




## Testes de hipóteses para regressão multivariada com dados não gaussianos em R: The `htmglm` Package

Lineu Alberto Cavazani de Freitas   
Paraná Federal University

Wagner Hugo Bonat   
Paraná Federal University

---

### Abstract

300 palavras

*Keywords:* keyword1, keyword2, keyword3, keyword4, keyword5, keyword6, keyword7, keyword8, keyword9, keyword10.

---

## 1. Introduction

The `htmglm` package for R (R Core Team 2022) provides functions to efetuar testes de hipóteses sobre parâmetros de multivariate covariance generalized linear models (McGLMs; Bonat and Jørgensen (2016)) ajustados usando o pacote `mcglm` (Bonat 2018).

McGLMs provide a general statistical modeling framework for normal and non-normal multivariate data analysis along with a wide range of correlation structures. Quando trabalhamos na classe dos McGLMs estimamos parâmetros de regressão, dispersão, potência e correlação. Cada conjunto de parâmetros possui interpretação prática bastante útil.

Por meio do estudo dos parâmetros de regressão é possível avaliar o efeito da(s) variável(is) explicativa(s) sobre a(s) resposta(s). Por meio do estudo dos parâmetros de dispersão pode-se avaliar o efeito da correlação entre unidades do estudo, muito útil em situações em que as observações do conjunto de dados são correlacionadas entre si, como por exemplo em estudos longitudinais, temporais e de medidas repetidas. Os parâmetros de potência nos fornecem um indicativo de qual distribuição de probabilidade melhor se adequa ao problema. E os parâmetros de correlação estimam a força da associação entre respostas em um problema multivariado.

O desenvolvimento de testes de hipóteses para fins de avaliação destas quantidades é de grande valia em problemas práticos e leva a formas procedurais para avaliação das quantidades resultantes do modelo. The `htmglm` package is a full R implementation e faz uso da estatística

de Wald para avaliar parâmetros de regressão e dispersão. The features include funções para testes de hipóteses lineares gerais, quadros de análise de variância uni e multivariados, bem como testes de comparações múltiplas.

**Package htmglm is available from the Comprehensive R Archive Network (CRAN) at <https://CRAN.R-project.org/package=mcglm> e complementam as já possíveis análises permitidas pelo pacote *mcglm***

No que diz respeito à implementações do teste Wald em outros contextos no R, o pacote **lmtest** (Zeileis and Hothorn 2002) possui uma função genérica para realizar testes Wald para comparar modelos lineares e lineares generalizados aninhados. Já o pacote **survey** (Lumley 2020, 2004, 2010) possui uma função que efetua testes Wald que, por padrão, testa se todos os coeficientes associados a um determinado termo de regressão são zero, mas é possível especificar hipóteses com outros valores.

O pacote **car** (Fox and Weisberg 2019) possui uma implementação para testar hipóteses lineares sobre parâmetros de modelos lineares, modelos lineares generalizados, modelos lineares multivariados, modelos de efeitos mistos, dentre outros; nesta implementação o usuário tem total controle de que parâmetros testar e com quais valores confrontar na hipótese nula.

Quanto aos quadros de análise de variância, o R possui a função `anova()` no pacote padrão **stats** (R Core Team 2022) aplicável a modelos lineares e lineares generalizados. Já o pacote **car** (Fox and Weisberg 2019) possui uma função que retorna quadros de análise de variância dos tipos II e III para diversos modelos. Para comparações múltiplas, um dos principais pacotes disponíveis é o **multcomp** (Hothorn, Bretz, and Westfall 2008) que fornece uma interface para testes de comparações múltiplas para modelos paramétricos.

Contudo, quando se trata de modelos multivariados de covariância linear generalizada ajustados no pacote **mcglm** existe apenas um tipo de análise de variância univariada implementada na biblioteca e não existem opções para realização de testes de hipóteses lineares gerais, nem testes de comparações múltiplas. Portanto, por se tratar de uma classe de modelos flexível e com alto poder de aplicação a problemas práticos, nosso objetivo geral é fornecer implementações que permitam efetuar testes de hipóteses para os McGLMs de tal modo que seja possível testar hipóteses lineares gerais, gerar quadros de análise variância, análise de variância multivariada e testes de comparações múltiplas.

The article is organized as follows. In [section 2](#) we present a revisão da estrutura geral e estimação dos parâmetros de um McGLM, baseado nas ideias de Bonat and Jørgensen (2016). In [section 3](#) são apresentados os detalhes do teste Wald para avaliar suposições sobre parâmetros de um McGLM. [section 4](#) introduces the R implementation discussing the main functions available in the **htmglm** package. [section 5](#) illustrates the package usage through some examples. Finally, [section 6](#) presents a discussion and directions for future work on the improvement of the **htmglm** package.

## 2. Multivariate covariance generalized linear models

Considere  $\mathbf{Y}_{N \times R} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_R\}$  uma matriz de variáveis resposta e  $\mathbf{M}_{N \times R} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_R\}$  uma matriz de valores esperados. A matriz de variância e covariância para cada resposta  $r$ ,  $r = 1, \dots, R$  é denotada por  $\Sigma_r$ , tem dimensão  $N \times N$ . Além disso, é necessária uma matriz de correlação  $\Sigma_b$ , de ordem  $R \times R$ , que descreve a correlação entre as variáveis resposta. Os McGLMs (Bonat and Jørgensen 2016) são definidos por:

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

em que as funções  $g_r()$  são as tradicionais funções de ligação;  $\mathbf{X}_r$  denota uma matriz de delineamento  $N \times k_r$ ;  $\boldsymbol{\beta}_r$  denota um vetor  $k_r \times 1$  de parâmetros de regressão.  $\boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b = \text{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \dots, \tilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b \otimes \mathbf{I})\text{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^\top, \dots, \tilde{\boldsymbol{\Sigma}}_R^\top)$  é o produto generalizado de Kronecker (Martinez-Beneito 2013), a matriz  $\tilde{\boldsymbol{\Sigma}}_r$  denota a matriz triangular inferior da decomposição de Cholesky da matriz  $\boldsymbol{\Sigma}_r$ . O operador  $\text{Bdiag}()$  denota a matriz bloco-diagonal e  $\mathbf{I}$  é uma matriz identidade  $N \times N$ .

Para variáveis resposta contínuas, binárias, binomiais, proporções ou índices a matriz de variância e covariância  $\boldsymbol{\Sigma}_r$  é dada por:

$$\boldsymbol{\Sigma}_r = \mathbf{V}(\boldsymbol{\mu}_r; p_r)^{1/2} (\boldsymbol{\Omega}(\boldsymbol{\tau}_r)) \mathbf{V}(\boldsymbol{\mu}_r; p_r)^{1/2}.$$

No caso de variáveis resposta que sejam contagens a matriz de variância e covariância para cada variável resposta fica dada por:

$$\boldsymbol{\Sigma}_r = \text{diag}(\boldsymbol{\mu}_r) + \mathbf{V}(\boldsymbol{\mu}_r; p_r)^{1/2} (\boldsymbol{\Omega}(\boldsymbol{\tau}_r)) \mathbf{V}(\boldsymbol{\mu}_r; p_r)^{1/2},$$

em que  $\mathbf{V}(\boldsymbol{\mu}_r; p_r) = \text{diag}(\vartheta(\boldsymbol{\mu}_r; p_r))$  denota uma matriz diagonal na qual as entradas são dadas pela função de variância  $\vartheta(\cdot; p_r)$  aplicada aos elementos do vetor  $\boldsymbol{\mu}_r$ . Diferentes escolhas de funções de variância  $\vartheta(\cdot; p_r)$  implicam em diferentes suposições a respeito da distribuição da variável resposta. Mencionaremos 3 opções de funções de variância: função de variância potência, função de dispersão Poisson–Tweedie e função de variância binomial.

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

A função de dispersão Poisson–Tweedie (Jørgensen and Kokonendji 2015) visa contornar a inflexibilidade da utilização da função de variância potência para respostas que caracterizam contagens. A função de dispersão é dada por  $\vartheta(\cdot; p) = \mu + \tau \mu^p$  em que  $\tau$  é o parâmetro de dispersão. Temos assim uma rica classe de modelos para lidar com respostas que caracterizam contagens, uma vez que muitas distribuições importantes aparecem como casos especiais, tais como: Hermite ( $p = 0$ ), Neyman tipo A ( $p = 1$ ), binomial negativa ( $p = 2$ ) e Poisson–inversa gaussiana ( $p = 3$ ).

Por fim, a função de variância binomial, dada por  $\vartheta(\cdot; p_r) = \mu_r^{p_{r1}}(1 - \mu_r)^{p_{r2}}$  é indicada quando a variável resposta é binária, restrita a um intervalo ou quando tem-se o número de sucessos em um número de tentativas.

É possível notar que o parâmetro de potência  $p$  aparece em todas as funções de variância discutidas. Este parâmetro tem especial importância pois trata-se de um índice que distingue diferentes distribuições de probabilidade importantes no contexto de modelagem e, por esta razão, pode ser utilizado como uma ferramenta para seleção automática da distribuição de probabilidade que mais se adequa ao problema.

A matriz de dispersão  $\boldsymbol{\Omega}(\boldsymbol{\tau})$  descreve a parte da covariância dentro de cada variável resposta que não depende da estrutura média, isto é, a estrutura de correlação entre as observações da

amostra. Baseando-se nas ideias de [Anderson et al. \(1973\)](#) e [Pourahmadi \(2000\)](#), [Bonat and Jørgensen \(2016\)](#) propuseram modelar a matriz de dispersão através de um preditor linear matricial combinado com uma função de ligação de covariância dada por:

$$h\{\boldsymbol{\Omega}(\boldsymbol{\tau}_r)\} = \tau_{r0}Z_0 + \dots + \tau_{rD}Z_D,$$

em que  $h()$  é a função de ligação de covariância,  $Z_{rd}$  com  $d = 0, \dots, D$  são matrizes que representam a estrutura de covariância presente em cada variável resposta  $r$  e  $\boldsymbol{\tau}_r = (\tau_{r0}, \dots, \tau_{rD})$  é um vetor  $(D + 1) \times 1$  de parâmetros de dispersão.

Algumas possíveis funções de ligação de covariância são a identidade, inversa e exponencial-matriz. A especificação da função de ligação de covariância é discutida por [Pinheiro and 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 \(2013\)](#).

Deste modo, os McGLMs configuram uma estrutura geral para análise via modelos de regressão para dados não gaussianos com múltiplas respostas em que não se faz suposições quanto à independência das observações. A classe é definida por 3 funções (de ligação, de variância e de covariância) além de um preditor linear e um preditor linear matricial para cada resposta sob análise.

## 2.1. Estimação e inferência

Os McGLMs são ajustados baseados no método de funções de estimação descritos em detalhes por [Bonat and Jørgensen \(2016\)](#) e [Jørgensen and Knudsen \(2004\)](#). Nesta subseção 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  $\boldsymbol{\lambda} = (\rho_1, \dots, \rho_{R(R-1)/2}, p_1, \dots, p_R, \boldsymbol{\tau}_1^\top, \dots, \boldsymbol{\tau}_R^\top)^\top$  é um vetor  $Q \times 1$  de parâmetros de dispersão. Além disso,  $\mathcal{Y} = (\mathbf{Y}_1^\top, \dots, \mathbf{Y}_R^\top)^\top$  denota o vetor empilhado de ordem  $NR \times 1$  da matriz de variáveis resposta  $\mathbf{Y}_{N \times R}$  e  $\mathcal{M} = (\boldsymbol{\mu}_1^\top, \dots, \boldsymbol{\mu}_R^\top)^\top$  denota o vetor empilhado de ordem  $NR \times 1$  da matriz de valores esperados  $\mathbf{M}_{N \times R}$ .

Para estimação dos parâmetros de regressão é utilizada a função quasi-score ([Liang and Zeger 1986](#)), representada por

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

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

$$\mathbf{S}_\beta = E(\nabla_\beta \psi_\beta) = -\mathbf{D}^\top \mathbf{C}^{-1} \mathbf{D},$$

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

$$\mathbf{V}_\beta = \text{VAR}(\psi_\beta) = \mathbf{D}^\top \mathbf{C}^{-1} \mathbf{D}.$$

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

$$\psi_{\lambda_i}(\beta, \lambda) = \text{tr}(W_{\lambda_i}(\mathbf{r}^\top \mathbf{r} - \mathbf{C})), i = 1, \dots, Q,$$

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

$$S_{\lambda_{ij}} = E \left( \frac{\partial}{\partial \lambda_i} \psi_{\lambda_j} \right) = -\text{tr}(W_{\lambda_i} \mathbf{C} W_{\lambda_j} \mathbf{C}).$$

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

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

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  $\beta$  e  $\lambda$ , Bonat and Jørgensen (2016) obtiveram as matrizes de sensibilidade e variabilidade cruzadas, denotadas por  $S_{\lambda\beta}$ ,  $S_{\beta\lambda}$  e  $V_{\lambda\beta}$ , mais detalhes em Bonat and Jørgensen (2016). As matrizes de sensibilidade e variabilidade conjuntas de  $\psi_\beta$  e  $\psi_\lambda$  são denotados por

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

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

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

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

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

$$\begin{aligned} \beta^{(i+1)} &= \beta^{(i)} - S_\beta^{-1} \psi_\beta(\beta^{(i)}, \lambda^{(i)}), \\ \lambda^{(i+1)} &= \lambda^{(i)} - S_\lambda^{-1} \psi_\lambda(\beta^{(i+1)}, \lambda^{(i)}). \end{aligned}$$

### 3. Teste Wald para McGLMs

Seguindo as ideias de REF MEU ARTIGO, considere  $\theta^*$  o vetor  $h \times 1$  de parâmetros desconsiderando os parâmetros de correlação, ou seja,  $\theta^*$  refere-se apenas a parâmetros de regressão, dispersão ou potência. As estimativas dos parâmetros de  $\theta^*$  são dadas por  $\hat{\theta}^*$ .

De maneira similar, considere  $J^{*-1}$  a inversa da matriz de informação de Godambe descon- siderando os parâmetros de correlação, de dimensão  $h \times h$ . Seja  $\mathbf{L}$  uma matriz de especificação de hipóteses a serem testadas, de dimensão  $s \times h$  e  $\mathbf{c}$  um vetor de dimensão  $s \times 1$  com os valores sob hipótese nula, em que  $s$  denota o número de restrições. As hipóteses a serem testadas podem ser escritas como:

$$H_0 : \mathbf{L}\boldsymbol{\theta}^* = \mathbf{c} \text{ vs } H_1 : \mathbf{L}\boldsymbol{\theta}^* \neq \mathbf{c}. \quad (1)$$

Desta forma, a generalização da estatística do teste Wald para verificar a validade de uma hipótese sobre parâmetros de um McGLM fica dada por:

$$W = (\mathbf{L}\hat{\boldsymbol{\theta}}^* - \mathbf{c})^T (\mathbf{L} \mathbf{J}^{*-1} \mathbf{L}^T)^{-1} (\mathbf{L}\hat{\boldsymbol{\theta}}^* - \mathbf{c}),$$

em que  $W \sim \chi_s^2$ , ou seja, independente do número de parâmetros nas hipóteses, a estatística de teste  $W$  é um único valor que segue assintoticamente distribuição  $\chi^2$  com graus de liberdade dados pelo número de restrições, isto é, o número de linhas da matriz  $\mathbf{L}$ , denotado por  $s$ .

Em geral, cada coluna da matriz  $\mathbf{L}$  corresponde a um dos  $h$  parâmetros de  $\boldsymbol{\theta}^*$  e cada linha a uma restrição. Sua construção consiste basicamente em preencher a matriz com 0, 1 e eventualmente -1 de tal modo que o produto  $\mathbf{L}\boldsymbol{\theta}^*$  represente corretamente as hipóteses de interesse. A correta especificação de  $\mathbf{L}$  permite testar qualquer parâmetro individualmente ou até mesmo formular hipóteses para diversos parâmetros.

**REF MEU ARTIGO** apresenta exemplos de como testar diferentes tipos de hipóteses de interesse que surgem em contextos práticos. Apresentaremos neste trabalho dois destes exemplos: hipóteses para múltiplos parâmetros e hipóteses sobre parâmetros de regressão ou dispersão para respostas sob mesmo preditor.

Para fins de ilustração, considere a situação em que deseja-se investigar se uma variável numérica  $X_1$  possui efeito sobre duas variáveis respostas, denotadas por  $Y_1$  e  $Y_2$ . Para tal tarefa coletou-se uma amostra com  $N$  observações e para cada observação registrou-se os valores de  $X_1$ ,  $Y_1$  e  $Y_2$ . Com base nos dados coletados ajustou-se um McGLM bivariado, com preditor dado por:

$$g_r(\mu_r) = \beta_{r0} + \beta_{r1}X_1, r = 1, 2, \quad (2)$$

em que o índice  $r$  denota a variável resposta,  $r = 1, 2$ ;  $\beta_{r0}$  representa o intercepto;  $\beta_{r1}$  um parâmetro de regressão associado a uma variável  $X_1$ . Considere que cada resposta possui apenas um parâmetro de dispersão  $\tau_{r0}$  e que os parâmetros de potência foram fixados. Portanto, trata-se de um problema em que há duas variáveis resposta e apenas uma variável explicativa. Considere que as unidades em estudo são independentes, logo  $Z_0 = I$ .

Suponha que o interesse seja avaliar se existe evidência suficiente para afirmar que há efeito da variável explicativa  $X_1$  em ambas as respostas simultaneamente. Neste caso teremos que testar 2 parâmetros:  $\beta_{11}$ , que associa  $X_1$  à primeira resposta; e  $\beta_{21}$ , que associa  $X_1$  à segunda resposta. Podemos escrever a hipótese da seguinte forma:

$$H_0 : \beta_{r1} = 0 \text{ vs } H_1 : \beta_{r1} \neq 0, \quad (3)$$

ou, de forma equivalente:

$$H_0 : \begin{pmatrix} \beta_{11} \\ \beta_{21} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ vs } H_1 : \begin{pmatrix} \beta_{11} \\ \beta_{21} \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

As hipóteses na forma da [Equation 1](#) possuem os seguintes elementos:

- $\boldsymbol{\theta}^{*T} = [\beta_{10} \ \beta_{11} \ \beta_{20} \ \beta_{21} \ \tau_{11} \ \tau_{21}]$ .
- $\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$ .
- $\mathbf{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ , é o valor sob hipótese nula.

O vetor  $\boldsymbol{\theta}^*$  possui seis elementos e a matriz  $\mathbf{L}$  seis colunas. Neste caso estamos testando dois parâmetros, portanto a matriz  $\mathbf{L}$  possui duas linhas. Essas linhas são compostas por zeros, exceto nas colunas referentes ao parâmetro de interesse. É simples verificar que o produto  $\mathbf{L}\boldsymbol{\theta}^*$  representa a hipótese de interesse inicialmente postulada na [Equation 3](#). Com isso, a distribuição assintótica do teste é  $\chi^2_2$ .

A [Equation 2](#) descreve um modelo bivariado genérico. É importante notar que neste exemplo ambas as respostas estão sujeitas ao mesmo preditor. Na prática, quando se trata dos McGLMs, preditores diferentes podem ser especificados entre variáveis respostas. Contudo, nos casos em que as respostas estão sujeitas a preditores idênticos e as hipóteses sobre os parâmetros não se alteram de resposta para resposta, uma especificação alternativa do procedimento é utilizando o produto Kronecker para testar uma mesma hipótese sobre múltiplas respostas tal como utilizado em [Bonat, Petterle, Balbinot, Mansur, and Graf \(2020\)](#).

Suponha que, neste exemplo, as hipóteses de interesse seguem sendo escritas tal como na [Equation 3](#). Contudo, como se trata de um modelo bivariado com mesmo preditor para as duas respostas, a hipótese de interesse é igual entre respostas e envolve apenas parâmetros de regressão, torna-se conveniente escrever a matriz  $\mathbf{L}$  como o produto Kronecker de duas matrizes: uma matriz  $\mathbf{G}$  e uma  $\mathbf{F}$ , ou seja,  $\mathbf{L} = \mathbf{G} \otimes \mathbf{F}$ . Desta forma, a matriz  $\mathbf{G}$  tem dimensão  $R \times R$  e especifica as hipóteses referentes às respostas, já a matriz  $\mathbf{F}$  especifica as hipóteses entre variáveis e tem dimensão  $s' \times h'$ , em que  $s'$  é o número de restrições lineares, ou seja, o número de parâmetros testados para uma única resposta, e  $h'$  é o número total de coeficientes de regressão ou dispersão da resposta. Portanto, a matriz  $\mathbf{L}$  tem dimensão  $(s'R \times h)$ .

Em geral, a matriz  $\mathbf{G}$  é uma matriz identidade de dimensão igual ao número de respostas analisadas no modelo. Enquanto que a matriz  $\mathbf{F}$  equivale a uma matriz  $\mathbf{L}$  caso houvesse apenas uma única resposta no modelo e apenas parâmetros de regressão ou dispersão. Utilizamos o produto Kronecker destas duas matrizes para garantir que a hipótese descrita na matriz  $\mathbf{F}$  seja testada nas  $R$  respostas do modelo.

Assim, considerando que se trata do caso em que se pode reescrever as hipóteses por meio da decomposição da matriz  $\mathbf{L}$ , os elementos do teste ficam dados por:

- $\boldsymbol{\beta}^T = [\beta_{10} \ \beta_{11} \ \beta_{20} \ \beta_{21}]$ : os parâmetros de regressão do modelo.

- $\mathbf{G} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ : matriz identidade com dimensão dada pelo número de respostas.
- $\mathbf{F} = \begin{bmatrix} 0 & 1 \end{bmatrix}$ : equivalente a um  $\mathbf{L}$  para uma única resposta.
- $\mathbf{L} = \mathbf{G} \otimes \mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ : matriz de especificação das hipóteses sobre todas as respostas.
- $\mathbf{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ , é o valor sob hipótese nula.

Deste modo, o produto  $\mathbf{L}\boldsymbol{\beta}$  representa a hipótese de interesse inicialmente postulada. Neste caso, a distribuição assintótica do teste é  $\chi^2_2$ . Esta especificação é bastante conveniente para a geração de quadros de análise de variância e todos os procedimentos são facilmente generalizados quando há interesse em avaliar hipóteses sobre os parâmetros de dispersão.

### 3.1. ANOVA e MANOVA via teste Wald

Com base na proposta de utilização do teste Wald para McGLMs, **REF MEU ARTIGO** propuseram três diferentes procedimentos para geração de quadros de ANOVA e MANOVA para parâmetros de regressão e um procedimento análogo à uma ANOVA e MANOVA para avaliação dos parâmetros de dispersão de um dado modelo. No caso das ANOVAs gera-se um quadro para cada variável resposta. Para as MANOVAs apenas um quadro é gerado, por isso, para que seja possível realizar as MANOVAs, as respostas do modelo devem estar sujeitas ao mesmo preditor.

Para fins de ilustração, considere a situação em que deseja-se investigar se duas variáveis numéricas denotadas por  $X_1$  e  $X_2$  possuem efeito sobre duas variáveis resposta denotadas por  $Y_1$  e  $Y_2$ . Para tal tarefa coletou-se uma amostra com  $N$  observações e para cada observação foram registrados os valores de  $X_1$ ,  $X_2$ ,  $Y_1$  e  $Y_2$ . Com base nos dados coletados ajustou-se um modelo bivariado, com preditor dado por:

$$g_r(\mu_r) = \beta_{r0} + \beta_{r1}X_1 + \beta_{r2}x_2 + \beta_{r3}X_1X_2.$$

em que o índice  $r$  denota a variável resposta,  $r = 1, 2$ ;  $\beta_{r0}$  representa o intercepto;  $\beta_{r1}$  um parâmetro de regressão associado à variável  $X_1$ ,  $\beta_{r2}$  um parâmetro de regressão associado à variável  $X_2$  e  $\beta_{r3}$  um parâmetro de regressão associado à interação entre  $X_1$  e  $X_2$ . Considere que as unidades em estudo são independentes, portanto cada resposta possui apenas um parâmetro de dispersão  $\tau_{r0}$  associado a uma matriz  $\mathbf{Z}_0 = \mathbf{I}$ . Além disso considere que os parâmetros de potência foram fixados.

A análise de variância do tipo II descrita em **REF MEU ARTIGO** testa, em cada linha, se o modelo completo difere do modelo sem uma variável. Caso haja interações no modelo, é testado o modelo completo contra o modelo sem o efeito principal e qualquer efeito de interação que envolva a variável. Deste modo se torna melhor interpretável o efeito daquela variável sobre o modelo completo, isto é, o impacto na qualidade do modelo caso retirássemos determinada variável. Considerando o preditor exemplo, a análise de variância do tipo II faria os seguintes testes:



1. Testa se o intercepto é igual a 0.
2. Testa se os parâmetros referentes a  $X_1$  são iguais a 0. Ou seja, é avaliado o impacto da retirada de  $X_1$  do modelo. Neste caso retira-se a interação pois nela há  $X_1$ .
3. Testa se os parâmetros referentes a  $X_2$  são iguais a 0. Ou seja, é avaliado o impacto da retirada de  $X_2$  do modelo. Neste caso retira-se a interação pois nela há  $X_2$ .
4. Testa se o efeito de interação é 0.

### 3.2. Teste de comparações múltiplas via teste Wald

Quando a ANOVA aponta para efeito significativo de uma variável categórica, costuma ser de interesse avaliar quais dos níveis diferem entre si. Para isso são empregados os testes de comparações múltiplas. Na literatura existem diversos procedimentos para efetuar tais testes, muitos deles descritos em ?.

Tal tipo de situação pode ser avaliada utilizando o teste Wald. Através da correta especificação da matriz  $\mathbf{L}$ , é possível avaliar hipóteses sobre qualquer possível contraste entre os níveis de uma determinada variável categórica. Portanto, é possível usar a estatística de Wald para efetuar também testes de comparações múltiplas.

O procedimento é baseado basicamente em 3 passos. O primeiro deles é obter a matriz de combinações lineares dos parâmetros do modelo que resultam nas médias ajustadas. Com esta matriz é possível gerar a matriz de contrastes, dada pela subtração duas a duas das linhas da matriz de combinações lineares. Por fim, basta selecionar as linhas de interesse desta matriz e usá-las como matriz de especificação de hipóteses do teste Wald, no lugar da matriz  $\mathbf{L}$ .

Por exemplo, suponha que há uma variável resposta  $Y$  sujeita a uma variável explicativa  $X$  de 4 níveis: A, B, C e D. Para avaliar o efeito da variável  $X$ , ajustou-se um modelo dado por:

$$g(\mu) = \beta_0 + \beta_1[X = B] + \beta_2[X = C] + \beta_3[X = D].$$

Nesta parametrização o primeiro nível da variável categórica é mantido como categoria de referência e, para os demais níveis, mede-se a mudança para a categoria de referência; este é o chamado contraste de tratamento. Neste contexto  $\beta_0$  representa a média ajustada do nível A, enquanto que  $\beta_1$  representa a diferença de A para B,  $\beta_2$  representa a diferença de A para C e  $\beta_3$  representa a diferença de A para D. Com esta parametrização é possível obter o valor predito para qualquer uma das categorias de tal modo que se o indivíduo pertencer à categoria A,  $\beta_0$  representa o predito; se o indivíduo pertencer à categoria B,  $\beta_0 + \beta_1$  representa o predito; para a categoria C,  $\beta_0 + \beta_2$  representa o predito e, por fim, para a categoria D,  $\beta_0 + \beta_3$  representa o predito.

Matricialmente, estes resultados podem ser descritos como

$$\mathbf{K}_0 = \begin{matrix} A \\ B \\ C \\ D \end{matrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

Note que o produto  $\mathbf{K}_0\boldsymbol{\beta}$  gera o vetor de preditos para cada nível de  $X$ . Por meio da subtração das linhas da matriz de combinações lineares  $\mathbf{K}_0$  podemos gerar uma matriz de contrastes  $\mathbf{K}_1$

$$\mathbf{K}_1 = \begin{matrix} A - B \\ A - C \\ A - D \\ B - C \\ B - D \\ C - D \end{matrix} \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

Para proceder um teste de comparações múltiplas basta selecionar os contrastes desejados nas linhas da matriz  $\mathbf{K}_1$  e utilizar estas linhas como matriz de especificação de hipóteses do teste Wald. Por fim, como usual em testes de comparações múltiplas, é recomendada a correção dos valores-p por meio da correção de Bonferroni.

Para efetuação deste procedimento para os McGLMs devemos lembrar que trata-se de uma classe de modelos multivariados. E tal como ocorre no caso das análises de variância, para os testes de comparações múltiplas existem duas possibilidades: testes para uma única resposta e testes para múltiplas respostas.

Na prática, se o interesse for um teste de comparações múltiplas multivariado, existe a necessidade de todas as respostas estarem sujeitas a um mesmo preditor e basta expandir a matriz de contrastes utilizando o produto Kronecker. No caso de um teste de comparações múltiplas para cada resposta, basta selecionar o vetor de estimativas e a partição correspondente ao vetor da matriz  $J_{\theta}^{-1}$  para a resposta específica e proceder com o teste.

## 4. Implementation

Todas as funções implementadas geram resultados mostrando graus de liberdade e valores-p baseados no teste Wald aplicado a um McGLM. A [Table 1](#) mostra os nomes e uma breve descrição das funções implementadas.

Função	Descrição
<code>mc_anova_I()</code>	ANOVA tipo I
<code>mc_anova_II()</code>	ANOVA tipo II
<code>mc_anova_III()</code>	ANOVA tipo III
<code>mc_manova_I()</code>	MANOVA tipo I
<code>mc_manova_II()</code>	MANOVA tipo II
<code>mc_manova_III()</code>	MANOVA tipo III
<code>mc_anova_dispersion()</code>	ANOVA tipo III para dispersão
<code>mc_manova_dispersion()</code>	MANOVA tipo III para dispersão
<code>mc_multcomp()</code>	Testes de comparações múltiplas por resposta
<code>mc_mult_multcomp()</code>	Testes de comparações múltiplas multivariado
<code>mc_linear_hypothesis()</code>	Hipóteses lineares gerais especificadas pelo usuário

Table 1: Funções implementadas

As funções `mc_anova_I()`, `mc_anova_II()` e `mc_anova_III()` são funções destinadas à avali-

ação dos parâmetros de regressão do modelo; elas geram quadros de análise de variância por resposta para um modelo *mcglm*. As funções *mc\_manova\_I()*, *mc\_manova\_II()* e *mc\_manova\_III()* também são funções destinadas à avaliação dos parâmetros de regressão do modelo; elas geram quadros de análise de variância multivariada para um modelo *mcglm*. Enquanto as funções de análise de variância univariadas visam avaliar o efeito das variáveis para cada resposta, as multivariadas visam avaliar o efeito das variáveis explicativas em todas as variáveis resposta simultaneamente. As nomenclaturas seguem o que foi exposto no ?? e as funções recebem como argumento apenas o objeto que armazena o modelo devidamente ajustado.

Tal como descrito no ??, a matriz  $\Omega(\tau)$  tem como objetivo modelar a correlação existente entre linhas do conjunto de dados por meio do chamado preditor linear matricial. Na prática temos, para cada matriz do preditor matricial, um parâmetro de dispersão  $\tau_d$ . De modo análogo ao que é feito para o preditor de média, podemos usar estes parâmetros para avaliar o efeito das unidades correlacionadas no estudo. Neste sentido implementamos as funções *mc\_anova\_dispersion()* e *mc\_manova\_dispersion()*.

A função *mc\_anova\_dispersion()* efetua uma análise de variância do tipo III para os parâmetros de dispersão do modelo. Tal como as demais funções com prefixo *mc\_anova*, é gerado um quadro para cada variável resposta, isto é, nos casos mais gerais avaliamos se há evidência que nos permita afirmar que determinado parâmetro de dispersão é igual a 0, ou seja, se existe efeito das medidas correlacionadas tal como especificado no preditor matricial para aquela resposta. A função recebe como argumento o objeto em que está armazenado o modelo, uma lista de índices indicando de que forma os parâmetros dispersão devem ser testados para cada resposta, de tal modo que parâmetros de dispersão que devem ser testados juntos compartilhem o mesmo índice; o último argumento são os nomes a serem mostrados no quadro final.

Já a função *mc\_manova\_dispersion()* pode ser utilizada em um modelo multivariado em que os preditores matriciais são iguais para todas as respostas e há o interesse em avaliar se o efeito das medidas correlacionadas é o mesmo para todas as respostas. Esta função recebe como argumento o objeto em que está armazenado o modelo, um vetor de índices indicando de que forma os parâmetros dispersão devem ser testados, de tal modo que parâmetros de dispersão que devem ser testados juntos compartilhem o mesmo índice; o último argumento são os nomes a serem mostrados no quadro final.

Para testes de comparações múltiplas foram implementadas as funções *mc\_multcomp()* e *mc\_mult\_multcomp()*. Estas funções devem ser usadas como complemento às funções de análise de variância e análise de variância multivariada quando estas apontam para efeito significativo de variáveis explicativas categóricas. As funções para comparações múltiplas são usadas para realizar comparações duas a duas e identificar quais níveis diferem entre si. Estas funções recebem como argumento o modelo, a variável ou variáveis em que há interesse em avaliar comparações entre níveis e também os dados usados para ajustar o modelo.

Por fim, a função *mc\_linear\_hypothesis()* é a implementação computacional em R que permite a execução de qualquer um dos testes apresentados no ??. É a função mais flexível que temos no conjunto de implementações. Com ela é possível especificar qualquer tipo de hipótese sobre parâmetros de regressão, dispersão ou potência de um modelo *mcglm*. Também é possível especificar hipóteses sobre múltiplos parâmetros e o vetor de valores da hipótese nula é definido pelo usuário. Esta função recebe como argumentos o modelo e um vetor contendo

os parâmetros que devem ser testados e os valores sob hipótese nula. Com algum trabalho, por meio da função de hipóteses lineares gerais, é possível replicar os resultados obtidos pelas funções de análise de variância.

## 5. Examples

### 5.1. Instalação

O pacote *mcglm* está disponível no Comprehensive R Archive Network (CRAN) em <https://CRAN.R-project.org/package=mcglm> e pode ser instalado por meio da função *install.packages()*.

```
install.packages("mcglm")  
library(mcglm)
```

As implementações referentes a este trabalho estão disponíveis publicamente na plataforma github em <https://github.com/lineu96/htmglm> e podem ser instaladas por meio da função *install\_github* do pacote *devtools*.

```
library(devtools)  
install_github("lineu96/htmglm")  
library(htmglm)
```

Nesta seção fornecemos exemplos práticos de utilização das funções implementadas com base em modelos multivariados ajustados com o pacote *mcglm*.

### 5.2. Exemplo 1: soya

Os dados são de um experimento feito em uma casa de vegetação com soja. O delineamento experimental conta com duas plantas por parcela em que cada unidade foi submetida a diferentes combinações de água e adubo. Existem três níveis de um fator correspondente à quantidade de água no solo (*water*) e cinco níveis de adubação com potássio (*pot*). Além disso as parcelas foram dispostas em cinco blocos (*block*). Três variáveis resposta foram avaliadas: a produtividade de grãos (*grain*), número de sementes (*seeds*) e número de ervilhas viáveis por planta (*viablepeas*).

Trata-se de um conjunto de dados interessante para exemplificar o uso das funções implementadas pois existem três variáveis resposta de tipos distintos: a produtividade de grãos é uma variável contínua, o número de sementes é uma contagem, e o número de ervilhas viáveis por planta é um exemplo de variável binomial. O conjunto de dados está disponível no pacote *mcglm*.

```
data("soya", package = "mcglm")
```

O objetivo da análise é avaliar o efeito de adubação e água sobre as três variáveis resposta de interesse. Para fins de análise consideramos como variáveis explicativas os níveis de água, adubação e também as interações entre estes dois fatores. Adicionalmente, o efeito de bloco foi acrescentado aos preditores.

Para ajustar o modelo o primeiro passo é especificar os preditores lineares.

```
form.grain <- grain ~ block + water * pot
form.seed <- seeds ~ block + water * pot

soya$viabilepeasP <- soya$viabilepeas / soya$totalpeas
form.peas <- viablepeasP ~ block + water * pot
```

O segundo passo é especificar as matrizes do preditor linear matricial. Consideramos neste caso que as observações são independentes, por isso incluímos apenas uma matriz identidade.

```
Z0 <- mc_id(soya)
```

Com os elementos definidos, podemos ajustar o modelo. Por meio da função `mcglm()` especificamos os preditores lineares para média, as matrizes dos preditores matriciais, as funções de ligação, de variância, o número de tentativas para a variável binomial e se temos interesse em estimar ou não os parâmetros de potência. Para mais detalhes sobre especificação de preditores e ajuste de McGLMs, consulte [Bonat and Jørgensen \(2016\)](#) e [Bonat \(2018\)](#).

```
fit_joint <- mcglm(linear_pred = c(form.grain,
                                   form.seed,
                                   form.peas),
                  matrix_pred = list(c(Z0),
                                     c(Z0),
                                     c(Z0)),
                  link = c("identity",
                           "log",
                           "logit"),
                  variance = c("constant",
                               "tweedie",
                               "binomialP"),
                  Ntrial = list(NULL,
                                NULL,
                                soya$totalpeas),
                  power_fixed = c(T,T,T),
                  data = soya)
```

Para avaliar alguns resultados do modelo é possível utilizar a função `summary()` que retorna a fórmula dos preditores lineares, as funções de ligação, de variância, de covariância especificadas para ajustar o modelo, as estimativas dos parâmetros de regressão e dispersão bem como os erros padrões.

### *Quadros de análise de variância para parâmetros de regressão*

Com o modelo ajustado podemos aplicar as funções implementadas para avaliar os parâmetros de regressão e dispersão do modelo. As funções de análise de variância dependem apenas do objeto que contém o modelo ajustado e retornam um quadro para cada resposta.

ANOVA tipo I

```
mc_anova_I(fit_joint)

## ANOVA type I using Wald statistic for fixed effects
##
## Call: grain ~ block + water * pot
##
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept 19 6283.6472    0e+00
## 2   block   18  419.6702    0e+00
## 3   water   14  405.1498    0e+00
## 4    pot    12  350.9316    0e+00
## 5 water:pot   8   30.4494    2e-04
##
## Call: seeds ~ block + water * pot
##
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept 19 127429.2620    0.0000
## 2   block   18   205.8174    0.0000
## 3   water   14   194.0161    0.0000
## 4    pot    12   130.2022    0.0000
## 5 water:pot   8    12.7366    0.1212
##
## Call: viablepeasP ~ block + water * pot
##
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept 19  971.1096    0.0000
## 2   block   18  300.2990    0.0000
## 3   water   14  297.4306    0.0000
## 4    pot    12  295.2420    0.0000
## 5 water:pot   8   20.0549    0.0101
```

## ANOVA tipo II

```
mc_anova_II(fit_joint)

## ANOVA type II using Wald statistic for fixed effects
##
## Call: grain ~ block + water * pot
##
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept   1 102.2961    0.0000
## 2   block     4  14.3051    0.0064
## 3   water    10  84.6677    0.0000
## 4    pot     12 350.9316    0.0000
## 5 water:pot   8  30.4494    0.0002
##
```

```
## Call: seeds ~ block + water * pot
##
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept   1 3993.9442  0.0000
## 2    block    4  11.6363  0.0203
## 3    water   10  70.8041  0.0000
## 4     pot    12 130.2022  0.0000
## 5 water:pot   8  12.7366  0.1212
##
## Call: viablepeasP ~ block + water * pot
##
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept   1  13.4353  0.0002
## 2    block    4   4.4305  0.3509
## 3    water   10 33.9928  0.0002
## 4     pot    12 295.2420  0.0000
## 5 water:pot   8  20.0549  0.0101
```

### ANOVA tipo III

```
mc_anova_III(fit_joint)

## ANOVA type III using Wald statistic for fixed effects
##
## Call: grain ~ block + water * pot
##
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept   1 102.2961  0.0000
## 2    block    4  14.3051  0.0064
## 3    water    2   2.3991  0.3013
## 4     pot     4  64.0038  0.0000
## 5 water:pot   8  30.4494  0.0002
##
## Call: seeds ~ block + water * pot
##
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept   1 3993.9442  0.0000
## 2    block    4  11.6363  0.0203
## 3    water    2   3.9399  0.1395
## 4     pot     4  19.1997  0.0007
## 5 water:pot   8  12.7366  0.1212
##
## Call: viablepeasP ~ block + water * pot
##
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept   1  13.4353  0.0002
```

```
## 2      block  4  4.4305  0.3509
## 3      water  2  5.2513  0.0724
## 4       pot  4 71.1026  0.0000
## 5 water:pot  8 20.0549  0.0101
```

De forma similar, as funções de análise de variância multivariadas também dependem apenas do modelo ajustado. É importante notar que para fins práticos as funções de análise de variância multivariada necessitam que os preditores para todas as respostas sejam os mesmos. MANOVA tipo I

```
mc_manova_I(fit_joint)

## MANOVA type I using Wald statistic for fixed effects
##
## Call: ~ block+water*pot
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept 57 168255.3139      0
## 2   block  54   816.7633      0
## 3   water  42   794.0601      0
## 4    pot  36   708.8164      0
## 5 water:pot 24    68.7879      0
```

MANOVA tipo II

```
mc_manova_II(fit_joint)

## MANOVA type II using Wald statistic for fixed effects
##
## Call: ~ block+water*pot
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept  3 5553.7954  0.000
## 2   block  12   23.7478  0.022
## 3   water  30  160.9564  0.000
## 4    pot  36  708.8164  0.000
## 5 water:pot 24    68.7879  0.000
```

MANOVA tipo III

```
mc_manova_III(fit_joint)

## MANOVA type III using Wald statistic for fixed effects
##
## Call: ~ block+water*pot
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept  3 5553.7954  0.0000
```



```
## 2    block 12    23.7478    0.0220
## 3    water  6     9.0173    0.1726
## 4     pot 12   149.0321    0.0000
## 5 water:pot 24    68.7879    0.0000
```

Para hipóteses lineares gerais sobre parâmetros de regressão basta especificar o modelo e a hipótese a ser testada. Para identificar os parâmetros de interesse, utilize a função `coef()`.

Teste sobre um único parâmetro de regressão

```
mc_linear_hypothesis(object = fit_joint,
                      hypothesis = c('beta11 = 0'))

## Linear hypothesis test
##
## Hypothesis:
## 1 beta11 = 0
##
## Results:
##   Df    Chi Pr(>Chi)
## 1  1  1.2362  0.2662
```

Teste sobre mais de um parâmetro de regressão

```
mc_linear_hypothesis(object = fit_joint,
                      hypothesis = c('beta11 = 0',
                                     'beta12 = 0'))

## Linear hypothesis test
##
## Hypothesis:
## 1 beta11 = 0
## 2 beta12 = 0
##
## Results:
##   Df    Chi Pr(>Chi)
## 1  2  3.5639  0.1683
```

Teste de igualdade de efeitos entre parâmetros de regressão

```
mc_linear_hypothesis(object = fit_joint,
                      hypothesis = c('beta11 = beta21'))

## Linear hypothesis test
##
## Hypothesis:
```

```
## 1 beta11 = beta21
##
## Results:
##   Df    Chi Pr(>Chi)
## 1   1 1.3491 0.2454
```

### 5.3. Exemplo 2: Hunting

O conjunto de dados Hunting, apresentados em ?, também está disponível no pacote *mcglm*. Os dados tratam de um problema em que as respostas são contagens bivariadas longitudinais sobre animais caçados na vila de Basile Fang, Bioko Norte Province, Bioko Island, Equatorial Guinea. As variáveis respostas são: números mensais de blue duikers (BD) e outros pequenos animais (OT) baleados ou capturados em uma amostra aleatória de 52 caçadores comerciais de agosto de 2010 a setembro de 2013. Consideremos que o interesse é avaliar o efeito de um fator com 2 níveis que indica se o animal foi caçado por meio de arma de fogo ou armadilha (*METHOD*) e um fator com 2 níveis que indica o sexo do animal (*SEX*).

```
data("Hunting", package = "mcglm")
```

Tal como no primeiro exemplo, para ajuste do modelo é necessário definir os preditores lineares para média, as matrizes dos preditores matriciais, as funções de ligação, de variância, se temos interesse em estimar ou não os parâmetros de potência. Para esta análise consideramos no preditor matricial a estrutura de medidas repetidas introduzidas pelas observações tomadas para o mesmo caçador e mês (HUNTER.MONTH) e o número de dias de caça por mês foi usado como termo offset.

```
form.OT <- OT ~ METHOD * SEX
form.BD <- BD ~ METHOD * SEX

Z0 <- mc_id(Hunting)
Z1 <- mc_mixed(~ 0 + HUNTER.MONTH, data = Hunting)

fit <- mcglm(linear_pred = c(form.BD, form.OT),
             matrix_pred = list(c(Z0, Z1),
                                c(Z0, Z1)),
             link = c("log", "log"),
             variance = c("poisson_tweedie",
                           "poisson_tweedie"),
             offset = list(log(Hunting$OFFSET),
                           log(Hunting$OFFSET)),
             data = Hunting)
```

Novamente, para avaliar alguns resultados do modelo é possível utilizar a função *summary()*. Podemos também aplicar as já apresentadas funções implementadas para ANOVAs, MANOVAs e testes de hipóteses lineares gerais sobre os parâmetros de regressão e dispersão do modelo.

Neste caso, como existe um preditor matricial especificado, pode ser de interesse um estudo aprofundado dos parâmetros de dispersão. Esta análise pode ser feita com a já utilizada função `mc_linear_hypothesis()`.

Teste sobre um único parâmetro de dispersão

```
mc_linear_hypothesis(object = fit,
                      hypothesis = c('tau11 = 0'))

## Linear hypothesis test
##
## Hypothesis:
## 1 tau11 = 0
##
## Results:
##   Df      Chi Pr(>Chi)
## 1   1 22.5613      0
```

Teste sobre mais de um parâmetro de dispersão

```
mc_linear_hypothesis(object = fit,
                      hypothesis = c('tau11 = 0',
                                     'tau21 = 0'))

## Linear hypothesis test
##
## Hypothesis:
## 1 tau11 = 0
## 2 tau21 = 0
##
## Results:
##   Df      Chi Pr(>Chi)
## 1   2 29.098      0
```

Teste de igualdade de efeitos entre parâmetros de dispersão

```
mc_linear_hypothesis(object = fit,
                      hypothesis = c('tau12 = tau21'))

## Linear hypothesis test
##
## Hypothesis:
## 1 tau12 = tau21
##
## Results:
##   Df      Chi Pr(>Chi)
## 1   1 97.0998      0
```

### Quadro de análise de variância para parâmetros de dispersão

As funções para avaliar os parâmetros de dispersão por meio de um procedimento análogo à análise de variância para parâmetros de regressão, requerem a especificação de mais argumentos: um deles que determina a relação entre parâmetros de dispersão e o outro que especifica os nomes que aparecerão na saída final.

#### ANOVA tipo III para dispersão

```
mc_anova_dispersion(fit,
                     p_var = list(c(0,1), c(0,1)),
                     names = list(c('tau10', 'tau11'),
                                   c('tau20', 'tau21')))
```

```
## ANOVA type III using Wald statistic for dispersion parameters
##
## Call: BD ~ METHOD * SEX
##
##   Dispersion Df      Chi Pr(>Chi)
## 1      tau10  1 22.5613         0
## 2      tau11  1 97.0998         0
##
## Call: OT ~ METHOD * SEX
##
##   Dispersion Df      Chi Pr(>Chi)
## 1      tau20  1  7.2008  0.0073
## 2      tau21  1 29.0133  0.0000
```

#### MANOVA tipo III para dispersão

```
mc_manova_dispersion(fit,
                      p_var = c(0,1),
                      names = c('tau0', 'tau1'))
```

```
## MANOVA type III using Wald statistic for dispersion parameters
##
## Call: ~ METHOD*SEX
##   Covariate Df      Chi Pr(>Chi)
## 1      tau0  2 29.0980         0
## 2      tau1  2 124.2049         0
```

### Comparações múltiplas

Por fim, podemos utilizar as funções para testes de comparações múltiplas para avaliar diferenças existentes entre níveis de variáveis explicativas categóricas incluídas no modelo. Esta tarefa pode ser feita por variável resposta:

```
mc_multcomp(object = fit,
             effect = list(c('METHOD', 'SEX'),
                           c('METHOD', 'SEX')),
             data = Hunting)

## Multiple comparisons test for each outcome using Wald statistic
##
## Call: BD ~ METHOD * SEX
##
##
```

	Contrast	Df	Chi	Pr(>Chi)
## 1	Escopeta:Female-Escopeta:Male	1	175.7657	0
## 2	Escopeta:Female-Trampa:Female	1	20.1379	0
## 3	Escopeta:Female-Trampa:Male	1	35.6372	0
## 4	Escopeta:Male-Trampa:Male	1	24.3946	0
## 5	Trampa:Female-Escopeta:Male	1	217.7398	0
## 6	Trampa:Female-Trampa:Male	1	132.6125	0

```
##
## Call: OT ~ METHOD * SEX
##
##
```

	Contrast	Df	Chi	Pr(>Chi)
## 1	Escopeta:Female-Escopeta:Male	1	14.3969	0.0009
## 2	Escopeta:Female-Trampa:Female	1	6.5843	0.0617
## 3	Escopeta:Female-Trampa:Male	1	5.6455	0.1050
## 4	Escopeta:Male-Trampa:Male	1	0.7480	1.0000
## 5	Trampa:Female-Escopeta:Male	1	31.3069	0.0000
## 6	Trampa:Female-Trampa:Male	1	25.3203	0.0000

Já no caso de preditores iguais para todas as respostas é possível realizar um teste de comparações múltiplas multivariado.

```
mc_mult_multcomp(object = fit,
                  effect = c('METHOD', 'SEX'),
                  data = Hunting)

## Multivariate multiple comparisons test using Wald statistic
##
## Call: ~ METHOD*SEX
##
```

	Contrast	Df	Chi	Pr(>Chi)
## 1	Escopeta:Female-Escopeta:Male	2	215.0490	0
## 2	Escopeta:Female-Trampa:Female	2	31.8503	0
## 3	Escopeta:Female-Trampa:Male	2	47.8804	0
## 4	Escopeta:Male-Trampa:Male	2	27.5459	0
## 5	Trampa:Female-Escopeta:Male	2	287.6161	0
## 6	Trampa:Female-Trampa:Male	2	184.8844	0

## 6. Concluding remarks

This article described the R implementation of procedimentos para realizar testes de hipóteses sobre parâmetros de McGLMs baseados na estatística de Wald. McGLMs contam com parâmetros de regressão, dispersão, potência e correlação; cada conjunto de parâmetros possui uma interpretação prática bastante relevante no contexto de análise de problemas com potenciais múltiplas respostas em função de um conjunto de variáveis explicativas.

Com base na proposta de utilização do teste Wald para McGLMs, desenvolvemos procedimentos para testes de hipóteses lineares gerais, geração de quadros de ANOVA e MANOVA para parâmetros de regressão e dispersão e também testes de comparações múltiplas. Todos estes procedimentos foram implementados na linguagem R e complementam as funcionalidades existentes na biblioteca *mcglm*.

The discussed examples illustrate the...

Possíveis extensões deste trabalho que seguem na linha de avaliação de parâmetros de McGLMs para um melhor entendimento do impacto dos elementos em problemas de modelagem são: explorar correções de valores-p de acordo com o tamanho das hipóteses testadas, explorar procedimentos além do teste Wald (como o teste Escore e o teste da razão de verossimilhanças), implementar novos procedimentos para comparações múltiplas, adaptar a proposta para lidar com contrastes alternativos aos usuais, explorar procedimentos para seleção automática de covariáveis (backward elimination, forward selection, stepwise selection) e também seleção de covariáveis por meio de inclusão de penalização no ajuste por complexidade (similar a ideia de regressão por splines).

## Acknowledgments

The authors thank the reviewers for their constructive and helpful comments, which greatly improved the article. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

## References

- Anderson T, *et al.* (1973). “Asymptotically efficient estimation of covariance matrices with linear structure.” *The Annals of Statistics*, **1**(1), 135–141.
- Bonat WH (2018). “Multiple Response Variables Regression Models in R: The *mcglm* Package.” *Journal of Statistical Software*, **84**(4), 1–30. doi:10.18637/jss.v084.i04.
- Bonat WH, Jørgensen B (2016). “Multivariate covariance generalized linear models.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **65**(5), 649–675.
- Bonat WH, Petterle RR, Balbinot P, Mansur A, Graf R (2020). “Modelling multiple outcomes in repeated measures studies: Comparing aesthetic eyelid surgery techniques.” *Statistical Modelling*, p. 1471082X20943312.
- Demidenko E (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.

- Fox J, Weisberg S (2019). *An R Companion to Applied Regression*. Third edition. Sage, Thousand Oaks CA. URL <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.
- Hothorn T, Bretz F, Westfall P (2008). “Simultaneous Inference in General Parametric Models.” *Biometrical Journal*, **50**(3), 346–363.
- 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, Knudsen SJ (2004). “Parameter orthogonality and bias adjustment for estimating functions.” *Scandinavian Journal of Statistics*, **31**(1), 93–114.
- Jørgensen B, Kokonendji CC (2015). “Discrete dispersion models and their Tweedie asymptotics.” *AStA Advances in Statistical Analysis*, **100**(1), 43–78.
- Liang KY, Zeger SL (1986). “Longitudinal data analysis using generalized linear models.” *Biometrika*, **73**(1), 13–22.
- Lumley T (2004). “Analysis of Complex Survey Samples.” *Journal of Statistical Software*, **9**(1), 1–19. R package version 2.2.
- Lumley T (2010). *Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R*. John Wiley and Sons.
- Lumley T (2020). “survey: analysis of complex survey samples.” R package version 4.0.
- Martinez-Beneito MA (2013). “A general modelling framework for multivariate disease mapping.” *Biometrika*, **100**(3), 539–553.
- Pinheiro JC, Bates DM (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 (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Zeileis A, Hothorn T (2002). “Diagnostic Checking in Regression Relationships.” *R News*, **2**(3), 7–10. URL <https://CRAN.R-project.org/doc/Rnews/>.

**Affiliation:**

Lineu Alberto Cavazani de Freitas  
Department of Informatics  
Paraná Federal University  
Centro Politécnico  
Curitiba 81531980, CP 19081, Paraná, Brazil.  
E-mail: [lineuacf@gmail.com](mailto:lineuacf@gmail.com)