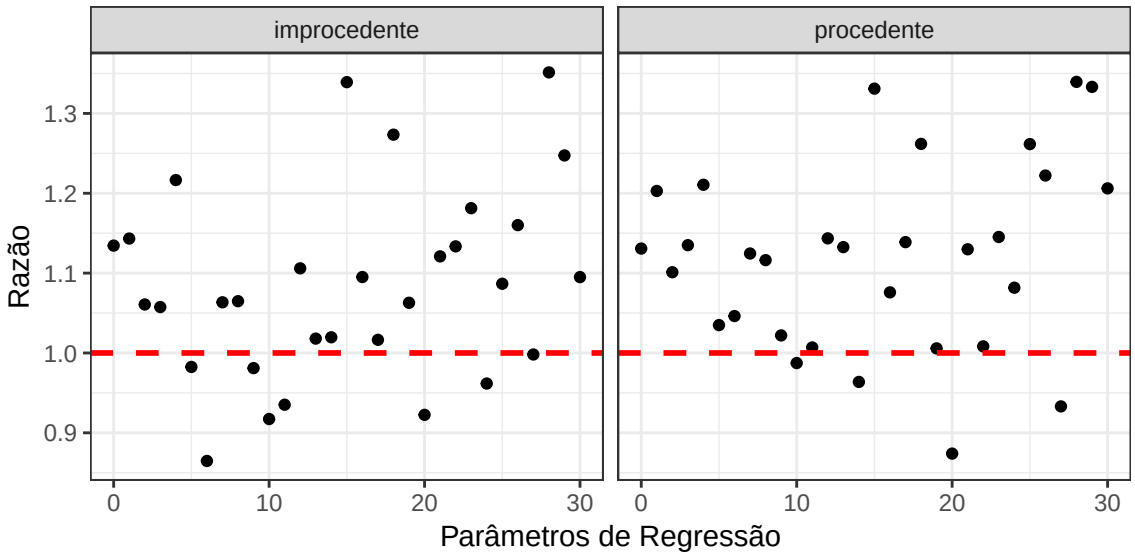


Figura 7 – Razão entre os erros bootstrap e originais do modelo para cada resposta.

Por fim, foi realizada a análise de variância multivariada com o objetivo de testar a nulidade simultânea de todos os parâmetros estimados no modelo. O resultado da MANOVA é sumarizado na Tabela 4. O resultado do teste de hipótese aponta para a não nulidade simultânea dos parâmetros.

Tabela 4 – Resumo do teste de hipóteses para o modelo ajustado

Efeitos	GL	W_s	Qui-quadrado	p-valor
Intercepto	2	0.018	155.965	< 0.01
foro	22	0.066	569.870	< 0.01
tipo_vara	4	0.009	76.136	< 0.01
adv_reu	2	0.002	21.220	< 0.01
empresa	18	0.067	570.791	< 0.01
tipo_dano	4	0.035	300.280	< 0.01
$\log_{10}$(valor_acao)	2	0.004	31.987	< 0.01
serasa	2	0.030	256.354	< 0.01
terceiro	2	0.042	360.509	< 0.01
consumo	2	0.097	835.402	< 0.01
gratuidade	2	0.032	272.629	< 0.01

Adicionalmente, para investigar os efeitos dos níveis de cada covariável na MANOVA, tem-se as estimativas pontuais e intervalares com 95% de confiança na Figura 8, em que cada cor representa uma covariável e cada símbolo uma resposta. O modelo, conforme descrito na Equação 5.1, é parametrizado em termos de efeitos para as variáveis categóricas, ou seja, os parâmetros representam a diferença com relação ao nível de referência na escala do preditor linear. Em termos práticos, espera-se que, a probabilidade de procedência aumente enquanto a de improcedência diminua (e vice-versa), ou seja, para covariáveis relevantes, espera-se que o sinal das estimativas

seja diferente para cada uma das respostas. Isso occorre para o valor da ação e para o efeito do nível Ipiranga da covariável foro, por exemplo. Intervalos de confiança que cobrem o valor zero, indicam que não há evidência de diferença em relação ao nível de referência, por exemplo, não há evidências de que a probabilidade de procedência e improcedência para a empresa Bradesco seja diferente da empresa Banco do Brasil.

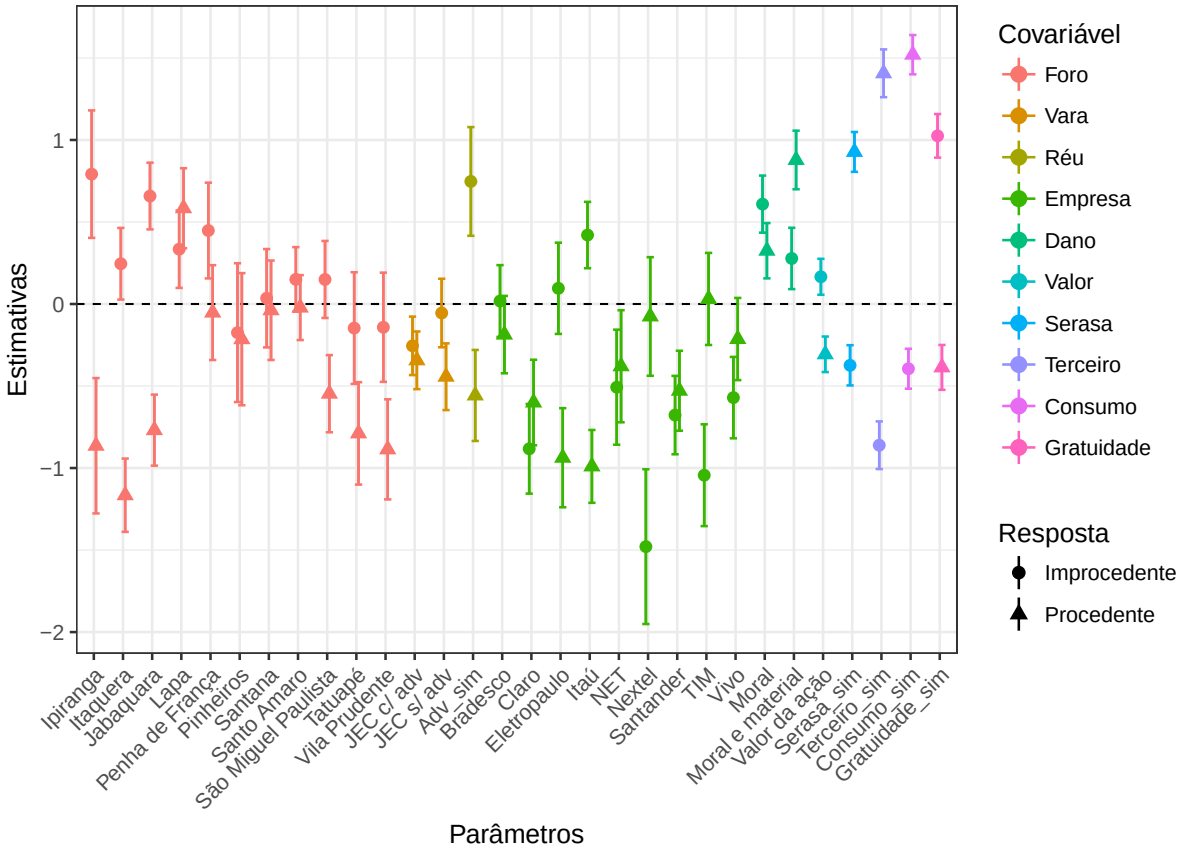


Figura 8 – Estimativas pontuais e intervalares para os níveies das covariáveis em cada resposta.

5.2 Análise Comportamental de Ovelhas Submetidas à Intervenção Humana

Para análise do experimento foram consideradas todas as 8 respostas apresentadas seção 2.2. Foi ajustado um modelo multivariado com os efeitos fixos das variáveis sessão experimental, momento e linhagem dispostos de forma aditiva.

Como já mencionado, trata-se de um experimento em que as observações não são independentes pois, devido a questões de delineamento, cada animal contribui com 9 medidas ao conjunto de dados. Faz-se então a necessidade de acomodar as medidas repetidas e as correlações existentes na especificação do modelo.

Existem dois conjuntos distintos de variáveis respostas: contagens e proporções. Para as contagens a função de ligação utilizada para média foi a logarítmica e para função de variância utilizou-se a Poisson-Tweedie. Já para as proporções foi utilizada a função de ligação logito com função de variância Binomial. Adicionalmente, estimou-se o parâmetro de potência para todas as respostas em análise. Os preditores lineares são descritos da forma

$$g_r(\mu_{ijklr}) = \beta_{0r} + \beta 1_{jr}\text{Sessao}_{ij} + \beta 2_{kr}\text{Momento}_{ik} + \beta 3_r\text{Linhagem}_i, \tag{5.2}$$

em que β_0 representa o intercepto, $\beta 1_j$ o efeito de sessão experimental, $\beta 2_k$ o efeito de momento experimental e $\beta 3$ o efeito de linhagem. O índice r refere-se às variáveis respostas do estudo, i varia de 1 até 180 e diz respeito ao número de observações, j indexa as 3 diferentes sessões experimentais: 1 (referência), 2 e 3. k assume os níveis da variável momento experimental: antes (referência), durante e depois. l assume os níveis da variável linhagem: reativo (referência) ou não reativo. E, por fim, g_1 a g_4 equivalem à função de ligação logaritmica, utilizada nas respostas de contagem e g_5 a g_8 representam a função de ligação logito, utilizadas nas respostas que configuram proporções.

Os preditores matriciais, comuns a todas as respostas, são dados por $h\{\boldsymbol{\Omega}(\boldsymbol{\tau})\} = \tau_0 Z_0 + \tau_1 Z_1 + \tau_2 Z_2$ e foram especificados de forma a explicitar que as medidas provenientes do mesmo animal são correlacionadas, bem como as medidas dentro da mesma sessão. A função $h(.)$ utilizada foi a identidade, τ_0, τ_1 e τ_2 representam os parâmetros de dispersão, Z_0 representa uma matriz identidade de ordem 180×180, Z_1 representa uma matriz 180×180 composta por blocos 9×9 de uns na diagonal principal especificando que as medidas provenientes do mesmo animal são correlacionadas, Z_2 representa uma matriz 180×180 composta por blocos 3×3 de uns na diagonal principal indicando que medidas na mesma sessão experimental são correlacionadas. Considerando um único animal, o preditor matricial tem a forma

$$h\{\boldsymbol{\Omega}(\boldsymbol{\tau})\} = \tau_0 \begin{bmatrix} 1&0&0&0&0&0&0&0&0\\ 0&1&0&0&0&0&0&0&0\\ 0&0&1&0&0&0&0&0&0\\ 0&0&0&1&0&0&0&0&0\\ 0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&0&1 \end{bmatrix} + \tau_1 \begin{bmatrix} 1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1 \end{bmatrix} + \tau_2 \begin{bmatrix} 1&1&1&0&0&0&0&0&0\\ 1&1&1&0&0&0&0&0&0\\ 1&1&1&0&0&0&0&0&0\\ 0&0&0&1&1&1&0&0&0\\ 0&0&0&1&1&1&0&0&0\\ 0&0&0&1&1&1&0&0&0\\ 0&0&0&0&0&0&1&1&1\\ 0&0&0&0&0&0&1&1&1\\ 0&0&0&0&0&0&1&1&1 \end{bmatrix}. \tag{5.3}$$

Para fins de análise, os diferentes tempos de observação nos momentos experimentais foram desconsiderados tendo em vista a opinião da pesquisadora responsável pelo experimento. Além disso foi ajustado o modelo com termo offset porém não foi verificada diferença que leve a justificar a inclusão do offset ao modelo.

A Tabela 5 mostra as estimativas dos parâmetros de dispersão τ_i, o qual testa a significância do preditor matricial e as estimativas do parâmento potência p que para

as contagens indicam qual distribuição se adequa e para proporções, valores próximos de 1 indicam adequação à distribuição beta.

Tabela 5 – Estimativas dos parâmetros de dispersão e potência, e erro-padrão (ep).

	τ_0	τ_1	τ_2	p
Resposta	Est. (ep)	Est. (ep)	Est. (ep)	Est. (ep)
ncorpo	2.34 (0.69)	0.42 (0.20)	-0.22 (0.22)	1.28 (0.26)
ncabeca	1.62 (0.38)	0.79 (0.32)	-0.11 (0.17)	1.11 (0.17)
norelha	1.75 (0.88)	0.23 (0.20)	0.14 (0.21)	1.42 (0.28)
nolho	1.67 (3.55)	0.16 (0.38)	0.18 (0.45)	1.21 (0.77)
resporelha2	0.75 (0.14)	-0.02 (0.03)	0.07 (0.07)	1.20 (0.04)
respcauda	0.59 (0.32)	-0.08 (0.06)	0.38 (0.18)	1.29 (0.11)
respolho	1.02 (0.48)	-0.05 (0.08)	0.40 (0.26)	1.54 (0.31)
resporlev	0.06 (0.09)	0.001 (0.01)	0.01 (0.02)	-0.27 (1.01)

Os valores estimados dos parâmetros mostram que o efeito das medidas repetidas observadas em um mesmo animal é mais acentuado no número de mudanças de postura de corpo e cabeça; para o número de variações de postura de orelha e olho e todas as respostas de proporção o parâmetro de dispersão associado é próximo de zero, indicando pouca correlação entre medidas obtidas em um mesmo animal.

Verificou-se ainda que há correlação entre medidas coletadas em uma mesma sessão experimental para a proporção do tempo em que o animal permaneceu com os olhos fechados ou semi-cerrados, proporção do tempo com a cauda em movimento e nas variações de postura de corpo. Para as demais respostas constatou-se que não há evidência de correlação entre medidas dentro das sessões experimentais.

As estimativas dos parâmetros de potência mostram que o número de mudanças de postura de corpo, cabeça e olho seguem distribuição próxima à Neyman Tipo A, já as variações de postura de orelha seguem distribuição próxima à Pólya–Aeppli. Como os parâmetros de potência estão próximos de 1 para a proporção do tempo que o animal permaneceu com as orelhas levantadas ou assimétricas, com a causa em movimento e com os olhos fechados ou semi-cerrados, há um indício de adequação à distribuição beta para estas respostas, o que não ocorre para a proporção do tempo com as orelhas levantadas. Contudo a proximidade de 0 do parâmetro de potência para esta resposta indica que, condicional às covariáveis, esta variável tem distribuição simétrica.

Na parte triangular inferior da Tabela 6 são mostradas as correlações estimadas pelo modelo e na triangular superior são mostrados os erros padrões das estimatiavas. Nota-se que, como esperado, as estimativas das correlações estimadas pelo modelo são mais próximas de 0 que as correlações de Pearson mostradas na Tabela 1. Tal redução deve-se ao fato de que parte da correlação entre as respostas é explicada pelas variáveis explicativas coletadas no estudo.

Tabela 6 – Matriz das correlações estimadas pelo modelo (triangular inferior) e erros padrões das estimatiavas (triangular superior).

	ncorpo	ncabeca	norelha	nolho	resporelha2	respcauda	respolho	resporlev
ncorpo	-	0.074	0.075	0.073	0.075	0.075	0.074	0.075
ncabeca	0.389	-	0.072	0.071	0.075	0.075	0.074	0.075
norelha	0.327	0.344	-	0.073	0.074	0.075	0.074	0.075
nolho	0.252	0.269	0.212	-	0.074	0.074	0.074	0.075
resporelha2	0.014	0.021	0.044	-0.027	-	0.075	0.073	0.068
respcauda	-0.021	-0.019	-0.01	-0.024	-0.004	-	0.074	0.074
respolho	-0.082	-0.057	-0.073	-0.029	-0.130	0.100	-	0.065
resporlev	0.026	-0.009	-0.0120	0.0130	0.283	-0.025	-0.308	-

Com o propósito de verificar a qualidade do ajuste do modelo, foi feita a análise de resíduos do modelo. Na Figura 9 são exibidos os histogramas dos resíduos de Pearson por resposta.

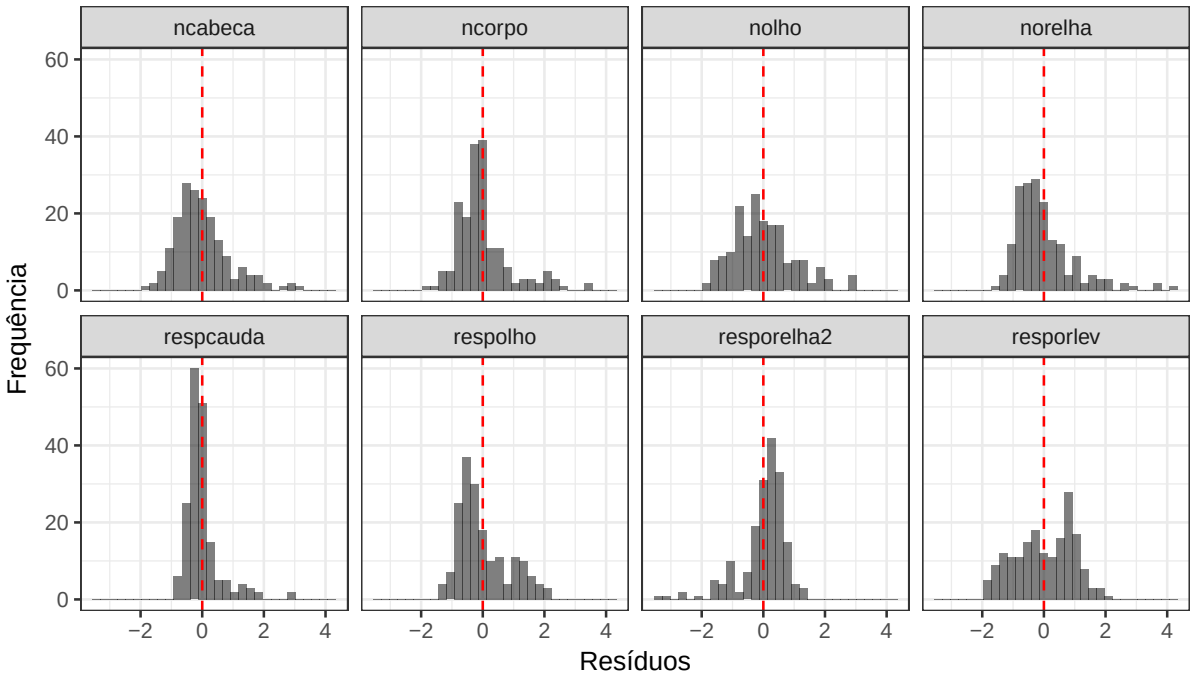


Figura 9 – Histograma dos resíduos de Pearson por resposta.

Devido à assimetria observada no comportamento dos resíduos de Pearson para a maior parte das respostas, investigou-se o comportamento dos erros padrões estimados pelo modelo bem como dos erros robustos e dos erros com vício corrigido. Na Figura 10 são mostradas as razões entre os erros robustos e de vício corrigido com os erros padrões originais. Nota-se que as razões entre os erros estão, na maior parte dos casos, próximas de 1, indicando que os erros Robusto e de Vício Corrigido são próximos aos erros originais estimados pelo modelo.

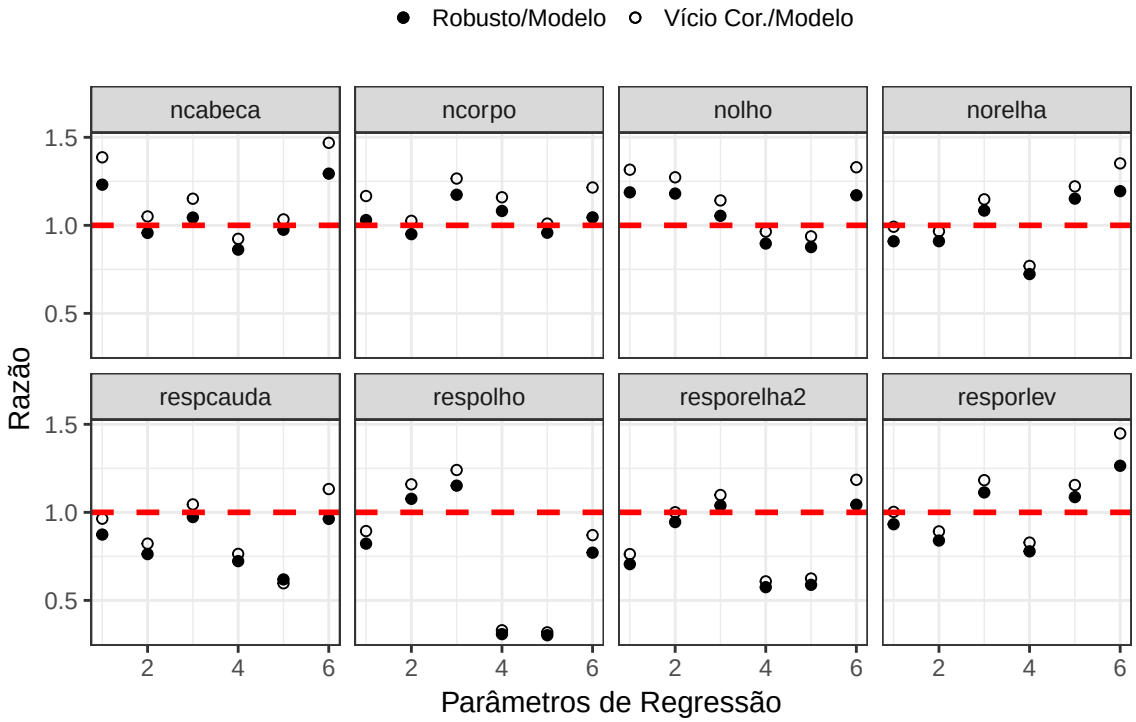


Figura 10 – Razão entre os erros robusto e vício corrigido com os erros do modelo.

Por fim, foi realziada a Análise de Variância Multivariada com o objetivo de testar a nulidade simultânea dos parâmetros estimados no modelo. O resultado da MANOVA é sumarizado na Tabela 7. O resultado do teste de hipótese aponta para a não nulidade simultânea dos parâmetros.

Tabela 7 – Resumo do teste de hipóteses para o modelo ajustado.

Efeitos	GL	W_s	Qui-quadrado	p-valor
Intercepto	8	6.98	1257.72	< 0.01
Sessão	16	0.18	33.27	< 0.01
Momento	16	0.78	140.57	< 0.01
Linhagem	8	0.10	19.41	< 0.01

Os resultados descritos na Tabela 8 mostram que para a variação de posturas de corpo, cabeça, proporção com as orelhas levantadas ou assimétricas e proporção com as orelhas levantadas apenas o efeito de momento foi detectado pelo teste de Wald na análise de variância univariada. Para o número de mudanças de postura de olho, sessão experimental e momento se mostraram significativas. Na proporção do tempo com a cauda em movimento apenas o efeito de linhagem foi detectado. No número de mudanças de postura de orelha todas as variáveis explicativas tiveram efeito constatado. Por fim, nenhuma das variáveis se mostrou significativa na proporção do tempo com os olhos fechados e semi-cerrados, este resultado já era esperado devido ao baixo poder de explicação das covariáveis constatado na Figura 3.

Tabela 8 – Medidas de ajuste de cada resposta para avaliação individual dos fatores.

Covariável	χ^2	GL	p-valor	χ^2	GL	p-valor	χ^2	GL	p-valor	χ^2	GL	p-valor
Contagens												
	ncorpo			ncabeça			norelha			nolho		
Sessão	0,83	2	0,65	3,47	2	0,17	8,97	2	0,01	16,37	2	<0,01
Momento	20,26	2	<0,01	54,69	2	<0,01	48,85	2	<0,01	6,25	2	0,04
Linhagem	0,22	1	0,63	0	1	0,99	4,05	1	0,04	1,39	1	0,23
Proporções												
	resporelha2			respcauda			respolho			resporlev		
Sessão	1,58	2	0,45	1,47	2	0,47	2,43	2	0,29	4,59	2	0,10
Momento	47,50	2	<0,01	0,02	2	0,98	2,97	2	0,22	19,13	2	<0,01
Linhagem	0,13	1	0,70	12,31	1	<0,01	0,34	1	0,55	0,01	1	0,91

6 Considerações Finais

O objetivo desse trabalho foi propor o teste de hipóteses multivariado para dados não gaussianos no contexto dos MCGLM construído para lidar com casos em que haja mais de uma variável resposta e que estas respostas sejam correlacionadas e não necessariamente apresentem distribuição gaussiana. O teste proposto é baseado na estatística de Wald para múltiplas respostas que, sob hipótese nula, segue distribuição Qui-quadrado com sR graus de liberdade. A construção do teste depende da especificação de uma matriz $\boldsymbol{L}$ e tem como objetivo verificar se há subsídio para aceitação da hipótese de nulidade conjunta de todos os parâmetros de um modelo multivariado, desde que todas as respostas estejam sujeitas às mesmas variáveis explicativas.

Com o intuito de verificar o poder do teste proposto sob diferentes distribuições de variável resposta, delineou-se um estudo de simulação com cenários em que se variou tamanho amostral e fixou-se diferentes valores para os parâmetros de regressão. Foram explorados cenários com 3 variáveis respostas, todas seguindo uma mesma distribuição de probabilidades: Poisson, Binomial com $n = 10$ ou Beta. As respostas estavam sujeitas a uma mesma matriz de correlação e todas a uma mesma variável explicativa categórica de 5 níveis. Para cada possível cenário foram gerados 1000 conjuntos de dados e, para cada um deles, ajustou-se um modelo multivariado e realizou-se o teste de hipótese proposto. Os resultados apontaram para o bom funcionamento do teste para tamanhos de amostra maior que 250. Nesses casos, para todas as distribuições de variável resposta, atingiu-se os 95% de confiança quando os parâmetros de regressão foram fixados em 0.

Comprovado o poder do teste, fez-se a aplicação do mesmo a dois conjuntos de dados reais, o primeiro deles refere-se à análise de processos movidos contra grandes litigantes e o segundo à análise comportamental de ovelhas submetidas à intervenção humana. A principal diferença entre ambas as análises está na estrutura dos dados, em que no primeiro caso tem-se 3 respostas binárias e observações independentes, e, no segundo caso tem-se respostas de diferentes naturezas, das quais 4 são proporções e 4 são contagens, além de apresentar uma estrutura de medidas repetidas. Para ambas as análises, o teste de hipóteses proposto aponta para a rejeição da hipótese nula.

Dado o escopo do trabalho, possíveis novos trabalhos compreendem a construção de outros testes multivariados para dados não gaussianos no contexto dos MCGLMs, como por exemplo, o análogo ao teste da razão de verossimilhanças. Ampliar o estudo de simulação, de forma a considerar cenários não abrangidos, como por exemplo ampliar o número de respostas, considerar casos com respostas seguindo diferentes distribuições, simular amostras com observações não independentes e aumentar o

tamanho amostral.

REFERÊNCIAS

ANDERSON, T. et al. Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 1, n. 1, p. 135–141, 1973. Citado na página 20.

BONAT, W. H. Multiple response variables regression models in R: The mcglm package. *Journal of Statistical Software*, v. 84, n. 4, p. 1–30, 2018. Citado na página 21.

BONAT, W. H.; JØRGENSEN, B. Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 65, n. 5, p. 649–675, 2016. Citado 4 vezes nas páginas 12, 19, 21 e 22.

BOX, G. E.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 211–252, 1964. Citado na página 11.

CARIO, M. C.; NELSON, B. L. *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix*. [S.l.], 1997. Citado na página 25.

DEMIDENKO, E. *Mixed models: theory and applications with R*. [S.l.]: John Wiley & Sons, 2013. Citado na página 20.

HOTELLING, H. *A generalized T test and measure of multivariate dispersion*. [S.l.], 1951. Citado 2 vezes nas páginas 12 e 23.

JØRGENSEN, B. Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 49, n. 2, p. 127–145, 1987. Citado na página 20.

JØRGENSEN, B. *The theory of dispersion models*. [S.l.]: CRC Press, 1997. Citado na página 20.

JØRGENSEN, B.; KNUDSEN, S. J. Parameter orthogonality and bias adjustment for estimating functions. *Scandinavian Journal of Statistics*, Wiley Online Library, v. 31, n. 1, p. 93–114, 2004. Citado 2 vezes nas páginas 21 e 22.

JØRGENSEN, B.; KOKONENDJI, C. C. Discrete dispersion models and their tweedie asymptotics. *AStA Advances in Statistical Analysis*, Springer, v. 100, n. 1, p. 43–78, 2015. Citado na página 20.

LAWLEY, D. A generalization of fisher's z test. *Biometrika*, JSTOR, v. 30, n. 1/2, p. 180–187, 1938. Citado 2 vezes nas páginas 12 e 23.

LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, Oxford University Press, v. 73, n. 1, p. 13–22, 1986. Citado 2 vezes nas páginas 20 e 21.

MARTINEZ-BENEITO, M. A. A general modelling framework for multivariate disease mapping. *Biometrika*, Oxford University Press, v. 100, n. 3, p. 539–553, 2013. Citado na página 21.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, v. 135, p. 370–384, 1972. Citado na página 11.

PILLAI, K. et al. Some new test criteria in multivariate analysis. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 26, n. 1, p. 117–121, 1955. Citado 2 vezes nas páginas 12 e 23.

PINHEIRO, J. C.; BATES, D. M. Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing*, Springer, v. 6, n. 3, p. 289–296, 1996. Citado na página 20.

POURAHMADI, M. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, Oxford University Press, v. 87, n. 2, p. 425–435, 2000. Citado na página 20.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>. Citado na página 21.

ROY, S. N. On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, JSTOR, p. 220–238, 1953. Citado 2 vezes nas páginas 12 e 23.

SMITH, H.; GNANADESIKAN, R.; HUGHES, J. Multivariate analysis of variance (manova). *Biometrics*, JSTOR, v. 18, n. 1, p. 22–41, 1962. Citado 2 vezes nas páginas 12 e 23.

TAMIOSO, P. R.; BOISSY, A.; BOIVIN, X.; CHANDEZE, H.; ANDANSON, S.; DELVAL, E.; HAZARD, D.; TACONELI, C. A.; MOLENTO, C. F. M. Does emotional reactivity influence behavioral and cardiac responses of ewes submitted to brushing? *Behavioural Processes*, p. np, 2017. Citado na página 15.

TRECENTI, J. A. Z. *Diagramas de influência: uma aplicação em Jurimetria*. Dissertação (Mestrado) — Instituto de Matemática e Estatística, Universidade de São Paulo, 2015. Citado na página 13.

WILKS, S. S. Certain generalizations in the analysis of variance. *Biometrika*, JSTOR, p. 471–494, 1932. Citado 2 vezes nas páginas 12 e 23.