LINEU ALBERTO CAVAZANI DE FREITAS

TESTES DE HIPÓTESE EM MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA (MCGLM)

(*versão pré-defesa, compilada em 30 de março de 2021*)

Documento apresentado como requisito parcial ao exame de qualificação de Mestrado no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Prof. Dr. Wagner Hugo Bonat.

Coorientador: Prof. Dr. Marco Antonio Zanata Alves.

CURITIBA PR

2021

SUMÁRIO

1 INTRODUÇÃO

Podemos entender como Ciência de Dados o estudo sistemático de conjuntos de dados com o objetivo de gerar conhecimento sobre determinado assunto. Em suma, o objetivo da Ciência de Dados é extrair informação. Seu processo é caracterizado por etapas como a definição do problema, planejamento do estudo, coleta e análise dos dados e, por fim, a interpretação dos resultados.

Trata-se de um campo de estudo extremamente interdisciplinar que envolve técnicas de áreas como Estatística, Ciência da Computação e Matemática. É uma área que vem ganhando destaque nos últimos anos devido a fatores tais como a popularização do uso de dados nas tomadas de decisão em diversos cenários, difusão do uso de grandes bancos de dados, desenvolvimento e propagação de técnicas modernas e eficientes de análise, sem mencionar o desenvolvimento computacional que permitiu a implementação de técnicas mais complexas para solução de problemas e também que mais pessoas tivessem acesso às técnicas e ferramentas necessárias para se analisar dados.

Alguns dos campos de interesse na Ciência de Dados são: métodos de amostragem, mineração de dados, bancos de dados, técnicas de análise exploratória, probabilidade, inferência, otimização, infraestrutura computacional, plataformas de Big Data, modelos estatísticos, dentre outros.

No contexto de modelos estatísticos, existem os chamados modelos de regressão, dentre os quais podemos citar: os modelos lineares, lineares generalizados, aditivos generalizados, de efeitos aleatórios, aditivos generalizados para locação, escala e forma e ainda os multivariados.

Os modelos de regressão são indicados a problemas nos quais temos interesse em verificar a associação entre uma ou mais variáveis resposta e um conjunto de variáveis explicativas; podemos ainda, além de verificar associação, utilizar o modelo para realizar predições para uma população.

Nos casos univariados mais gerais, estes modelos associam uma única variável resposta, também chamada de variável dependente, a uma ou mais variáveis explicativas, conhecidas como variáveis independentes, covariáveis ou preditoras.

De forma geral, um modelo de regressão é uma expressão matemática que relaciona a média da variável resposta às variáveis preditoras, 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, no qual um dos pressupostos é de que a variável resposta, condicional às variáveis explicativas, siga distribuição Normal. Todavia, não são raras as situações em que a suposição de normalidade não é atendida. Uma alternativa, por muito tempo adotada, foi buscar uma transformação da variável resposta a fim de atender os pressupostos do modelo, tal como a família de transformações Box-Cox (Box e Cox, 1964). Contudo, este tipo de solução leva a dificuldades na interpretação dos resultados.

Neste contexto, a proposta de maior renome para contornar tais restrições foram os Modelos Lineares Generalizados (GLM) propostos por Nelder e Wedderburn (1972). Essa classe de modelos permitiu a flexibilização da distribuição da variável resposta de tal modo 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. 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 (McGLM) propostos por Bonat e Jørgensen (2016). Essa classe permite lidar com múltiplas respostas de diferentes naturezas e, de alguma forma, correlacionadas. Além disso, não há nesta classe suposições quanto à independência entre as observações da amostra, pois a correlação entre observações pode ser modelada por um preditor linear matricial que envolve matrizes conhecidas.

De forma geral, o McGLM é uma estrutura para modelagem de múltiplas respostas, de diferentes naturezas, em que não há necessidade de observações independentes. Estas características tornam o McGLM uma classe flexível ao ponto de ser possível chegar a extensões multivariadas para modelos de medidas repetidas, séries temporais, dados longitudinais, espaciais e espaço-temporais.

Quando trabalhamos com modelos de regressão, por diversas vezes há o interesse em avaliar os parâmetros do modelo. Isto é, verificar se os valores que associam as variáveis explicativas às variáveis respostas são iguais a determinados valores de interesse. Isto é feito através dos chamados testes de hipótese.

Em geral, existe o interesse em avaliar se há evidência suficiente para afirmar que o parâmetro que associa a variável explicativa à variável resposta é igual a 0, pois, caso esta afirmação seja verdadeira, podemos concluir que a variável explicativa não está associada à variável resposta. Contudo, através dos testes de hipótese podemos avaliar outros valores diferentes de 0.

Para o caso dos modelos lineares tradicionais existem técnicas como a Análise de Variância (ANOVA), na qual o objetivo é analisar o efeito de cada uma das variáveis explicativas, isto é, avaliar se a retirada de cada variável gera perda ao modelo ajustado. Em outras palavras, na Análise de Variância realizamos sucessivos testes de hipótese para verificar se o parâmetro que associa a variável explicativa à variável resposta é igual a 0.

Quando se está na classe de modelos multivariados para dados gaussianos, extende-se o conceito de Análise de Variância (ANOVA) para a Análise de Variância Multivariada (Smith et al., 1962), a MANOVA. 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) e (Hotelling, 1951), traço de Pillai (Pillai et al., 1955) e maior raiz de Roy (Roy, 1953).

No entanto, considerando o cenário com múltiplas respostas não gaussianas, são escassas as discussões na literatura a respeito de testes de hipótese sobre os parâmetros do modelo. Deste modo, nosso objetivo geral é o desenvolvimento destes testes de hipótese para os Modelos Multivariados de Covariância Linear Generalizada (McGLM) por se tratar de uma classe de modelos flexível e com alto poder de aplicação a problemas práticos em que se fazem necessários tais testes para avaliação do modelo.

Portanto, este trabalho tem os seguintes objetivos específicos:

1. Adaptar e implementar testes de hipótese gerais para os Modelos Multivariados de Covariância Linear Generalizada.

2. Desenvolver um procedimento para efetuar Análises de Variância Multivariadas para os Modelos Multivariados de Covariância Linear Generalizada.

3. Demonstrar as propriedades e comportamento dos testes propostos com base em estudos de simulação.

4. Demonstrar o potencial de aplicação das metodologias discutidas com base na aplicação a conjuntos de dados reais.

Este trabalho está organizado em sete capítulos: na atual seção foi exposto o tema de forma a enfatizar as características dos modelos lineares e testes de hipóteses. O Capítulo 2 é dedicado à revisão bibliográfica da estrutura dos McGLM. No Capítulo 3 são apresentados e discutidos os testes de hipótese já no contexto dos McGLM. No capítulo 4 são mostradas as funções implementadas. O Capítulo 5 apresenta o escopo do estudo de simulação para verificar as principais propriedades dos testes propostos. O Capítulo 6 apresenta os conjuntos de dados que serão usados no trabalho com o objetivo de discutir a aplicação do método a conjuntos de dados reais. E, por fim, no Capítulo 7 são apresentados os comentários finais e são discutidos os resultados esperados do estudo.

2 MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA

Os Modelos Linerares Generalizados (GLM), propostos por Nelder e Wedderburn (1972), são uma forma de modelagem univariada para dados de diferentes naturezas, tais como respostas contínuas, binárias e contagens. Tais características tornam essa classe de modelos uma flexível ferramenta de modelagem aplicável a diversos tipos de problema. Contudo, por mais flexível e discutida na literatura, essa classe apresenta duas principais restrições:

1. A incapacidade de lidar com observações dependentes.
2. E/ou a incapacidade de lidar com múltiplas respostas simultaneamente.

Com o objetivo de contornar estas restrições, 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: os chamados Modelos Multivariados de Covariância Linear Generalizada (McGLM).

Vamos discutir os McGLM como uma extensão dos GLM. Vale ressaltar que é usada uma especificação menos usual de um Modelo Linear Generalizado, porém trata-se de uma notação mais conveniente para chegar à uma especificação melhor construída de um Modelo Multivariado de Covariância Linear Generalizada.

2.1 GLM

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 GLM pode ser descrito da forma

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}) &= \boldsymbol{\mu} = g^{-1}(\boldsymbol{X}\boldsymbol{\beta}), \\ \mathrm{Var}(\boldsymbol{Y}) &= \Sigma = \mathrm{V}\left(\boldsymbol{\mu}; p\right)^{1/2} \left(\tau_0 \boldsymbol{I}\right) \mathrm{V}\left(\boldsymbol{\mu}; p\right)^{1/2}, \end{aligned} \tag{2.1}$$

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 GLM 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:

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

2. 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 e Kokonendji, 2015).

3. 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.

Lembre-se que o GLM é uma classe de modelos de regressão univariados em que um dos pressupostos é a indepenência entre as observações. Esta independência é especificada na matriz identidade no centro da equação que define a matriz de variância e covariância. Podemos imaginar que, substituindo esta matriz identidade por uma matriz qualquer que reflita a relação entre os indivíduos da amostra teremos uma extensão do Modelo Linear Generalizado para observações dependentes. É justamente essa a ideia dos Modelos de Covariância Linear Generalizada, o cGLM.

2.2 CGLM

Os cGLM são 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 $\boldsymbol{\Omega(\tau)}$ 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 (Equações de Estimativas Generalizadas), em que utiliza-se uma matriz de correlação de trabalho para considerar a dependência entre as observações. A matriz $\boldsymbol{\Omega(\tau)}$ 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

$$h\left\{\boldsymbol{\Omega}(\boldsymbol{\tau})\right\} = \tau_0 Z_0 + \ldots + \tau_D Z_D, \tag{2.2}$$

em que $h(.)$ é a função de ligação de covariância, Z_d com $d = 0,\ldots,$ 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 modelos da literatura 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, temos uma classe com toda a flexibilidade dos GLM, porém contornando a restrição da independência entre as observações desde que o preditor linear matricial seja adequadamente especificado.

Deste modo, é contornada a primeira restrição dos GLM. A segunda restrição diz respeito às múltiplas respostas e, resolvendo esta restrição, chegamos ao McGLM.

2.3 MCGLM

O McGLM pode ser entendido como uma extensão multivariada do cGLM e contorna as duas principais restrições presentes nos GLM, pois além de permitir a modelagem de dados com estrutura de covariância, permite modelar múltiplas respostas.

Considre $\boldsymbol{Y}_{N\times R} = \{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_R\}$ uma matriz de variáveis resposta e $\boldsymbol{M}_{N\times R} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_R\}$ uma matriz de valores esperados. Cada uma das variáveis resposta tem sua própria matriz de variância e covariância, responsável por modelar a covariância dentro de cada resposta, sendo expressa por

$$\Sigma_r = \mathrm{V}_r\left(\boldsymbol{\mu}_r; p\right)^{1/2} \boldsymbol{\Omega}_r\left(\boldsymbol{\tau}\right) \mathrm{V}_r\left(\boldsymbol{\mu}_r; p\right)^{1/2}. \tag{2.3}$$

Além disso, é necessária uma matriz de correlação Σ_b, de ordem $R \times R$, que descreve a correlação entre as variáveis resposta. 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

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}) &= \boldsymbol{M} = \{g_1^{-1}(\boldsymbol{X}_1\boldsymbol{\beta}_1), \dots, g_R^{-1}(\boldsymbol{X}_R\boldsymbol{\beta}_R)\} \\ \mathrm{Var}(\boldsymbol{Y}) &= \boldsymbol{C} = \boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b, \end{aligned} \tag{2.4}$$

em que $\boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b = \mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \dots, \tilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b \otimes \boldsymbol{I})\mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^\top, \dots, \tilde{\boldsymbol{\Sigma}}_R^\top)$ 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).

2.3.1 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 McGLM permitem a divisão dos parâmetros em dois conjuntos: $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\lambda}^\top)^\top$. Desta forma, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_R^\top)^\top$ 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 resposta $\boldsymbol{Y}_{N\times R}$ 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 e Zeger, 1986), representada por

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \boldsymbol{D}^\top \boldsymbol{C}^{-1}(\mathcal{Y} - \mathcal{M}), \tag{2.5}$$

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

$$S_{\boldsymbol{\beta}} = E(\nabla_{\boldsymbol{\beta}\psi\boldsymbol{\beta}}) = -\boldsymbol{D}^\top \boldsymbol{C}^{-1} \boldsymbol{D}, \tag{2.6}$$

enquanto que a matriz $K \times K$ de variabilidade de $\psi_{\boldsymbol{\beta}}$ é escrita como

$$V_{\boldsymbol{\beta}} = VAR(\psi\boldsymbol{\beta}) = \boldsymbol{D}^\top \boldsymbol{C}^{-1} \boldsymbol{D}. \tag{2.7}$$

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

$$\psi_{\lambda_i}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathrm{tr}(W_{\lambda i}(\boldsymbol{r}^\top \boldsymbol{r} - \boldsymbol{C})), i = 1, .., Q, \tag{2.8}$$

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

$$S_{\lambda_{ij}} = E\left(\frac{\partial}{\partial \boldsymbol{\lambda_i}} \psi \boldsymbol{\lambda_j}\right) = -tr(W_{\lambda_i} C W_{\lambda_j} C). \tag{2.9}$$

Já a entrada (i, j) da matriz de variabilidade $Q \times Q$ de ψ_λ é definida por

$$V_{\lambda_{ij}} = Cov\left(\psi_{\lambda_i}, \psi_{\lambda_j}\right) = 2tr(W_{\lambda_i} C W_{\lambda_j} C) + \sum_{l=1}^{NR} k_l^{(4)} (W_{\lambda_i})_{ll} (W_{\lambda_j})_{ll}, \tag{2.10}$$

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

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

$$S_\theta = \begin{bmatrix} S_\beta & S_{\beta\lambda} \\ S_{\lambda\beta} & S_\lambda \end{bmatrix} \text{ e } V_\theta = \begin{bmatrix} V_\beta & V_{\lambda\beta}^\top \\ V_{\lambda\beta} & V_\lambda \end{bmatrix}. \tag{2.11}$$

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

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, J_{\boldsymbol{\theta}}^{-1}), \tag{2.12}$$

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

$$\begin{aligned} \boldsymbol{\beta}^{(i+1)} &= \boldsymbol{\beta}^{(i)} - S_{\boldsymbol{\beta}}^{-1} \psi \boldsymbol{\beta}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\lambda}^{(i)}), \\ \boldsymbol{\lambda}^{(i+1)} &= \boldsymbol{\lambda}^{(i)} \alpha S_{\boldsymbol{\lambda}}^{-1} \psi \boldsymbol{\lambda}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}). \end{aligned} \tag{2.13}$$

REFERÊNCIAS

Anderson, T. et al. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, 1(1):135–141.

Bonat, W. H. (2018). Multiple response variables regression models in R: The mcglm package. *Journal of Statistical Software*, 84(4):1–30.

Bonat, W. H. e Jørgensen, B. (2016). Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(5):649–675.

Box, G. E. e Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, páginas 211–252.

Demidenko, E. (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.

Hotelling, H. (1951). A generalized t test and measure of multivariate dispersion. Relatório técnico, UNIVERSITY OF NORTH CAROLINA Chapel Hill United States.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(2):127–145.

Jørgensen, B. (1997). *The theory of dispersion models*. CRC Press.

Jørgensen, B. e Knudsen, S. J. (2004). Parameter orthogonality and bias adjustment for estimating functions. *Scandinavian Journal of Statistics*, 31(1):93–114.

Jørgensen, B. e Kokonendji, C. C. (2015). Discrete dispersion models and their tweedie asymptotics. *AStA Advances in Statistical Analysis*, 100(1):43–78.

Lawley, D. (1938). A generalization of fisher's z test. *Biometrika*, 30(1/2):180–187.

Liang, K.-Y. e Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Martinez-Beneito, M. A. (2013). A general modelling framework for multivariate disease mapping. *Biometrika*, 100(3):539–553.

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

Pillai, K. et al. (1955). Some new test criteria in multivariate analysis. *The Annals of Mathematical Statistics*, 26(1):117–121.

Pinheiro, J. C. e Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing*, 6(3):289–296.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 87(2):425–435.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, páginas 220–238.

Smith, H., Gnanadesikan, R. e Hughes, J. (1962). Multivariate analysis of variance (manova). *Biometrics*, 18(1):22–41.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, páginas 471–494.