Hypothesis tests for multivariate regression with non-Gaussian data in **R**: The htmcglm Package

Lineu Alberto Cavazani de Freitas 🆔
Paraná Federal University

Wagner Hugo Bonat 🆔
Paraná Federal University

Abstract

This article describes the R package htmcglm implemented for performing hypothesis tests on regression and dispersion parameters of multivariate covariance generalized linear models (McGLMs) fiited by package mcglm. Our implementations are based on the proposal to use the Wald test to perform general hypothesis tests, ANOVAs, MANOVAs and multiple comparison tests. McGLMs provide a general statistical modeling framework for normal and non-normal multivariate data analysis along with a wide range of correlation structures. Our contribution is aimed at providing tools for a better interpretation of the estimated parameters and thus extracting more information and conclusions about the problems modeled through the class. By studying the regression parameters, it is possible to assess the effect of the explanatory variable(s) on the response(s). By studying the dispersion parameters, the effect of the correlation between study units can be evaluated, very useful in situations where the observations of the data set are correlated with each other, such as in longitudinal, temporal and repeated measures studies. Our implementations provide a procedural and safe way to answer common questions in the context of multivariate modeling, such as: which variables are associated with the outcome of the phenomenon of interest, if there is an effect of the correlation structure between individuals in the study, if the effect of a certain variable is the same regardless of the response, among others. Illustrations in this article cover the most common hypothesis tests of interest in practical contexts in modeling multivariate regression problems.

Keywords: multivariate regression models, McGLM, hypothesis tests, Wald test, ANOVA, MANOVA, Multiple comparisons, R.

1. Introduction

The **htmcglm** package for R (R Core Team 2022) provides functions for performing hypothesis testing on parameters of multivariate covariance generalized linear models (McGLMs; Bonat

and Jørgensen (2016)) fitted using the **mcglm** package (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. When we work on the McGLMs class we estimate parameters of regression, dispersion, power and correlation. Each set of parameters has a very useful practical interpretation.

By studying the regression parameters, it is possible to assess the effect of the explanatory variable(s) on the response(s). By studying the dispersion parameters, the effect of the correlation between study units can be evaluated, very useful in situations where the observations of the data set are correlated with each other, such as in longitudinal, temporal and repeated measures studies. The power parameters give us an indication of which probability distribution best fits the problem. And correlation parameters estimate the strength of association between responses in a multivariate problem.

The development of hypothesis tests for the purpose of evaluating these quantities is of great value in practical problems and leads to procedural forms for evaluating the resulting quantities of the model. The **htmcglm** package is a full R implementation with functions based on Wald statistics to evaluate regression and dispersion parameters. The features include functions for general linear hypothesis testing, univariate and multivariate analysis of variance tables, as well as multiple comparison tests.

Package **htmcglm** is available on github at https://github.com/lineu96/htmcglm and complement the functions available in the package **mcglm**.

There are several implementations of the Wald test in different contexts in R. The package **lmtest** (Zeileis and Hothorn 2002) has a generic function to perform Wald tests to compare nested linear and generalized linear models. The package **survey** (Lumley 2020, 2004, 2010) has a function that performs Wald tests that, by default, tests whether all coefficients associated with a given regression term are zero, but it is possible to specify hypotheses with other values.

The package **car** (Fox and Weisberg 2019) has an implementation to test linear hypotheses about parameters of linear models, generalized linear models, multivariate linear models, mixed effects models, among others; in this implementation, the user has full control of which parameters to test and with which values to compare in the null hypothesis.

For analysis of variance tables, R has the function `anova()` in the standard package **stats** (R Core Team 2022) applicable to linear and generalized linear models. The package **car** (Fox and Weisberg 2019) has a function that returns analysis of variance tables of types II and III for different models. For multiple comparisons, one of the main packages available is **multcomp** (Hothorn, Bretz, and Westfall 2008) which provides an interface for testing multiple comparisons for parametric models.

However, when dealing with multivariate covariance generalized linear models fitted in the **mcglm** package, there is only one type of analysis of variance implemented in the library and there are no options for performing general linear hypothesis tests, nor multiple comparison tests. Therefore, as it is a flexible class of models with high power of application to practical problems, our general objective is to provide implementations that allow performing hypothesis tests for McGLMs in such a way that it is possible to test general linear hypotheses, generate analysis of variance tables, multivariate analysis of variance tables and multiple comparisons tests.

The article is organized as follows. In section 2 we present a review of the general structure

and estimation of the parameters of a McGLM, based on the ideas of Bonat and Jørgensen (2016). In section 3 the details of the Wald test to evaluate assumptions about parameters of a McGLM are presented. In section 4 we introduce the R implementation discussing the main functions available in the **htmcglm** package. We illustrate the package usage through some examples in section 5. Finally, section 6 presents a discussion and directions for future work on the improvement of the **htmcglm** package.

2. Multivariate covariance generalized linear models

Consider $\boldsymbol{Y}_{N \times R} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_R\}$ a matrix of response variables and $\boldsymbol{M}_{N \times R} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_R\}$ a matrix of expected values. The variance and covariance matrix for each response r, $r = 1, \ldots, R$, is denoted by Σ_r and has dimension $N \times N$. In addition, consider a $R \times R$ correlation matrix, denoted by Σ_b, to describe the correlation between the response variables. The McGLMs (Bonat and Jørgensen 2016) are defined by:

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{M} = \{g_1^{-1}(\boldsymbol{X}_1\boldsymbol{\beta}_1), \ldots, g_R^{-1}(\boldsymbol{X}_R\boldsymbol{\beta}_R)\}$$

$$\mathrm{Var}(\boldsymbol{Y}) = \boldsymbol{C} = \boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b,$$

where the functions $g_r()$ are the traditional link functions; $\boldsymbol{X}_r$ denotes a design matrix $N \times k_r$; $\boldsymbol{\beta}_r$ denotes a vector $k_r \times 1$ of regression parameters. $\boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b = \mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \ldots, \tilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b \otimes \boldsymbol{I})\mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^\top, \ldots, \tilde{\boldsymbol{\Sigma}}_R^\top)$ is the Generalized Kronecker product (Martinez-Beneito 2013), the matrix $\tilde{\boldsymbol{\Sigma}}_r$ denotes the lower triangular matrix of the Cholesky decomposition of the matrix $\boldsymbol{\Sigma}_r$. the operator Bdiag() denotes the block-diagonal matrix and $\boldsymbol{I}$ is an identity matrix $N \times N$.

For continuous, binary, binomial, proportions or indices, the variance and covariance matrix $\boldsymbol{\Sigma}_r$ is given by:

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

In the case of response variables that are counts, the variance and covariance matrix for each response variable is given by:

$$\Sigma_r = diag(\boldsymbol{\mu}_r) + \mathrm{V}\left(\boldsymbol{\mu}_r; p_r\right)^{1/2}\left(\boldsymbol{\Omega}\left(\boldsymbol{\tau}_r\right)\right)\mathrm{V}\left(\boldsymbol{\mu}_r; p_r\right)^{1/2},$$

where $\mathrm{V}\left(\boldsymbol{\mu}_r; p_r\right) = diag(\vartheta(\boldsymbol{\mu}_r; p_r))$ denotes a diagonal matrix in which the entries are given by the variance function $\vartheta(\cdot; p_r)$ applied to vector elements $\boldsymbol{\mu}_r$. Different choices of variance functions $\vartheta(\cdot; p_r)$ imply different assumptions about the distribution of the response variable. We will mention 3 options of variance functions: the power variance function, the Poisson-Tweedie dispersion function and the binomial variance function.

The power variance function characterizes the Tweedie family of distributions, is given by $\vartheta\left(\cdot; p_r\right) = \mu_r^{p_r}$, in which some distributions stand out: Normal ($p = 0$), Poisson ($p = 1$), gama ($p = 2$) and inverse Gaussian ($p = 3$) (Jørgensen 1987, 1997).

The Poisson-Tweedie dispersion function (Jørgensen and Kokonendji 2015) is indicated for events defined by counts. Is given by $\vartheta\left(\cdot; p\right) = \mu + \tau\mu^p$ where τ is the dispersion parameter. We thus have a rich class of models for dealing with responses that characterize counts, since

many important distributions appear as special cases, such as: Hermite ($p = 0$), Neyman type A ($p = 1$), negative binomial ($p = 2$) and Poisson–inverse Gaussian (p = 3).

Finally, the binomial variance function, given by $\vartheta\left(\cdot; p_r\right) = \mu_r^{p_{r1}}(1 - \mu_r)^{p_{r2}}$ is indicated when the response variable is binary, restricted to an interval or when there is a number of successes in a number of trials.

It is possible to notice that the power parameter p appears in all the variance functions discussed. This parameter is especially important because it is an index that distinguishes different probability distributions that are important in the modeling context and, for this reason, can be used as a tool for automatic selection of the probability distribution that best suits the problem.

The dispersion matrix $\boldsymbol{\Omega}(\boldsymbol{\tau})$ describes the part of the covariance within each response variable that does not depend on the mean structure, that is, the correlation structure between the observations in the sample. Based on the ideas of Anderson *et al.* (1973) and Pourahmadi (2000), Bonat and Jørgensen (2016) proposed to model the dispersion matrix through a matrix linear predictor combined with a covariance link function given by:

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

where $h()$ is the covariance link function, Z_{rd} with $d = 0,\ldots,$ D are matrices that represent the covariance structure present in each response variable r and $\boldsymbol{\tau_r} = (\tau_{r0},\ldots,\tau_{rD})$ is a vector $(D + 1) \times 1$ of dispersion parameters.

Some possible covariance link functions are identity, inverse and exponential-matrix. The specification of the covariance link function is discussed by Pinheiro and Bates (1996) and it is possible to select combinations of matrices to obtain the best known models in the literature for longitudinal data, time series, spatial and spatio-temporal data. Further details are discussed by Demidenko (2013).

In this way, the McGLMs configure a general framework for analysis via regression models for non-Gaussian data with multiple responses, in which no assumptions are made regarding the independence of the observations. The class is defined by 3 functions (link, variance and covariance), in addition to a linear predictor and a matrix linear predictor for each response under analysis.

2.1. Estimation and inference

McGLMs are fitted based on the estimating function approach described in detail by Bonat and Jørgensen (2016) and Jørgensen and Knudsen (2004). This subsection presents an overview of the algorithm and the asymptotic distribution of estimators based on estimating functions.

McGLM's second-moment assumptions allow the division of parameters into two subsets: $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\lambda}^\top)^\top$. This way, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top,\ldots,\boldsymbol{\beta}_R^\top)^\top$ is a vector $K \times 1$ of regression parameters and $\boldsymbol{\lambda} = (\rho_1,\ldots,\rho_{R(R-1)/2}, p_1,\ldots,p_R, \boldsymbol{\tau}_1^\top,\ldots,\boldsymbol{\tau}_R^\top)^\top$ is a vector $Q \times 1$ of dispersion parameters. Furthermore, $\mathcal{Y} = (\boldsymbol{Y}_1^\top,\ldots,\boldsymbol{Y}_R^\top)^\top$ denotes the stacked vector $NR \times 1$ of the matrix of response variables $\boldsymbol{Y}_{N \times R}$ and $\mathcal{M} = (\boldsymbol{\mu}_1^\top,\ldots,\boldsymbol{\mu}_R^\top)^\top$ denotes the stacked vector $NR \times 1$ of the matrix of expected values $\boldsymbol{M}_{N \times R}$.

To estimate the regression parameters, the quasi-score function (Liang and Zeger 1986) is used, represented by

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

where $\boldsymbol{D} = \nabla_\beta \mathcal{M}$ is a matrix $NR \times K$, and ∇_β denotes the gradient operator. Using the quasi-score function the sensitivity matrix $K \times K$ of ψ_β is given by

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

whereas the $K \times K$ variability matrix of ψ_β is written as

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

For the dispersion parameters, the Pearson estimating function is used, defined in the form

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

where $W_{\lambda i} = -\frac{\partial \boldsymbol{C}^{-1}}{\partial \lambda_i}$ and $\boldsymbol{r} = (\mathcal{Y} - \mathcal{M})$. The entry (i, j) of the $Q \times Q$ sensitivity matrix of ψ_λ is given by

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

The entry (i, j) of the variability matrix $Q \times Q$ of ψ_λ is defined by

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

where $k_l^{(4)}$ denotes the fourth cumulant of $\mathcal{Y}_l$. In the McGLM estimation process, empirical versions are used.

To take into account the covariance between the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, Bonat and Jørgensen (2016) obtained the cross-sensitivity and variability matrices, denoted by $S_{\lambda\beta}$, $S_{\beta\lambda}$ and $V_{\lambda\beta}$, more details in Bonat and Jørgensen (2016). The joint sensitivity and variability matrices of ψ_β and ψ_λ are denoted by

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

Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\lambda}}^\top)^\top$ the estimator based on estimating functions of $\boldsymbol{\theta}$. Then, the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is

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

where $J_{\boldsymbol{\theta}}^{-1}$ is the inverse of the Godambe information matrix, given by $J_{\boldsymbol{\theta}}^{-1} = S_{\boldsymbol{\theta}}^{-1} V_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}^{-\top}$, where $S_{\boldsymbol{\theta}}^{-\top} = (S_{\boldsymbol{\theta}}^{-1})^\top$.

To solve the system of equations $\psi_\beta = 0$ and $\psi_\lambda = 0$ the modified Chaser algorithm is used, proposed by Jørgensen and Knudsen (2004), which is defined as

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - S_{\boldsymbol{\beta}}^{-1}\psi\boldsymbol{\beta}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\lambda}^{(i)}),$$
$$\boldsymbol{\lambda}^{(i+1)} = \boldsymbol{\lambda}^{(i)}\alpha S_{\boldsymbol{\lambda}}^{-1}\psi\boldsymbol{\lambda}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}).$$

3. Wald Test for McGLMs

Following the ideas of **REF MEU ARTIGO**, consider $\boldsymbol{\theta}^*$ the vector $h \times 1$ of parameters disregarding the correlation parameters, ie, $\boldsymbol{\theta}^*$ only refers to regression, dispersion or power parameters. The estimates of the parameters of $\boldsymbol{\theta}^*$ are given by $\hat{\boldsymbol{\theta}}^*$. In a similar way, consider J^{*-1} the inverse of the Godambe information matrix disregarding the correlation parameters, of dimension $h \times h$. Let $\boldsymbol{L}$ a specification matrix of hypotheses to be tested, of dimension $s \times h$ and $\boldsymbol{c}$ a vector of dimension $s \times 1$ with the values under the null hypothesis, where s denotes the number of restrictions. The hypotheses to be tested can be written as:

$$H_0 : \boldsymbol{L}\boldsymbol{\theta}^* = \boldsymbol{c} \; vs \; H_1 : \boldsymbol{L}\boldsymbol{\theta}^* \neq \boldsymbol{c}. \tag{1}$$

In this way, the generalization of the Wald test statistic to verify the validity of a hypothesis about parameters of a McGLM is given by:

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

where $W \sim \chi_s^2$, that is, regardless of the number of parameters in the hypotheses, the test statistic W is a single value that asymptotically follows the χ^2 distribution with degrees of freedom given by the number of constraints, that is, the number of rows in the matrix $\boldsymbol{L}$, denoted by s.

In general, each column of the matrix $\boldsymbol{L}$ corresponds to one of the h parameters of $\boldsymbol{\theta}^*$ and each row to a constraint. Its construction basically consists of filling the matrix with 0, 1 and eventually -1 in such a way that the product $\boldsymbol{L}\boldsymbol{\theta}^*$ correctly represents the hypotheses of interest. The correct specification of $\boldsymbol{L}$ allows testing any parameter individually or even formulating hypotheses for several parameters.

REF MEU ARTIGO presents examples of how to test different types of hypotheses of interest that arise in practical contexts. We will present in this work two of these examples: hypotheses for multiple parameters and hypotheses about regression or dispersion parameters for responses under the same predictor.

For purposes of illustration, consider the situation in which you want to investigate whether a numeric variable X_1 has an effect on two response variables, denoted by Y_1 and Y_2. A bivariate McGLM for this problem may have a predictor given by:

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

where the index r denotes the response variable, r = 1,2; β_{r0} represents the intercept; β_{r1} a regression parameter associated with the variable X_1. Assume that each response has only one dispersion parameter τ_{r0} and that the power parameters were fixed. Therefore, it is a problem in which there are two response variables and only one explanatory variable. Assume that the units under study are independent, so $Z_0 = I$.

Suppose the interest is to assess whether there is sufficient evidence to state that there is an effect of the explanatory variable X_1 on both responses simultaneously. In this case we will have to test 2 parameters: β_{11}, which associates X_1 with the first response; and β_{21}, which associates X_1 with the second response. We can write the hypothesis as follows:

$$H_0 : \beta_{r1} = 0 \ vs \ H_1 : \beta_{r1} \neq 0, \tag{3}$$

or, equivalently:

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

The hypotheses in the form of Equation 1 have the following elements:

- $\boldsymbol{\theta^{*T}} = \begin{bmatrix} \beta_{10} & \beta_{11} & \beta_{20} & \beta_{21} & \tau_{11} & \tau_{21} \end{bmatrix}$.

- $\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$.

- $\boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

The vector $\boldsymbol{\theta^*}$ has six elements and the matrix $\boldsymbol{L}$ has six columns. In this case we are testing two parameters, so the matrix $\boldsymbol{L}$ has two rows. These lines are composed of zeros, except in the columns referring to the parameter of interest. It is simple to verify that the product $\boldsymbol{L\theta^*}$ represents the hypothesis of interest initially postulated in Equation 3. Thus, the asymptotic distribution of the test is χ_2^2.

The Equation 2 describes a generic bivariate model. It is important to note that in this example both responses are subject to the same predictor. In practice, when it comes to McGLMs, different predictors can be specified between response variables. However, in cases where the responses are subject to identical predictors and the hypothesis about the parameters do not change from response to response, an alternative specification of the procedure is to use the Kronecker product to test the same hypothesis on multiple responses as used in Bonat, Petterle, Balbinot, Mansur, and Graf (2020).

Suppose that, in this example, the hypotheses of interest are still written as in the form of Equation 3. However, as this is a bivariate model with the same predictor for the two responses, the hypothesis of interest are the same between responses and involves only regression parameters, it is convenient to write the matrix $\boldsymbol{L}$ as the Kronecker product of two matrices: a matrix $\boldsymbol{G}$ and a matrix $\boldsymbol{F}$, ie, $\boldsymbol{L} = \boldsymbol{G} \otimes \boldsymbol{F}$. In this way, the matrix $\boldsymbol{G}$ has dimension $R \times R$ and specifies the hypotheses about the responses, whereas the matrix $\boldsymbol{F}$ specifies the hypotheses between variables and has dimension $s' \times h'$, where s' is the number of linear constraints, that is, the number of parameters tested for a single response, and h' is the total number of coefficients of regression or dispersion of the response. Therefore, the matrix $\boldsymbol{L}$ has dimension $(s'R \times h)$.

In general, the matrix $\boldsymbol{G}$ is an identity matrix with a dimension equal to the number of responses analyzed in the model. Whereas the matrix $\boldsymbol{F}$ is equivalent to a matrix $\boldsymbol{L}$ if there

was only a single response in the model and only regression or dispersion parameters. We use the Kronecker product of these two matrices to ensure that the hypothesis described in the $\boldsymbol{F}$ matrix will be tested on the R model responses.

Thus, considering that this is the case in which the hypotheses can be rewritten by decomposing the $\boldsymbol{L}$ matrix, the test elements are given by:

- $\boldsymbol{\beta^T} = \begin{bmatrix} \beta_{10} & \beta_{11} & \beta_{20} & \beta_{21} \end{bmatrix}$: the model regression parameters.

- $\boldsymbol{G} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$: identity matrix with dimension given by the number of responses.

- $\boldsymbol{F} = \begin{bmatrix} 0 & 1 \end{bmatrix}$: equivalent to a $\boldsymbol{L}$ for a single response.

- $\boldsymbol{L} = \boldsymbol{G} \otimes \boldsymbol{F} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$: matrix specifying the hypotheses on all responses.

- $\boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$: matrix with the values under the null hypothesis.

Thus, the product $\boldsymbol{L\beta}$ represents the initially postulated hypothesis of interest. In this case, the asymptotic distribution of the test is χ_2^2. This specification is very convenient for generating analysis of variance tables and all procedures are easily generalized when there is interest in evaluating hypotheses about the dispersion parameters.

3.1. ANOVA and MANOVA via Wald test

Based on the proposed use of the Wald test for McGLMs, **REF MEU ARTIGO** proposed three different procedures for generating ANOVA and MANOVA tables for regression parameters, and a procedure similar to ANOVA and MANOVA to evaluate the dispersion parameters of a model. In the case of ANOVAs, a table is generated for each response variable. For MANOVAs only one table is generated, therefore, in order to be able to perform MANOVAs, the responses must be subject to the same predictor.

For purposes of illustration, consider the situation where the objective is to investigate whether two numeric variables denoted by X_1 and X_2 have an effect on two response variables denoted by Y_1 and Y_2. For this case, consider the following predictor:

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

where the index r denotes the response variable, $r = 1, 2$; β_{r0} represents the intercept; β_{r1} a regression parameter associated with the variable X_1, β_{r2} a regression parameter associated with the variable X_2 and β_{r3} a regression parameter associated with the interaction between X_1 and X_2. Assume that the units under study are independent, so each response has only one dispersion parameter τ_{r0} associated with a matrix $Z_0 = I$. Also consider that the power parameters have been fixed.

The type II analysis of variance described in **REF MEU ARTIGO** tests, on each line, whether the complete model differs from the model without a variable. If there are interactions

in the model, the complete model is tested against the model without the main effect and any interaction effect involving the variable. In this way, the effect of that variable on the complete model becomes better interpretable, that is, the impact on the quality of the model if we removed a certain variable. Considering the example predictor, the type II analysis of variance would do the following tests:

1. Tests if the intercept is equal to 0.

2. Tests if the parameters referring to X_1 are equal to 0. That is, the impact of removing X_1 from the model is evaluated. In this case, the interaction is removed because it contains X_1.

3. Tests if the parameters referring to X_2 are equal to 0. That is, the impact of removing X_2 from the model is evaluated. In this case, the interaction is removed because it contains X_2.

4. Tests if the interaction effect is 0.

3.2. Multiple comparisons test via Wald test

When ANOVA shows a significant effect of a categorical variable, it is usually of interest to assess which of the levels differ from each other. For this, multiple comparison tests are used. In the literature there are several procedures to perform such tests, many of them described in Hsu (1996).

Such a situation can be evaluated using the Wald test. By correctly specifying the $\boldsymbol{L}$ matrix, it is possible to evaluate hypotheses about any possible contrast between the levels of a given categorical variable. Therefore, it is possible to use Wald's statistics to perform multiple comparison tests as well.

The procedure is basically based on 3 steps: (i) obtain the matrix of linear combinations of the model parameters that result in the adjusted means; (ii) generate the matrix of contrasts, given by subtracting each pair of lines from the matrix of linear combinations; and (iii) select the lines of interest from this matrix and use them as the Wald test hypothesis specification matrix, instead of the $\boldsymbol{L}$ matrix.

For example, suppose there is a response variable Y subject to an explanatory variable X of 4 levels: A, B, C and D. To evaluate the effect of the variable X, a model given by:

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

In this parameterization, the first level of the categorical variable is the reference category and, for the other levels, the change to the reference category is measured; this is called the treatment contrast. In this context β_0 represents the adjusted mean of level A, while β_1 represents the difference from A to B, β_2 represents the difference from A to C and β_3 represents the difference from A to D. With this parameterization it is possible to obtain the predicted value for any of the categories in such a way that if the individual belongs to category A, β_0 represents the predicted value; if the individual belongs to category B, $\beta_0 + \beta_1$ represents the predicted; for category C, $\beta_0 + \beta_2$ represents the predicted, and finally, for category D, $\beta_0 + \beta_3$ represents the predicted.

In the matrix, these results can be described as:

$$
\boldsymbol{K_0} = \begin{matrix} A \\ B \\ C \\ D \end{matrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}
$$

Note that the product $\boldsymbol{K_0}\boldsymbol{\beta}$ generates the vector of predictions for each level of X. By subtracting the rows from the matrix of linear combinations $\boldsymbol{K_0}$ we can generate a matrix of contrasts $\boldsymbol{K_1}$:

$$
\boldsymbol{K_1} = \begin{matrix} A-B \\ A-C \\ A-D \\ B-C \\ B-D \\ C-D \end{matrix} \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}
$$

To carry out a test of multiple comparisons, we just select the desired contrasts in the lines of the matrix $\boldsymbol{K_1}$ and use these lines as a matrix for specifying the hypotheses of the Wald test. Finally, as usual in tests of multiple comparisons, correction of p-values by means of Bonferroni correction is recommended.

To carry out this procedure for McGLMs, we must remember that this is a class of multivariate models. And as in the case of analysis of variance, for tests of multiple comparisons there are two possibilities: tests for a single response and tests for multiple responses.

In practice, if the interest is a multivariate multiple comparison test, there is a need for all responses to be subject to the same predictor and it is enough to expand the contrast matrix using the Kronecker product. In the case of a multiple comparison test for each response, simply select the vector of estimates and the partition corresponding to the matrix $J_{\boldsymbol{\theta}}^{-1}$ for the specific response and proceed with the test.

4. Implementation

All functions implemented in the package **htmcglm** generate results showing degrees of freedom and p-values based on the Wald test applied to a McGLM. The Table 1 shows the names and a brief description of the implemented functions.

The functions `mc_anova_I()`, `mc_anova_II()` and `mc_anova_III()` are functions designed to evaluate regression parameters; they generate analysis of variance tables per response for a McGLM fitted in the **mcglm** package. The functions `mc_manova_I()`, `mc_manova_II()` and `mc_manova_III()` are also functions designed to evaluate the regression parameters of the model; they generate multivariate analysis of variance tables for a McGLM where the responses are subject to the same predictor. While univariate analysis of variance functions aim to assess the effect of variables for each response, multivariate ones aim to assess the effect of explanatory variables on all response variables simultaneously. The nomenclatures follow what is presented in **REF MY ARTICLE** and the functions receive as an argument only the object that stores the properly adjusted model.

function	Description
`mc_anova_I()`	ANOVA type I
`mc_anova_II()`	ANOVA type II
`mc_anova_III()`	ANOVA type III
`mc_manova_I()`	MANOVA type I
`mc_manova_II()`	MANOVA type II
`mc_manova_III()`	MANOVA type III
`mc_anova_dispersion()`	ANOVA type III for dispersion
`mc_manova_dispersion()`	MANOVA type III for dispersion
`mc_multcomp()`	Multiple comparison tests per response
`mc_mult_multcomp()`	Multivariate multiple comparison tests
`mc_linear_hypothesis()`	User-specified general linear hypothesis

Table 1: Functions implemented in the htmcglm package.

As described in subsection 2.1, the $\mathbf{\Omega(\tau)}$ matrix aims to model the correlation between rows of the data set through the so-called matrix linear predictor. In practice we have, for each matrix of the matrix predictor, a dispersion parameter τ_d. Similar to what is done for the mean predictor, we can use these parameters to assess the effect of correlated units in the study. For this, we implement the functions `mc_anova_dispersion()` and `mc_manova_dispersion()`.

The `mc_anova_dispersion()` function performs an analysis of variance for the model's dispersion parameters. Similar to the other functions with the prefix `mc_anova`, a table is generated for each response variable, that is, in the most general cases, we evaluate whether there is evidence that allows us to say that a given dispersion parameter is equal to 0, that is, whether there is an effect of the correlated measures as specified in the matrix linear predictor for that response. The function receives as argument the object in which the model is stored, a list of indices indicating how the dispersion parameters must be tested for each response, in such a way that the dispersion parameters that must be tested together share the same index; the last argument is the set of names to be shown in the final table.

The `mc_manova_dispersion()` function can be used in a multivariate model in which the matrix linear predictors are the same for all responses and there is an interest in evaluating whether the effect of correlated measures is the same for all responses. This function receives as argument the object in which the model is stored, a vector of indices indicating how the dispersion parameters must be tested, in such a way that dispersion parameters that must be tested together share the same index; the last argument is the set of names to be shown in the final table.

For multiple comparisons tests, the functions `mc_multcomp()` and `mc_mult_multcomp()` were implemented. These functions should be used as a complement to the analysis of variance and multivariate analysis of variance functions when they show a significant effect of categorical explanatory variables. The functions for multiple comparisons are used to perform two-by-two comparisons and identify which levels differ from each other. These functions receive as an argument the model, the variable or variables in which there is interest in evaluating comparisons between levels and also the data used to fit the model.

Finally, the `mc_linear_hypothesis()` function is the most flexible we have in the set of implementations. With it, it is possible to specify any type of hypothesis about regression, dispersion or power parameters of a McGLM. It is also possible to specify hypotheses on

multiple parameters and the vector of null hypothesis values is user defined. This function receives as arguments the model, a vector containing the parameters to be tested and the values under the null hypothesis. With some work, using the general linear hypotheses function, it is possible to replicate the results obtained by the analysis of variance functions.

5. Examples

In this section we provide practical examples of using the functions implemented in the **htmcglm** package based on multivariate models fitted with the **mcglm** package.

5.1. Example 1: soya

The data are from an experiment carried out in a greenhouse with soybeans. The experimental design has two plants per plot in which each unit was subjected to different combinations of water and fertilizer. There are three levels of a factor corresponding to the amount of water in the soil (`water`) and five levels of potassium fertilization (`pot`). In addition, the plots were arranged in five blocks (`block`). Three response variables were evaluated: grain yield (`grain`), number of seeds (`seeds`) and number of viable peas per plant (`viablepeas`).

This is an interesting dataset to exemplify the use of the implemented functions because there are three response variables of different types: grain yield is a continuous variable, the number of seeds is a count, and the number of viable peas per plant is an example of a binomial variable. The dataset is available in the `mcglm` package.

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

The objective of the analysis is to evaluate the effect of fertilization and water on the three response variables of interest. For the purposes of analysis, we considered as explanatory variables the levels of water, fertilization and also the interactions between these two factors. Additionally, the block effect was added to the predictors. To fit the model, the first step is to specify the linear predictors.

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

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

The second step is to specify the matrix linear predictor. We consider in this case that the observations are independent, so we include only one identity matrix.

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

With the elements defined, we can adjust the model. Through the function `mcglm()` we specify the linear predictors for the mean, the matrices of the matrix linear predictors, the link and variance functions, the number of trials for the binomial variable and whether or not we are interested in estimating the power parameters. For more details on specifying predictors and fit McGLMs, see Bonat and Jørgensen (2016) and 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)
```

To evaluate some results of the model it is possible to use the function `summary()` that returns the formula of the linear predictors, the link, variance and covariance functions specified to fit the model, the estimates of the regression and dispersion parameters as well as standard errors.

With the model adjusted, we can apply the implemented functions to evaluate the regression and dispersion parameters of the model. The analysis of variance functions depend only on the object that contains the fitted model and return a table for each response.

ANOVA type 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 type 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 type 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
```

Similarly, multivariate analysis of variance functions also depend only on the fitted model. It is important to note that for practical purposes the multivariate analysis of variance functions require the predictors for all responses to be the same.

MANOVA type 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 type 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 type 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
```

For general linear hypotheses about regression parameters, it is sufficient to specify the model and the hypothesis to be tested. To identify the parameters of interest, use the `coef()` function.

Test on a single regression parameter

```
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
```

Test on more than one regression parameter

```
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
```

Test of equality of effects between regression parameters

```
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
```

5.2. Example 2: Hunting

The Hunting dataset, presented in Bonat, Olivero, Grande-Vega, Farfán, and Fa (2017), is

also available in the package **mcglm**. The data addresses a problem where responses are longitudinal bivariate counts on animals hunted in Basile Fang village, Bioko North Province, Bioko Island, Equatorial Guinea. The response variables are: monthly numbers of blue duikers (`BD`) and other small animals (`OT`) shot or captured in a random sample of 52 commercial hunters from August 2010 to September 2013. Assume that the interest is to evaluate the effect of a factor with 2 levels that indicates if the animal was hunted by means of a firearm or trap (`METHOD`) and a factor with 2 levels that indicates the sex of the animal (`SEX`).

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

As in the first example, to fit the model it is necessary to define the linear predictors for the mean, the matrices of the linear matrix predictors, the link and variance functions, whether or not we are interested in estimating the power parameters. For this analysis, we considered in the matrix predictor the structure of repeated measures introduced by the observations taken for the same hunter and month (`HUNTER.MONTH`) and the number of hunting days per month was used as an offset term.

```
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)
```

Again, to evaluate some model results it is possible to use the `summary()` function. We can also apply the already presented functions implemented for ANOVAs, MANOVAs and tests of general linear hypotheses on the regression parameters of the model.

In this case, as there is a specified matrix predictor, an in-depth study of the dispersion parameters may be of interest. This analysis can be done with the already used function `mc_linear_hypothesis()`.

Test on a single dispersion parameter

```
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
```

Test on more than one dispersion parameter

```
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
```

Test of equality of effects between dispersion parameters

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

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

To evaluate dispersion parameters, we have the procedure analogous to the analysis of variance for regression parameters. These functions require specifying more arguments: one that determines the relationship between dispersion parameters and the other that specifies the names that will appear in the final output.

ANOVA type III for dispersion

```
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 type III for dispersion

```
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
```

Finally, we can use the functions for testing multiple comparisons to assess differences between levels of categorical explanatory variables included in the model.

Univariate multiple comparisons test

```
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
```

Multivariate multiple comparisons test

```
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 procedures to perform hypothesis tests on McGLMs parameters based on Wald statistics. McGLMs have regression, dispersion, power and correlation parameters; each set of parameters has a very relevant practical interpretation in the context of problem analysis with potential multiple responses as a function of a set of explanatory variables.

Based on the proposed use of the Wald test for McGLMs, we developed the **htmcglm** with procedures for testing general linear hypotheses, generating ANOVA and MANOVA tables for regression and dispersion parameters and also multiple comparisons tests. All these procedures were implemented in the R language and complement the existing functionalities in the **mcglm** library.

The discussed examples illustrate how to evaluate the most common hypotheses that arise in regression problems: evaluating parameters individually and evaluating sets of parameters. We focused our efforts on tools to evaluate regression and dispersion parameters, because by studying regression parameters it is possible to identify the variables that have a significant effect on the response; on the other hand, the dispersion parameters allow assessing whether there is an effect of correlated observations. In this way, the study of these quantities provides valuable information about the importance of the elements of a multivariate regression problem.

Possible extensions of the **mcglm** package follow the idea of evaluation of McGLMs parameters for a better understanding of the impact of elements in modeling problems. Some possibilities are: exploring corrections of p-values according to the size of the tested hypotheses, exploring procedures beyond the Wald test (such as the Score test and the likelihood ratio test), implementing new procedures for multiple comparisons, adapting the proposal to deal with alternative contrasts to the usual ones, explore procedures for automatic selection of covariates (backward elimination, forward selection, stepwise selection) and also covariate selection through the inclusion of penalty in the complexity adjustment (similar to the idea of spline regression).

Acknowledgments

References

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

Bonat W, Olivero J, Grande-Vega M, Farfán M, Fa J (2017). "Modelling the covariance structure in marginal multivariate count models: Hunting in Bioko Island." *Journal of Agricultural, Biological and Environmental Statistics*, **22**(4), 446–464.

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.

Hsu J (1996). *Multiple comparisons: theory and methods*. CRC Press.

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 verson 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