LINEU ALBERTO CAVAZANI DE FREITAS

TESTES DE HIPÓTESES EM MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA

(*versão pré-defesa, compilada em 28 de julho de 2021*)

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

Área de concentração: *Ciência da Computação*.

Orientador: Prof. Dr. Wagner Hugo Bonat.

Coorientador: Prof. Dr. Marco Antonio Zanata Alves.

CURITIBA PR

2021

RESUMO

Ciência de dados é um campo de estudo interdisciplinar que compreende áreas como estatística, ciência da computação e matemática. Neste contexto, métodos estatísticos são de fundamental importância sendo que, dentre as possíveis técnicas disponíveis para análise de dados, os modelos de regressão tem papel importante. Tais modelos são indicados a problemas nos quais existe interesse em verificar a associação entre uma ou mais variáveis respostas e um conjunto de variáveis explicativas; isto é feito através da obtenção de uma equação que explique a relação entre as variáveis explicativas e a(s) resposta(s). Existem modelos uni e multivariados: nos modelos univariados há apenas uma variável resposta; já em modelos multivariados há mais de uma resposta. Dentre as classes de modelos multivariados estão os modelos multivariados de covariância linear generalizada (McGLMs). Trata-se de uma flexível clase que permite lidar com múltiplas respostas de diferentes naturezas, correlacionadas entre si em que é possível modelar também a correlação entre indivíduos do conjunto de dados. No contexto de modelos de regressão, um interesse comum costuma ser verificar se a retirada de determinada variável explicativa gera um modelo significativamente pior, ou seja, avalia-se se há evidência suficiente nos dados para afirmar que determinada variável explicativa não possui efeito sobre a resposta. Tais conjecturas são avaliadas através dos chamados testes de hipóteses. Três testes de hipóteses são comuns em regressão: o teste da razão de verossimilhanças, o teste Wald e o teste do multiplicador de lagrange, também conhecido como teste escore. Existem ainda técnicas baseadas em testes de hipóteses tais como a análise de variância (ANOVA) em que o objetivo é a avaliação do efeito de cada uma das variáveis explicativas sobre a(s) resposta(s); isto é feito através da comparação via testes de hipóteses entre modelos com e sem cada uma das variáveis explicativas. Para o caso multivariado estende-se a técnica de análise de variância (ANOVA) para a análise de variância multivariada (MANOVA). No entanto, considerando os modelos multivariados de covariância linear generalizada, não há discussão a respeito da construção destes testes para a classe. Assim, por se tratar de uma classe de modelos flexível e com alto poder de aplicação a problemas práticos, nosso objetivo geral é o desenvolvimento de testes de hipóteses para os McGLMs. Nossa proposta é adaptar o teste Wald para a realização de testes de hipóteses gerais sobre parâmetros de McGLMs. Temos como objetivos implementar funções para efetuar tais testes, bem como funções para efetuar ANOVAs e MANOVAs. As propriedades e comportamento dos testes propostos serão verificados com base em estudos de simulação e o potencial de aplicação das metodologias discutidas será apresentado com base na aplicação a conjuntos de dados reais.

Palavras-chave: Modelos multivariados de covariância linear generalizada. Testes de hipóteses. Teste Wald. ANOVA. MANOVA.

ABSTRACT

Data science is an interdisciplinary field of study comprising areas such as statistics, computer science and mathematics. In this context, statistical methods are of fundamental importance and, among the possible techniques available for data analysis, regression models play an important role. These models are suitable for problems in which there is an interest in verifying the association between one or more response variables and a set of explanatory variables; this is done by obtaining an equation that explains the relationship between the explanatory variables and the answer(s). There are univariate and multivariate models: in univariate models there is only one response variable; in multivariate models there are more than one answer. Among the classes of multivariate covariance generalized linear models (McGLMs). It is a flexible class that allows dealing with multiple responses of different types, correlated with each other, in which it is also possible to model the correlation between individuals of the data set. In the context of regression models, a common interest is usually to verify whether the removal of a certain explanatory variable generates a significantly worse model, that is, it is evaluated whether there is enough evidence in the data to state that a certain explanatory variable has no effect on the response. These conjectures are evaluated through so-called hypothesis tests. Three hypothesis tests are common in regression: the likelihood ratio test, the Wald test, and the lagrange multiplier test, also known as the score test. There are also techniques based on hypothesis tests such as analysis of variance (ANOVA) in which the objective is to evaluate the effect of each of the explanatory variables on the answer(s); this is done through the comparison via hypothesis tests between models with and without each of the explanatory variables. For the multivariate case, the analysis of variance technique (ANOVA) is extended to the multivariate analysis of variance (MANOVA). However, considering the multivariate covariance generalized linear models, there is no discussion about the construction of these tests for the class. Thus, as it is a flexible class of models with high application power to practical problems, our general goal is the development of hypothesis tests for McGLMs. Our proposal is to adapt the Wald test to perform general hypothesis tests on McGLMs parameters. We aim to implement functions to perform such tests, as well as functions to perform ANOVAs and MANOVAs. The properties and behavior of the proposed tests will be verified based on simulation studies and the potential application of the discussed methodologies will be presented based on their application to real data sets.

Keywords: Multivariate covariance generalized linear models. Hypothesis tests. Wald test. ANOVA. MANOVA.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Desde o surgimento do termo *data science* por volta de 1996 (Press, 2013) a discussão sobre o tema atrai pesquisadores das mais diversas áreas (Cao, 2016). A ciência de dados é vista como um campo de estudo de natureza interdisciplinar que incorpora conhecimento de grandes áreas como estatística, ciência da computação e matemática (Ley e Bordas, 2018). Weihs e Ickstadt (2018) afirmam que a ciência de dados é um campo em muito influenciado por áreas como informática, ciência da computação, matemática, pesquisa operacional, estatística e ciências aplicadas. Em (Cao, 2016) é dito que ciência de dados engloba técnicas de como: estatística, aprendizado de máquina, gerenciamento de *big data*, dentre outras.

Alguns dos campos de interesse da ciência de dados são: métodos de amostragem, mineração de dados, bancos de dados, técnicas de análise exploratória, probabilidade, inferência, otimização, infraestrutura computacional, plataformas de *big data*, modelos estatísticos, dentre outros. Weihs e Ickstadt (2018) afirmam que os métodos estatísticos são de fundamental importância em grande parte das etapas da ciência de dados. Neste sentido, os modelos de regressão tem papel importante. Tais modelos são indicados a problemas nos quais existe interesse em verificar a associação entre uma ou mais variáveis resposta (também chamadas de variáveis dependentes) e um conjunto de variáveis explicativas (também chamadas de variáveis independentes, covariáveis ou preditoras).

Para entender minimamente um modelo de regressão, é necessário compreender o conceito de fenômeno aleatório, variável aleatória e distribuição de probabilidade. Um fenômeno aleatório é uma situação na qual diferentes observações podem fornecer diferentes desfechos. Estes fenômenos podem ser descritos por variáveis aleatórias que associam um valor numérico a cada desfecho possível do fenômeno. Os desfechos deste fenômeno podem ser descritos por uma escala que pode ser discreta ou contínua. Uma variável aleatória é considerada discreta quando os possíveis desfechos estão dentro de um conjunto enumerável de valores. Já uma variável aleatória contínua ocorre quando os possíveis resultados estão em um conjunto não enumerável de valores. Na prática existem probabilidades associadas aos valores de uma variável aleatória, e estas probabilidades podem ser descritas através de funções. No caso das variáveis discretas, a função que associa probabilidades aos valores da variável aleatória é chamada de função de probabilidade. No caso das contínuas, esta função é chamada de função densidade de probabilidade.

Existem ainda modelos probabilísticos que buscam descrever as probabilidades de variáveis aleatórias, as chamadas distribuições de probabilidade. Portanto, em problemas práticos, podemos buscar uma distribuição de probabilidades que melhor descreva o fenômeno de interesse. Estas distribuições são descritas por funções e tais funções possuem parâmetros que controlam aspectos da distribuição como escala e forma, tais parâmetros são quantidades desconhecidas

estimadas através dos dados. Na análise de regressão busca-se modelar os parâmetros das distribuições de probabilidade como uma função de outras variáveis. Isto é feito através da decomposição do parâmetro da distribuição em outros parâmetros, chamados de parâmetros de regressão, que dependem de variáveis conhecidas e fixas, as variáveis explicativas.

Assim, o objetivo dos modelos de regressão consiste em obter uma equação que explique a relação entre as variáveis explicativas e o parâmetro de interesse da distribuição de probabilidades selecionada para modelar a variável aleatória. Em geral, o parâmetro de interesse da distribuição de probabilidades modelado em função das variávis explicativas é a média. Fazendo uso da equação resultante do processo de análise de regressão, é possível estudar a importância das variáveis explicativas sobre a resposta e realizar predições da variável resposta com base nos valores observados das variáveis explicativas.

Em contextos práticos o processo de análise via modelo de regressão parte de um conjunto de dados. Neste contexto, um conjunto de dados é uma representação tabular em que unidades amostrais são representadas nas linhas e seus atributos (variáveis) são representados nas colunas. Pode-se usar um modelo de regressão para, por exemplo, modelar a relação entre a média de uma variável aleatória e um conjunto de variáveis explicativas. Assume-se então que a variável aleatória segue uma distribuição de probabilidades e que o parâmetro de média desta distribuição pode ser descrito por uma combinação linear de parâmetros de regressão associados às variáveis explicativas. Sendo assim, o conhecimento a respeito da influência de uma variável explicativa sobre a resposta vem do estudo das estimativas dos parâmetros de regressão. A obtenção destes parâmetros estimados se dá na chamada etapa de ajuste do modelo, e isto gera a equação da regressão ajustada.

Existem na prática modelos uni e multivariados. Nos modelos univariados há apenas uma variável resposta e temos interesse em avaliar o efeito das variáveis explicativas sobre essa única resposta. No caso dos modelos multivariados há mais de uma resposta e o interesse passa a ser avaliar o efeito dessas variáveis sobre todas as respostas. Existem inúmeras classes de modelos de regressão, mencionaremos neste trabalho três importantes classes: os modelos lineares, os lineares generalizados e os multivariados de covariância linear generalizada. No cenário univariado, durante muitos anos o modelo linear normal (Galton, 1886) teve papel de destaque no contexto dos modelos de regressão devido principalmente as suas facilidades computacionais. Um dos pressupostos do modelo linear normal é de que a variável resposta, condicional às variáveis explicativas, segue a distribuição normal. Todavia, não são raras as situações em que a suposição de normalidade não é atendida. Uma alternativa, por muito tempo adotada, foi buscar uma transformação da variável resposta a fim de atender os pressupostos do modelo, tal como a família de transformações proposta por Box e Cox (1964). Contudo, este tipo de solução leva a dificuldades na interpretação dos resultados.

Com o passar o tempo, o avanço computacional permitiu a proposição de modelos mais complexos, que necessitavam de processos iterativos para estimação dos parâmetros (Paula, 2004). A proposta de maior renome foram os modelos lineares generalizados (GLM) propostos

por Nelder e Wedderburn (1972). Essa classe de modelos permitiu a flexibilização da distribuição da variável resposta de tal modo que esta pertença à família exponencial de distribuições. Em meio aos casos especiais de distribuições possíveis nesta classe de modelos estão a Bernoulli, binomial, Poisson, normal, gama, normal inversa, entre outras. Trata-se portanto, de uma classe de modelos de regressão univariados para dados de diferentes naturezas, tais como: dados contínuos simétricos e assimétricos, contagens e assim por diante. Tais características tornam esta classe uma flexível ferramenta de modelagem aplicável a diversos tipos de problema.

Petterle et al. (2017) fez uso de modelos lineares generalizados em um problema de resposta binária em que o objetivo era avaliar a probabilidade de uso incorreto do sistema de retenção de diferentes capacetes de motociclistas. Michelon et al. (2019) avaliou diferentes modelos para dados de contagem na classe dos GLM para modelar o número de sementes de Eucalyptus cloeziana. Larsen et al. (2011) utilizou um modelo linear generalizado com distribuição gama com o objetivo de avaliar que características influenciam os níveis da pressão parcial de dióxido de carbono (pCO_2) em lagos localizados ao sul e centro da Noruega. Mais a respeito dos modelos lineares generalizados pode ser visto em Paula (2004) e Cordeiro e Demétrio (2008).

Embora as técnicas citadas sejam úteis, há casos em que são coletadas mais de uma resposta por unidade experimental e há o interesse de modelá-las em função de um conjunto de variáveis explicativas. Neste cenário surgem os modelos multivariados de covariância linear generalizada (McGLM) propostos por Bonat e Jørgensen (2016). Essa classe pode ser vista com uma extensão multivariada dos GLMs que permite lidar com múltiplas respostas de diferentes naturezas e, de alguma forma, correlacionadas. Além disso, não há nesta classe suposições quanto à independência entre as observações, pois a correlação entre observações pode ser modelada por um preditor linear matricial que envolve matrizes conhecidas. Estas características tornam o McGLM uma classe flexível ao ponto de ser possível chegar a extensões multivariadas para modelos de medidas repetidas, séries temporais, dados longitudinais, espaciais e espaço-temporais.

Quando trabalha-se com modelos de regressão, um interesse comum aos analistas é o de verificar se a retirada de determinada variável explicativa do modelo geraria uma perda no ajuste. Ou seja, uma conjectura de interesse é avaliar se há evidência suficiente nos dados para afirmar que determinada variável explicativa não possui efeito sobre a resposta. Isto é feito através dos chamados testes de hipóteses. Testes de hipóteses são ferramentas estatísticas que auxiliam no processo de tomada de decisão sobre valores desconhecidos (parâmetros) estimados por meio de uma amostra (estimativas). Tal procedimento permite verificar se existe evidência nos dados amostrais que apoiem ou não uma hipótese estatística formulada a respeito de um parâmetro. As suposições a respeito de um parâmetro desconhecido estimado com base nos dados são denominadas hipóteses estatísticas, estas hipóteses podem ser rejeitadas ou não rejeitadas com base nos dados. Segundo Lehmann (1993) podemos atribuir a teoria, formalização e filosofia dos

testes de hipótese a Neyman e Pearson (1928a), Neyman e Pearson (1928b) e Fisher (1925). A teoria clássica de testes de hipóteses é apresentada formalmente em Lehmann e Romano (2006).

No contexto de modelos de regressão, três testes de hipóteses são comuns: o teste da razão de verossimilhanças, o teste Wald e o teste do multiplicador de lagrange, também conhecido como teste escore. Engle (1984) descreve a formulação geral dos três testes. Todos eles são baseados na função de verossimilhança dos modelos. Um modelo de regressão busca encontrar o valor dos parâmetros que associam variáveis explicativas às respostas que maximizam a função de verossimilhança, ou seja, buscam encontrar um conjunto de parâmetros desconhecidos que façam com o que o dado seja provável (verossímil).

O teste da razão de verossimilhanças, inicialmente proposto por Wilks (1938), é efetuado a partir de dois modelos com o objetivo de compará-los. A ideia consiste em obter um modelo com todas as variáveis explicativas e um segundo modelo sem algumas dessas variáveis. O teste é usado para comparar estes modelos através da diferença do logaritmo da função de verossimilhança. Caso essa diferença seja estatísticamente significativa, significa que a retirada das variáveis do modelo completo prejudicam o ajuste. Caso não seja observada diferença entre o modelo completo e o restrito, significa que as variáveis retiradas não geram perda na qualidade e, por este motivo, tais variáveis podem ser descartadas.

Já o teste Wald, proposto por Wald (1943), requer apenas um modelo ajustado. A ideia consiste em verificar se existe evidência para afirmar que um ou mais parâmetros são iguais a valores postulados. O teste avalia quão longe o valor estimado está do valor postulado. Utilizando o teste Wald é possível formular hipóteses para múltiplos parâmetros, e costuma ser de especial interesse verificar se há evidência que permita afirmar que os parâmetros que associam determinada variável explicativa a variável resposta são iguais a zero. Caso tal hipótese não seja rejeitada, significa que caso estas variáveis sejam retiradas, não existirá perda de qualidade no modelo.

O teste do multiplicador de lagrange ou teste escore (Aitchison e Silvey, 1958), (Silvey, 1959), (Rao, 1948), tal como o teste Wald, requer apenas um modelo ajustado. No caso do teste escore o modelo ajustado não possui o parâmetro de interesse e o que é feito é testar se adicionar esta variável omitida resultará em uma melhora significativa no modelo. Isto é feito com base na inclinação da função de verossimilhança, esta inclinação é usada para estimar a melhoria no modelo caso as variáveis omitidas fossem incluídas.

De certo modo, os três testes podem ser usados para verificar se a retirada de determinada variável do modelo prejudica o ajuste. No caso do teste de razão de verossimilhanças, dois modelos precisam ser ajustados. Já o teste Wald e o escore necessitam de apenas um modelo. Além disso, os testes são assintóticamente equivalentes. Em amostras finitas estes testes podem apresentar resultados diferentes como discutido por Evans e Savin (1982).

Para o caso dos modelos lineares tradicionais existem técnicas como a análise de variância (ANOVA), proposta inicialmente por Fisher e Mackenzie (1923). Segundo St et al. (1989), a ANOVA é um dos métodos estatísticos mais amplamente usados para testar hipóteses e

que está presente em praticamente todos os materiais introdutórios de estatística. O objetivo da técnica é a avaliação do efeito de cada uma das variáveis explicativas sobre a resposta. Isto é feito através da comparação via testes de hipóteses entre modelos com e sem cada uma das variáveis explicativas. Logo, tal procedimento permite que seja possível avaliar se a retirada de cada uma das variáveis gera um modelo significativamente pior quando comparado ao modelo com a variável. Para o caso multivariado estende-se a técnica de análise de variância (ANOVA) para a análise de variância multivariada (Smith et al., 1962), a MANOVA. E dentre os testes de hipóteses multivariados já discutidos na literatura, destacam-se o λ de Wilk's (Wilks, 1932), traço de Hotelling-Lawley (Lawley, 1938), (Hotelling, 1951), traço de Pillai (Pillai et al., 1955) e maior raiz de Roy (Roy, 1953).

Buscamos até aqui enfatizar a importância dos modelos de regressão no contexto de ciência de dados e sua relevância na análise de problemas práticos. Além disso ressaltamos a importância dos testes de hipóteses e também de procedimentos baseados em tais testes para fins de avaliação da importância das variáveis incluídas nos modelos. No entanto, considerando os modelos multivariados de covariância linear generalizada, não há discussão a respeito da construção destes testes para a classe. Assim, por se tratar de uma classe de modelos flexível e com alto poder de aplicação a problemas práticos, nosso objetivo geral é o desenvolvimento de testes de hipóteses para os McGLMs.

Em nosso trabalho buscamos propor uma adaptação do teste de Wald clássico utilizado em modelos lineares para os McGLMs. A construção do teste Wald em modelos tradicionais é baseada nas estimativas de máxima verossimilhança. Contudo a estatística de teste usada não depende da máxima verossimilhança, e sim de um vetor de estimativas dos parâmetros e uma matriz de variância e covariância destas estimativas. Assim, por mais que os McGLMs não sejam ajustados com base na maximização da função de verossimilhança para obtenção dos parâmetros do modelo, o método de estimação fornece os componentes necessários para a construção do teste. Neste sentido, das três opções clássicas de testes de hipóteses comumente aplicados a problemas de regressão, o teste Wald se torna o mais atrativo pois basta adaptar a estatística de teste.

Nosso trabalho tem os seguintes objetivos específicos: adaptar o teste Wald para realização de testes de hipóteses gerais sobre parâmetros de modelos multivariados de covariância linear generalizada, implementar funções para efetuar tais testes, bem como funções para efetuar análises de variância e análises de variância multivariadas para os McGLMs, avaliar as propriedades e comportamento dos testes propostos com base em estudos de simulação e avaliar o potencial de aplicação das metodologias discutidas com base na aplicação a conjuntos de dados reais.

Este é um projeto de qualificação e está organizado em cinco capítulos: na atual seção foi exposto o tema de forma a enfatizar as características dos modelos de regressão e a utilidade dos testes de hipóteses neste contexto. O Capítulo 2 é dedicado à revisão bibliográfica da estrutura dos McGLMs e testes de hipótese. No Capítulo 3 é apresentada nossa proposta de adaptação do

teste Wald para avaliar suposições sobre parâmetros de um McGLM. No capítulo 4 são mostrados os resultados preliminares. E por fim, no Capítulo 5 são discutidas as tarefas a serem cumpridas até o fim do mestrado, desafios, resultados esperados e o cronograma das atividades.

2 REVISÃO DE LITARATURA

Nossa revisão de literatura aborda predominantemente três temas. O primeiro deles é uma revisão da estrutura geral e estimação dos parâmetros de um modelo multivariado de covariância linear generalizada, baseado nas ideias de Bonat e Jørgensen (2016). A segunda parte da revisão de literatura diz respeito ao procedimento dos chamados testes de hipóteses com o foco de tratar do objetivo, notação, componentes, aplicação deste tipo de procedimento no contexto de modelos de regressão e ainda quais os testes mais comuns. Por fim, a última parte da revisão diz respeito às análises de variância, que podem ser vistos como procedimentos baseados em testes de hipóteses sequenciais para avaliar os parâmetros de um modelo de regressão.

2.1 MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA

Os modelos linerares generalizados (GLM), propostos por Nelder e Wedderburn (1972), são uma forma de modelagem que lida exclusivamente com uma resposta e com respostas de diferentes naturezas, tais como respostas contínuas, binárias e contagens. Tais características tornam essa classe de modelos uma flexível ferramenta de modelagem aplicável a diversos tipos de problemas. Contudo, por mais flexível e discutida na literatura, essa classe apresenta ao menos três importantes restrições: um leque restrito de distribuições disponíveis para modelagem, a incapacidade de lidar com observações dependentes e a incapacidade de lidar com múltiplas respostas simultaneamente.

Com o objetivo de contornar estas restrições, foi proposta por Bonat e Jørgensen (2016), uma estrutura geral para análise de dados não gaussianos com múltiplas respostas em que não se faz suposições quanto à independência das observações: os chamados modelos multivariados de covariância linear generalizada (McGLMs). Tais modelos, levam em conta a não normalidade por meio de uma função de variância. Além disso, a estrutura média é modelada por meio de uma função de ligação e um preditor linear. Os parâmetros dos modelos são obtidos através de funções de estimação baseadas em suposições de segundo momento.

Vamos discutir os McGLMs como uma extensão dos GLMs. Vale ressaltar que é usada uma especificação menos usual de um modelo linear generalizado, porém trata-se de uma notação mais conveniente para chegar à uma especificação mais simples de um modelo multivariado de covariância linear generalizada.

2.1.1 Modelo linear generalizado

Para definição da extensão de um modelo linear generalizdo (GLM) apresentada por Bonat e Jørgensen (2016), considere $\boldsymbol{Y}$ um vetor $N \times 1$ de valores observados da variável resposta,

$\boldsymbol{X}$ uma matriz de delineamento $N \times k$ e $\boldsymbol{\beta}$ um vetor de parâmetros de regressão $k \times 1$. Com isso, um GLM pode ser escrito da seguinte forma

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

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

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

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

2. A função de dispersão Poisson–Tweedie, a qual caracteriza a família Poisson-Tweedie de distribuições, que visa contornar a inflexibilidade da utilização da função de variância potência para respostas discretas. A família Poisson-Tweedie tem função de dispersão dada por $\vartheta\left(\boldsymbol{\mu}; p\right) = \mu + \tau\mu^p$, em que τ é o parâmetro de dispersão. A função de dispersão Poisson-Tweedie tem como casos particulares os mais famosos modelos para dados de contagem: Hermite ($p = 0$), Neyman tipo A ($p = 1$), binomial negativa ($p = 2$) e Poisson–inversa gaussiana (p = 3) (Jørgensen e Kokonendji, 2015). Não se trata de uma função de variância usual, mas é uma função que caracteriza o relacionamento entre média e variância.

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

Lembre-se que o GLM é uma classe de modelos de regressão univariados em que um dos pressupostos é a independência entre as observações. Esta independência é especificada na matriz identidade no centro Equação 2.1. Podemos imaginar que, substituindo esta matriz identidade por uma matriz qualquer que reflita a relação entre os indivíduos da amostra teremos uma extensão do Modelo Linear Generalizado para observações dependentes. É justamente essa a ideia dos modelos de covariância linear generalizada, o cGLM.

2.1.2 Modelo de covariância linear generalizada

Os modelos de covariância linear generalizada (cGLM) são uma alternativa para problemas em que a suposição de independência entre as observações não é atendida. Neste caso, a solução proposta é substituir a matriz identidade $\boldsymbol{I}$ da Equação 2.1 por uma matriz não

diagonal $\boldsymbol{\Omega(\tau)}$ que descreva adequadamente a estrutura de correlação entre as observações. Trata-se de uma ideia similar à proposta de Liang e Zeger (1986) nos modelos GEE (Equações de Estimação Generalizadas), em que utiliza-se uma matriz de correlação de trabalho para considerar a dependência entre as observações. A matriz $\boldsymbol{\Omega(\tau)}$ é descrita como uma combinação de matrizes conhecidas tal como nas propostas de Anderson et al. (1973) e Pourahmadi (2000), podendo ser escrita da forma

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

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

Com isso, substituindo a matriz identidade da Equação 2.1 pela Equação 2.2, temos uma classe com toda a flexibilidade dos GLMs, porém contornando a restrição da independência entre as observações desde que o preditor linear matricial seja adequadamente especificado. Deste modo, é contornada a restrição da incapacidade de lidar com observações dependentes. Outra restrição diz respeito às múltiplas respostas e, contornando este problema, chegamos ao McGLM.

2.1.3 Modelos multivariados de covariância linear generalizada

Os modelos multivariados de covariância linear generalizada (McGLMs) podem ser entendidos como uma extensão multivariada dos cGLMs que contornam as principais restrições presentes nos GLMs. Para definição de um McGLM considre $\boldsymbol{Y}_{N\times R} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_R\}$ uma matriz de variáveis resposta e $\boldsymbol{M}_{N\times R} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_R\}$ uma matriz de valores esperados. Cada uma das variáveis resposta tem sua própria matriz de variância e covariância, responsável por modelar a covariância dentro de cada resposta, sendo expressa por

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

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

Finalmente, um MCGLM é descrito como

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

em que $\boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b = \text{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \ldots, \tilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b \otimes \boldsymbol{I})\text{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^\top, \ldots, \tilde{\boldsymbol{\Sigma}}_R^\top)$ é o produto generalizado de Kronecker, a matriz $\tilde{\boldsymbol{\Sigma}}_r$ denota a matriz triangular inferior da decomposição de Cholesky da matriz $\boldsymbol{\Sigma}_r$, o operador Bdiag denota a matriz bloco-diagonal e $\boldsymbol{I}$ uma matriz identidade $N \times N$. Com isso, chega-se a uma classe de modelos na qual através da especificação da função de variâcia têm-se um leque maior de distribuições disponíveis, através do preditor matricial se torna possível a modelagem de dados com estrutura de covariância, e ainda é possível a modelagem de múltiplas respostas.

2.1.4 Estimação e inferência

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

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

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

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

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

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

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

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

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

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

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

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

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

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

em que $k_l^{(4)}$ denota a quarta cumulante de $\boldsymbol{\mathcal{Y}}_l$. No processo de estimação dos McGLMs é usada sua versão empírica.

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

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

Seja $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^{\top}, \hat{\boldsymbol{\lambda}}^{\top})^{\top}$ o estimador baseado na Equação 2.5 e Equação 2.8, a distribuição assintótica de $\hat{\boldsymbol{\theta}}$ é

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

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

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

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

Toda metodologia do McGLM está implementada no pacote *mcglm* (Bonat, 2018) do software estatístico R (R Core Team, 2020).

2.2 TESTES DE HIPÓTESES

A palavra "inferir"significa tirar conclusão. O campo de estudo chamado de inferência estatística tem como objetivo o desenvolvimento e discussão de métodos e procedimentos que permitem, com certo grau de confiança, fazer afirmações sobre uma popupalação com base em informação amostral. Na prática, costuma ser inviável trabalhar com uma população. Assim, a alternativa usada é coletar uma amostra e utilizar esta amostra para tirar conclusões. Neste

sentido, a inferência estatística fornece ferramentas para estudar quantidades populacionais (parâmetros) por meio de estimativas destas quantidades obtidas através da amostra.

Contudo, é importante notar que diferentes amostras podem fornecer diferentes resultados. Por exemplo, se há interesse em estudar a média de determinada característica na população mas não há condições de se observar a característica em todas as unidades, usa-se uma amostra. E é totalmente plausível que diferentes amostras apresentem médias amostrais diferentes. Portanto, os métodos de inferência estatística sempre apresentarão determinado grau de incerteza.

Campos importantes da inferência estatística são a estimação de quantidades (por ponto e intervalo) e testes de hipóteses. O objetivo desta revisão é apresentar uma visão geral a respeito de testes de hipóteses estatísticas e os principais componentes. Mais sobre inferência estatística pode ser visto em Barndorff-Nielsen e Cox (2017), Silvey (2017), Azzalini (2017), Wasserman (2013), entre outros.

2.2.1 Elementos de um teste de hipóteses

A atual teoria dos testes de hipóteses é resultado da combinação de trabalhos conduzidos predominantemente na década de 1920 por Ronald Fisher, Jerzy Neyman e Egon Pearson em publicações como Fisher (1992), Fisher (1929), Neyman e Pearson (2020a), Neyman e Pearson (2020b) e Neyman e Pearson (1933).

Entende-se por hipótese estatística uma afirmação a respeito de um ou mais parâmetros (desconhecidos) que são estimados com base em uma amostra. Já um teste de hipóteses é o procedimento que permite responder perguntas como: com base na evidência amostral, podemos considerar que dado parâmetro é igual a determinado valor? Alguns dos componentes de um teste de hipóteses são: as hipóteses, a estatística de teste, a distribuição da estatística de teste, o nível de significância, o poder do teste, a região crítica e o p-valor.

Para definição dos elementos necessários para condução de um teste de hipóteses, considere que uma amostra foi tomada com o intuito de estudar determinada característica de uma população. Considere $\hat{\theta}$ a estimativa de um parâmetro θ da população. Neste contexto, uma hipótese estatística é uma afirmação a respeito do valor do parâmetro θ que é estudado através da estimativa $\hat{\theta}$ a fim de concluir algo sobre a população de interesse.

Na prática, sempre são definidas duas hipóteses de interesse. A primeira delas é chamada de hipótese nula (H_0) e trata-se da hipótese de que o valor de um parâmetro populacional é igual a algum valor especificado. A segunda hipótese é chamada de hipótese alternativa (H_1) e trata-se da hipótese de que o parâmetro tem um valor diferente daquele especificado na hipótese nula. Deste modo, através do estudo da quantidade $\hat{\theta}$ verificamos a plausibilidade de se afirmar que θ é igual a um valor θ_0. Portanto, três tipos de hipóteses podem ser especificadas:

1. $H_0 : \theta = \theta_0 \; vs \; H_1 : \theta \neq \theta_0$.

2. $H_0 : \theta = \theta_0 \; vs \; H_1 : \theta > \theta_0$.

3. $H_0 : \theta = \theta_0 \; vs \; H_1 : \theta < \theta_0$.

Com as hipóteses definidas, dois resultados são possíveis em termos de H_0: rejeição ou não rejeição. O uso do termo "aceitar"a hipótese nula não é recomendado tendo em vista que a decisão a favor ou contra a hipótese se dá por meio de informação amostral. Ainda, por se tratar de um procedimento baseado em informação amostral, existe um risco associado a decisões equivocadas. Os possíveis desfechos de um teste de hipóteses estão descritos na Tabela 2.1, que mostra que existem dois casos nos quais toma-se uma decisão equivocada. Em uma delas rejeita-se uma hipótese nula verdadeira (erro do tipo I) e na outra não rejeita-se uma hipótese nula falsa (erro do tipo II).

A probabilidade do erro do tipo I é denotada por α e chamada de nível de significância, já a probabilidade do erro do tipo II é denotada por β. O cenário ideal é aquele que minimiza tanto α quanto β, contudo, em geral, à medida que α reduz, β tende a aumentar. Por este motivo busca-se controlar o erro do tipo I. Além disso temos que a probabilidade de se rejeitar a hipótese nula quando a hipótese alternativa é verdadeira (rejeitar corretamente H_0) recebe o nome de poder do teste.

	Rejeita H_0	**Não Rejeita** H_0
H_0 **verdadeira**	Erro tipo I	Decisão correta
H_0 **falsa**	Decisão correta	Erro tipo II

Tabela 2.1: Desfechos possíveis em um teste de hipóteses

A decisão acerca da rejeição ou não rejeição de H_0 se dá por meio da avaliação de uma estatística de teste, uma região crítica e um valor crítico. A estatística de teste é um valor obtido através de operações da estimativa do parâmetro de interesse e, em alguns casos, envolve outras quantidades vindas da amostra. Esta estatística segue uma distribuição de probabilidade e esta distribuição é usada para definir a região e o valor crítico.

Considerando a distribuição da estatística de teste, define-se um conjunto de valores que podem ser assumidos pela estatística de teste para os quais rejeita-se a hipótese nula, a chamada região de rejeição. Já o valor crítico é o valor que divide a área de rejeição da área de não rejeição de H_0. Caso a estatística de teste esteja dentro da região crítica, significa que as evidências amostrais apontam para a rejeição de H_0. Por outro lado, se a estatística de teste estiver fora da região crítica, quer dizer que os dados apontam para uma não rejeição de H_0. O já mencionado nível de significância (α) tem importante papel no processo, pois trata-se de um valor fixado e, reduzindo o nível de significância, torna-se cada vez mais difícil rejeitar a hipótese nula.

O último conceito importante para compreensão do procedimento geral de testes de hipóteses é chamado de nível descritivo, p-valor ou ainda α^*. Basicamente, trata-se da probabilidade de a estatística de teste tomar um valor igual ou mais extremo do que aquele que foi observado, supondo que a hipótese nula é verdadeira. Deste modo, o p-valor pode ser visto como uma quantidade que fornece informação quanto ao grau que os dados vão contra a hipótese nula.

Esta quantidade pode ainda ser utilizada como parte da regra decisão, uma vez que um p-valor menor que o nível de significância sugere que há evidência nos dados em favor da rejeição da hipótese nula.

Assim, o procedimento geral para condução de um teste de hipóteses consiste em:

1. Definir H_0 e H_1.
2. Identificar o teste a ser efetuado, sua estatística de teste e distribuição.
3. Obter as quantidades necessárias para o cálculo da estatística de teste.
4. Fixar o nível de significância.
5. Definir o valor e a região crítica.
6. Confrontar o valor e região crítica com a estatística de teste.
7. Obter o p-valor.
8. Concluir pela rejeição ou não rejeição da hipótese nula.

2.2.2 Testes de hipóteses em modelos de regressão

Em modelos de regressão modela-se parâmetros de distribuições de probabilidade como uma função de outras variáveis. Basicamente, o parâmetro de interesse é reescrito como uma combinação linear de novos parâmetros associados a vetores numéricos que contém o valor de variáveis explicativas.

Os parâmetros desta combinação linear são estimados com base nos dados e, como estão associados a variáveis explicativas, pode ser de interesse verificar se a retirada de uma ou mais variáveis do modelo gera um modelo significativamente pior que o original. Em outros termos, uma hipótese de interesse costuma ser verificar se há evidência suficiente nos dados para afirmar que determinada variável explicativa não possui efeito sobre a resposta.

Neste contexto, testes de hipóteses são amplamente empregados, sendo que, quando se trata de modelos de regressão, três testes são usualmente utilizados: o teste da razão de verossimilhanças, o teste Wald e o teste multiplicador de lagrange, também conhecido como teste escore. Engle (1984) descreve a formulação geral dos três testes.

O teste da razão de verossimilhanças (Wilks, 1938) busca comparar a verossimilhança de dois modelos ajustados: um deles com todas as variáveis explicativas e outro sem alguma ou algumas destas variáveis. O objetivo consiste em verificar se a diferença entre estas verossimilhanças sugere que a retirada das variáveis não gera um modelo pior.

O teste Wald (Wald, 1943) necessita de apenas um modelo e avalia-se o quão distante, com base na verossimilhança, o valor estimado está de valor postulado, isto é, um valor sob hipótese nula. Em geral testa-se o quão distante a estimativa do parâmetro de interesse está de zero, contudo pode-se executar o teste utilizando outros valores.

O teste do multiplicador de lagrange ou teste escore (Aitchison e Silvey, 1958), (Silvey, 1959), (Rao, 1948) também requer apenas um modelo ajustado. Contudo este modelo não possui determinadas variáveis explicativas e estima-se, com base na inclinação da função de verossimilhança se a adição destas variáveis ao modelo gera uma melhora significativa.

Com isto, apresentamos nesta seção uma visão geral destes três testes com suas características, hipóteses, estatísticas de teste e similaridades.

2.2.2.1 Teste da razão de verossimilhanças

Considere dois modelos de regressão ajustados para os quais as estimativas dos parâmetros foram obtidas por meio da maximização da função de verossimilhança. O primeiro destes modelos (M_1) é chamado de modelo irrestrito, isto é, com todas as variáveis explicativas. Já o segundo modelo (M_0) é chamado de modelo restrito, isto é, sem alguma(s) da(s) variável(is) explicativa(s). Considere que M_0 possui q parâmetros a menos que M_1.

Denotamos L_1 a verossimilhança do modelo M_1 e L_0 a verossimilhança do modelo M_0, ou seja, as verossimilhanças dos modelos restrito e irrestrito. Com isso, o teste consiste em verificar se a diferença entre as verossimilhanças dos modelos é grande ao ponto de afirmar que a retirada das variáveis tal como feita no modelo restrito gera um modelo com pior ajuste.

A razão L_0/L_1 resulta em um valor que sempre estará entre 0 e 1. Quanto mais próxima de 1 for esta razão, maior é a evidência a favor da hipótese que a verossimilhança dos modelos não difere. Com isso em mente, caso tivermos em mãos dois modelos ajustados, um deles com um número de maior de parâmetros e outro com número menor com verossimilhança próxima, deve-se optar pelo modelo menor.

Se denotarmos por $\boldsymbol{\beta}$ um vetor de parâmetros de regressão e $\mathbf{0}$ o vetor nulo. As hipóteses a serem testadas podem ser descritas como:

$$H_0 : \boldsymbol{\beta} = \mathbf{0} \;\; vs \;\; H_1 : \boldsymbol{\beta} \neq \mathbf{0},$$

a estatística de teste é dada por:

$$LRT = -2log\,(L_0/L_1),$$

em que LRT $\sim \chi^2_q$. É possível notar que quanto menor for a razão de verossimilhanças maior será a estatística de teste, ou seja, mais provável será a rejeição de H_0 pois L_0 e L_1 se afastam.

2.2.2.2 Teste Wald

Considere um único modelo de regressão ajustado com p parâmetros obtidas por meio da maximização da função de verossimilhança. $\boldsymbol{\beta}$ representa o vetor de parâmetros de regressão deste modelo, em que as estimativas são dadas por $\hat{\boldsymbol{\beta}}$.

O teste Wald avalia a distância entre as estimativas dos parâmetros e um conjunto de valores postulados. Esta diferença é ainda padronizada por medidas de precisão das estimativas do parâmetros testados. Quanto mais distante de 0 for o valor da distância padronizada, menores são as evidências a favor da hipótese de que os valores estimados são iguais aos valores postulados.

Com isso, a ideia do teste consiste em verificar se existe evidência suficiente nos dados para afirmar que um ou mais parâmetros são iguais a valores especificados. Em geral, os valores especificados se tratam de um vetor nulo para verificar se há evidência para afirmar que os valores dos parâmetros são iguais a 0, contudo existe a possibilidade de especificar hipóteses para qualquer valor.

Para definir as hipóteses do teste Wald, considere que há interesse em testar q restrições ao modelo original, isto é, há interesse em avaliar q dos p parâmetros originais. As hipóteses são especificadas por meio de uma matriz $\boldsymbol{L}$ de dimensão qxp e um vetor $\boldsymbol{c}$ de valores postulados de dimensão q. Com base nestes elementos, as hipóteses podem ser descritas como:

$$H_0 : \boldsymbol{L\beta} = \boldsymbol{c} \;\; vs \;\; H_1 : \boldsymbol{L\beta} \neq \boldsymbol{c},$$

a estatística de teste é dada por:

$$WT = (\boldsymbol{L\hat{\beta} - c})^T \, (\boldsymbol{L \, Var^{-1}(\hat{\beta}) \, L^T})^{-1} \, (\boldsymbol{L\hat{\beta} - c}),$$

em que $WT \sim \chi^2_q$. Note que a estatística de teste necessita de elementos que devem ser especificados pelo pesquisador e quantidades facilmente obtidas após ajuste do modelo: as estimativas dos parâmetros e a matriz de variância e covariância das estimativas.

2.2.2.3 Teste escore

Considere um único modelo de regressão ajustado com p parâmetros obtidas por meio da maximização da função de verossimilhança. $\boldsymbol{\beta}$ representa o vetor de parâmetros de regressão deste modelo, em que as estimativas são dadas por $\boldsymbol{\hat{\beta}}$.

O teste do multiplicador de lagrange ou teste escore, tal como o teste Wald, requer apenas um modelo ajustado. No caso do teste escore o modelo ajustado não possui o parâmetro de interesse e o que é feito é testar se adicionar esta variável omitida resultará em uma melhora significativa no modelo. Isto é feito com base na inclinação da função de verossimilhança.

As hipóteses do teste são dadas por:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{0} \;\; vs \;\; H_1 : \boldsymbol{\beta} \neq \boldsymbol{0},$$

a estatística de teste é dada por:

$$LMT = \boldsymbol{S'(\hat{\beta}) \, Var(\hat{\beta}) \, S(\hat{\beta})},$$

em que $S(\hat{\beta})$ representa a função escore e $\boldsymbol{Var(\hat{\beta})}$ a matriz de variâncias avaliadas sob o modelo restrito (H_0). $LMT \sim \chi^2_q$, em que q representa o número de parâmetros fixados sob H_0.

2.2.2.4 Visão geral

Os três testes podem ser utilizados para avaliar restrições aos parâmetros de modelos lineares no sentido de avaliar se a retirada de variáveis explicativas do modelo reduz significativamente o ajuste. No caso do teste de razão de verossimilhanças, dois modelos encaixados precisam ser ajustados a fim de verificar a diferença entre eles. Já o teste Wald e o escore necessitam de apenas um modelo. Além disso os testes são assintóticamente equivalentes. Em amostras finitas estes testes podem apresentar resultados diferentes de tal modo que a estatística de teste WT é maior que a estatística LRT que, por sua vez, é maior que LMT (Evans e Savin, 1982).

A figura Figura 2.1 ilustra o que cada um dos três testes faz. No eixo horizontal estão representados os valores possíveis de um parâmetro de interesse. No eixo vertical estão os valores do logaritmo da verossimilhança para cada valor possível do parâmetro de interesse. O teste da razão de verossimilhanças (LRT) busca comparar os logaritmos das verossimilhanças entre um modelo em que restringe-se o valor do parâmetro em zero, com um modelo em que o parâmetro é estimado. Esta comparação se dá através das alturas das verossimilhanças para os dois modelos, se essa distância for pequena, há indício de que há pouca diferença entre os modelos. O teste Wald (WT) busca comparar a estimativa de máxima verossimilhança do parâmetro com um valor postulado sob hipótese nula. Caso esta diferença seja pequena, há indício de que o valor da estimativa original não difere estatísticamente do valor postulado. Já o teste escore (LMT) avalia a inclinação na curva que descreve o logaritmo da verossimilhança quando o parâmetro de interesse é fixado em zero. Ou seja, avalia-se a mudança na verossimilhança no valor hipotético do parâmetro.

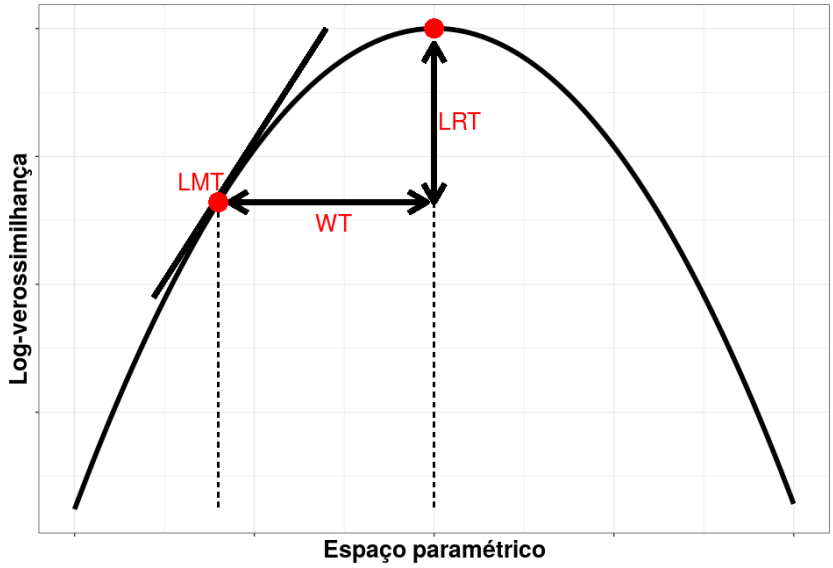

Figura 2.1: Representação gráfica do teste da razão de verossimilhanças (LRT), teste Wald (WT) e teste escore (LMT)

2.2.3 ANOVA e MANOVA

Quando trabalhamos com modelos univariados, uma das formas de avaliar a significância de cada uma das variáveis de uma forma procedural é através da análise de variância (ANOVA) (Fisher e Mackenzie, 1923). Este método consiste em efetuar testes de hipóteses sucessivos impondo restrições ao modelo original. O objetivo é testar se a ausência de determinada variável gera um modelo significativamente inferior que o modelo com determinada variável. Os resultados destes sucessivos testes são sumarizados numa tabela: o chamado quadro de análise de variância. Em geral, este quadro contêm em cada linha: a variável, o valor de uma estatística de teste referente à hipótese de nulidade de todos os parâmetros associados à esta variável, os graus de liberdade desta hipótese, e um p-valor associado à hipótese testada naquela linha do quadro.

Trata-se de um interessante procedimento para avaliar a relevância de uma variável ao problema, contudo, cuidados devem ser tomados no que diz respeito à forma como o quadro foi elaborado. Como já mencionado, cada linha do quadro refere-se a uma hipótese e estas hipóteses podem ser formuladas de formas distintas. Formas conhecidas de se elaborar o quadro são as chamadas ANOVAs do tipo I, II e III. Esta nomenclatura vem do software estatístico SAS (Institute, 1985), contudo as implementações existentes em outros softwares que seguem esta nomenclatura não necessariamente correspondem ao que está implementado no SAS. No software R (R Core Team, 2020) as implementações dos diferentes tipos de análise de variância podem ser obtidas e usadas no pacote *car* (Fox e Weisberg, 2019).

Geralmente, no contexto de modelos de regressão, para gerar o quadro de análise de variância, uma sequência de testes de testes da razão de verossimilhanças são realizados para avaliar o efeito de cada variável explicativa do modelo. Contudo é possível gerar quadros de análise de variância através do teste Wald.

Do mesmo modo que é feito para um modelo univariado, podemos chegar também a uma análise de variância multivariada (MANOVA) realizando sucessivos testes de hipóteses nos quais existe o interesse em avaliar o efeito de determinada variável em todas as respostas simultaneamente. Portanto, a pergunta que a ser respondida seria: esta variável tem efeito diferente de 0 para todas as respostas? A MANOVA clássica (Smith et al., 1962) é um assunto com vasta discussão na literatura e possui diversas propostas com o objetivo de verificar a nulidade dos parâmetros de um modelo de regressão multivariado, como o λ de Wilk's (Wilks, 1932), traço de Hotelling-Lawley (Lawley, 1938); (Hotelling, 1951), traço de Pillai (Pillai et al., 1955) e maior raiz de Roy (Roy, 1953). Tal como no caso univariado basta, para cada linha do quadro de análise de variância, especificar corretamente uma matriz $\boldsymbol{L}$ que represente de forma adequada a hipótese a ser testada.

3 PROPOSTA: TESTE WALD EM MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA

Tal como descrito no Capítulo 2, a construção do teste Wald é baseada nas estimativas de máxima verossimilhança. Porém, ao avaliar a estatística de teste é possível verificar que ela não faz uso explícito da função de verossimilhança, e sim de um vetor de estimativas dos parâmetros e uma matriz de variância e covariância destas estimativas. Assim, por mais que os McGLMs não sejam ajustados com base na maximização da função de verossimilhança para obtenção dos parâmetros do modelo, o método de estimação apresentado no Capítulo 2 fornece os componentes necessários para uma adaptação do teste.

Sendo assim, das três opções clássicas de testes de hipóteses comumente aplicados a problemas de regressão (razão de verossimilhanças, Wald e escore), o teste Wald se torna o mais atrativo no contexto dos McGLMs pois é o mais simples de se adaptar. Outra vantagem do teste Wald em relação a seus concorrentes é que existe a possibilidade de formular hipóteses para testar qualquer valor. Quando se trata dos McGLMs, esta ideia se torna especialmente atrativa pois forncece ferramentas para avaliar os parâmetros de potência.

Quando trabalhamos na classe dos McGLMs estimamos parâmetros de regressão, dispersão e potência. Os parâmetros de regressão são aqueles que associam a variável explicativa à variável resposta, através do estudo destes parâmetros é possível avaliar o efeito da variável explicativa sobre a resposta. Já os parâmetros de dispersão estão associados ao preditor matricial, através destes parâmetros pode-se avaliar o efeito da correlação entre unidades do estudo. E os parâmetros de potência nos fornecem um indicativo de qual distribuição de probabilidade melhor se adequa ao problema de acordo com a função de variância escolhida.

Com isso, nosso objetivo consiste em adaptar o teste Wald para realização de testes de hipóteses gerais sobre qualquer parâmetro dos McGLMs, sejam eles de regressão, dispersão ou potência. Com base nesta adaptação, temos ainda como objetivo chegar a procedimentos análogos às análises de variância e análises de variâncias multivariadas para parâmetros de regressão e ainda estender o conceito para parâmetros de dispersão.

Nossa adaptação visa uma de responder questões comuns no contexto de modelagem, como: quais variáveis influenciam a resposta? Existe efeito da estrutura de correlação entre indivíduos no estudo? Qual a distribuição de probabilidade que melhor se adequa ao problema? O efeito de determinada variável é o mesmo independente da resposta? Dentre outras.

Vale ressaltar que por si só, os McGLMs já contornam importantes restrições encontradas nas classes clássicas de modelos, como a impossibilidade de modelar múltiplas respostas e modelar a dependência entre indivíduos. Nossa contribuição vai no sentido de fornecer ferramentas para uma melhor interpretação dos parâmetros estimados e assim extrair mais informações e conclusões a respeito dos problemas modelados através da classe.

3.1 HIPÓTESES E ESTATÍSTICA DE TESTE

Considere um McGLM com h parâmetros estimados, sejam eles de regressão, dispersão e potência. Seja $\boldsymbol{L}$ uma matriz de especificação de hipóteses a serem testadas, de dimensão $s \times h$, $\boldsymbol{\theta_{\beta,\tau,p}}$ um vetor de dimensão $h \times 1$ de parâmetros de regressão, dispersão e potência do modelo, $\boldsymbol{c}$ um vetor de dimensão $s \times 1$ com os valores sob hipótese nula. As hipóteses a serem testadas podem ser escritas como:

$$H_0 : \boldsymbol{L\theta_{\beta,\tau,p}} = \boldsymbol{c} \; vs \; H_1 : \boldsymbol{L\theta_{\beta,\tau,p}} \neq \boldsymbol{c}, \tag{3.1}$$

considere $\hat{\boldsymbol{\theta}}_{\boldsymbol{\beta,\tau,p}}$ um vetor de dimensão $h \times 1$ com todas as estimativas dos parâmetros de regressão, dispersão e potência do modelo e $J^{-1}_{\boldsymbol{\beta,\tau,p}}$ a inversa da matriz de informação de Godambe desconsiderando os parâmetros de correlação, de dimensão $h \times h$. A generalização da estatística de teste do teste Wald para verificar a validade de uma hipótese sobre parâmetros de um McGLM fica dada por:

$$W = (\boldsymbol{L\hat{\theta}_{\beta,\tau,p}} - \boldsymbol{c})^T \, (\boldsymbol{L} \, \boldsymbol{J^{-1}_{\beta,\tau,p}} \, \boldsymbol{L^T})^{-1} \, (\boldsymbol{L\hat{\theta}_{\beta,\tau,p}} - \boldsymbol{c}), \tag{3.2}$$

em que $W \sim \chi^2_s$, ou seja, independente do número de parâmetros testados, a estatística de teste W é um único valor que segue assintóticamente distribuição χ^2 com graus de liberdade dados pelo número de parâmetros testados, isto é, o número de linhas da matriz $\boldsymbol{L}$, denotado por s.

Cada coluna da matriz $\boldsymbol{L}$ corresponde a um dos h parâmetros do modelo e cada linha a uma hipótese. Sua construção consiste basicamente em preencher a matriz com 0, 1 e eventualmente -1 de tal modo que o produto $\boldsymbol{L\theta_{\beta,\tau,p}}$ represente corretamente a hipótese de interesse. A correta especificação de $\boldsymbol{L}$ permite testar qualquer parâmetro individualmente ou até mesmo formular hipóteses para diversos parâmetros, sejam eles de regressão, dispersão ou potência.

Em um contexto prático, após a obtenção das estimativas dos parâmetros do modelo podemos estar interessados em três tipos de hipóteses: a primeira delas diz respeito a quando o interesse está em avaliar se existe evidência que permita afirmar que apenas um único parâmetro é igual a um valor postulado; a segunda delas ocorre quando há interesse em avaliar se existe evidência para afirmar que um conjunto de parâmetros é igual a um vetor de valores postulado; já a terceira hipótese diz respeito a situações em que o analista está interessado em saber se a diferença entre os efeitos de duas variáveis é igual a 0.

Para fins de ilustração dos tipos de hipóteses mencionadas considere a situação em que deseja-se investigar se uma variável numérica x_1 possui efeito sobre duas variáveis resposta, denotadas por Y_1 e Y_2. Para tal tarefa coletou-se uma amostra com n indivíduos e para cada

indivíduo observou-se o valor de x_1, Y_1 e Y_2. Com base nos dados coletados ajustou-se um modelo bivariado, com preditor dado por:

$$g_r(\mu_r) = \beta_{r0} + \beta_{r1} x_1, \tag{3.3}$$

em que o índice r denota a variável resposta, r = 1,2; β_{r0} representa o intercepto; β_{r1} um parâmetro de regressão associado a uma variável x_1. Considere que cada resposta possui apenas um parâmetro de dispersão: τ_{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 $Z_0 = I$.

Neste cenário poderiam ser perguntas de interesse: será que a variável x_1 tem efeito apenas sobre a primeira resposta? Ou apenas sobre a segunda resposta? Será que a variável x_1 possui efeito sobre as duas respostas ao mesmo tempo? Será que o efeito da variável é o mesmo para ambas as respostas? Todas essas perguntas podem ser respondidas através de testes de hipóteses sobre os parâmetros do modelo e especificadas por meio da Equação 3.1. Nas subseções a seguir são apresentados os elementos para responder cada uma destas perguntas.

3.1.1 Exemplo 1: hipótese para um único parâmetro

Considere o primeiro tipo de hipótese: há interesse em avaliar se existe efeito da variável x_1 apenas na primeira resposta. A hipótese pode ser escrita da seguinte forma:

$$H_0 : \beta_{11} = 0 \; vs \; H_1 : \beta_{11} \neq 0. \tag{3.4}$$

Esta mesma hipótese pode ser reescrita na notação mais conveniente para aplicação da estatística do teste Wald:

$$H_0 : \boldsymbol{L\theta_{\beta,\tau,p}} = \boldsymbol{c} \; vs \; H_1 : \boldsymbol{L\theta_{\beta,\tau,p}} \neq \boldsymbol{c}, \tag{3.5}$$

em que:

- $\boldsymbol{\theta^T_{\beta,\tau,p}} = \begin{bmatrix} \beta_{10} & \beta_{11} & \beta_{20} & \beta_{21} & \tau_{11} & \tau_{21} \end{bmatrix}$.
- $\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$.
- $\boldsymbol{c} = \begin{bmatrix} 0 \end{bmatrix}$, é o valor sob hipótese nula.

Note que o vetor $\boldsymbol{\theta_{\beta,\tau,p}}$ possui seis elementos, consequentemente a matriz $\boldsymbol{L}$ contém seis colunas (uma para cada elemento) e apenas uma linha, pois apenas um único parâmetro está sendo testado. Essa única linha é composta por zeros, exceto a coluna referente ao parâmetro de interesse que recebe 1. É simples verificar que o produto $\boldsymbol{L\theta_{\beta,\tau,p}}$ representa a hipótese de interesse inicialmente postulada. Com isso, a distribuição assintótica do teste é χ^2_1

3.1.2 Exemplo 2: hipótese para múltiplos parâmetros

Suponha agora que o interesse neste problema genérico não é mais testar o efeito da variável explicativa apenas em uma resposta. Suponha que o interesse é avaliar se existe evidência suficiente para afirmar que há efeito da variável explicativa x_1 em ambas as respostas simultâneamente. Neste caso teremos que testar 2 parâmetros: β_{11}, que associa x_1 à primeira resposta; e β_{21}, que associa x_1 à segunda resposta. Podemos escrever a hipótese da seguinte forma:

$$H_0 : \beta_{r1} = 0 \; vs \; H_1 : \beta_{r1} \neq 0, \tag{3.6}$$

ou, de forma equivalente:

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

A hipótese pode ainda ser reescrita na notação conveniente para o teste Wald:

$$H_0 : \boldsymbol{L\theta_{\beta,\tau,p}} = \boldsymbol{c} \; vs \; H_1 : \boldsymbol{L\theta_{\beta,\tau,p}} \neq \boldsymbol{c}, \tag{3.7}$$

em que:

- $\boldsymbol{\theta^T_{\beta,\tau,p}} = \begin{bmatrix} \beta_{10} & \beta_{11} & \beta_{20} & \beta_{21} & \tau_{11} & \tau_{21} \end{bmatrix}$.
- $\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$.
- $\boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, é o valor sob hipótese nula.

O vetor $\boldsymbol{\theta_{\beta,\tau,p}}$ se mantém com seis elementos e a matriz $\boldsymbol{L}$ com seis colunas. Neste caso estamos testando dois parâmetros, portanto a matriz $\boldsymbol{L}$ possui duas linhas. Novamente, essas linhas são compostas por zeros, exceto nas colunas referentes ao parâmetro de interesse. É simples verificar que o produto $\boldsymbol{L\theta_{\beta,\tau,p}}$ representa a hipótese de interesse inicialmente postulada. Com isso, a distribuição assintótica do teste é χ^2_2.

3.1.3 Exemplo 3: hipótese de igualdade de parâmetros

Suponha que a hipótese de interesse não envolve testar se o valor do parâmetro é igual a um valor postulado mas sim verificar se, no caso deste problema genérico, o efeito da variável x_1 é o mesmo independente da resposta. Nesta situação formularíamos uma hipótese de igualdade entre os parâmetros, ou em outros termos, se a diferença dos efeitos é nula:

$$H_0 : \beta_{11} - \beta_{21} = 0 \; vs \; H_1 : \beta_{11} - \beta_{21} \neq 0, \tag{3.8}$$

esta hipótese pode ser reescrita na seguinte notação:

$$H_0 : \boldsymbol{L\theta_{\beta,\tau,p}} = \boldsymbol{c} \; vs \; H_1 : \boldsymbol{L\theta_{\beta,\tau,p}} \neq \boldsymbol{c},$$

em que:

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

Como existe apenas uma hipótese, a matriz $\boldsymbol{L}$ possui apenas uma linha. Para a matriz $\boldsymbol{L}$ ser corretamente especificada no caso de uma hipótese de igualdade precisamos colocar 1 na coluna referente a um parâmetro, e -1 na coluna referente ao outro parâmetro, de tal modo que o produto $\boldsymbol{L\theta_{\beta,\tau,p}}$ representa a hipótese de interesse inicialmente postulada. Neste caso, a distribuição assintótica do teste é χ^2_1.

3.1.4 Exemplo 4: hipótese sobre parâmetros de regressão ou dispersão para respostas sob mesmo preditor

A Equação 3.3 descreve um modelo bivariado genérico. É importante notar que neste exemplo ambas as respostas estão sujeitas às mesmas combinações lineares de parâmetros de regressão e dispersão, isto é, o preditor para Y_1 e Y_2 é idêntico. Na prática, quando se trata dos McGLMs, preditores diferentes podem ser especificados entre variáveis respostas. Deste modo, o que foi exposto na Subseção 3.1.2 serve para qualquer caso em que haja interesse em testar hipóteses sobre mais de um parâmetro do modelo, sejam eles na mesma resposta ou em respostas diferentes ou ainda respostas sob diferentes preditores.

Nos casos em que as respostas estão sujeitas ao mesmo preditor e as hipóteses sobre os parâmetros de regressão ou dispersão a serem testadas não se alteram de resposta para resposta, uma especificação alternativa do procedimento é utilizando o produto Kronecker para comportar as hipóteses sobre as múltiplas respostas tal como utilizado em Bonat et al. (2020).

Suponha que, neste exemplo, as hipóteses de interesse seguem sendo sendo escritas tal como na Equação 3.6. Contudo, como se trata de um modelo bivariado, com mesmo preditor, a hipótese de interesse é igual entre respostas e envolve apenas parâmetros de regressão, torna-se conveniente escrever a matriz $\boldsymbol{L}$ como o produto Kronecker de duas matrizes: uma matriz $\boldsymbol{G}$ e uma $\boldsymbol{F}$, ou seja, $\boldsymbol{L} = \boldsymbol{G} \otimes \boldsymbol{F}$. A matriz $\boldsymbol{G}$ tem dimensão $R \times R$ e especifica as hipóteses referentes às respostas, já a matriz $\boldsymbol{F}$ 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 $\boldsymbol{L}$ tem dimensão $(s'R \times h)$.

A matriz $\boldsymbol{G}$ é uma matriz identidade de dimensão igual ao número de respostas analisadas no modelo. Enquanto que a matriz $\boldsymbol{F}$ equivale a uma matriz $\boldsymbol{L}$ 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 $\boldsymbol{F}$ seja testada nas R respostas do modelo.

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

- $\boldsymbol{\beta^T} = \begin{bmatrix}\beta_{10} & \beta_{11} & \beta_{20} & \beta_{21}\end{bmatrix}$: os parâmetros de regressão do modelo.
- $\boldsymbol{G} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$: matriz identidade com dimensão dada pelo número de respostas.
- $\boldsymbol{F} = \begin{bmatrix}0 & 1\end{bmatrix}$: equivalente a um $\boldsymbol{L}$ para uma única resposta.
- $\boldsymbol{L} = \boldsymbol{G} \otimes \boldsymbol{F} = \begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}$: matriz de especificação das hipóteses sobre todas as respostas.
- $\boldsymbol{c} = \begin{bmatrix}0\\ 0\end{bmatrix}$, é o valor sob hipótese nula.

Assim, o produto $\boldsymbol{L\beta}$ representa a hipótese de interesse inicialmente postulada. Neste caso, a distribuição assintótica do teste é χ^2_2. O procedimento é facilmente generalizado quando há interesse em avaliar uma hipótesse sobre os parâmetros de dispersão. Esta especificação é bastante conveniente para a geração de quadros de análise de variância.

3.2 ANOVA E MANOVA VIA TESTE WALD

Mostramos nos exemplos como é possível testar qualquer parâmetro de um McGLM seja ele de regressão, dispersão ou potência. É possível testar hipóteses sobre parâmetros individualmente, formular hipóteses para múltiplos parâmetros, formular hipóteses para combinações entre parâmetros e ainda testar valores diferentes de zero. Como explicitado nos exemplos, basta uma correta especificação da matriz $\boldsymbol{L}$. Independente do número de parâmetros testados, a estatística de teste W é um único valor que segue assintóticamente distribuição χ^2 em que os graus de liberdade são dados pelo número de hipóteses, isto é, o número de linhas da matriz $\boldsymbol{L}$.

Com base na adaptação do teste Wald para aplicação a McGLMs, buscamos propor neste trabalho três diferentes procedimentos para geração de quadros de ANOVA e MANOVA para parâmetros de regressão e um procedimento para parâmetros de dispersão de um dado modelo, seguimos a nomenclatura tipos I, II e III. 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 dos testes feitos por cada tipo das análise de variância proposta, considere a situação em que deseja-se investigar se duas variáveis numéricas (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 indivíduos e para cada indivíduo observou-se o valor de x_1, x_2, Y_1 e Y_2. Com base nos dados coletados ajustou-se um modelo bivariado, com preditor dado por:

$$g_r(\mu_r) = \beta_{r0} + \beta_{r1}x_1 + \beta_{r2}x_2 + \beta_{r3}x_1x_2. \tag{3.9}$$

em que o índice r denota a variável resposta, r = 1,2; β_{r0} representa o intercepto; β_{r1} um parâmetro de regressão associado a uma variável x_1, β_{r2} um parâmetro de regressão associado a uma variável x_2 e β_{r3} um parâmetro de regressão associado a interação entre x_1 e x_2. Considere que cada resposta possui apenas um parâmetro de dispersão: τ_{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 $Z_0 = I$.

3.2.1 ANOVA e MANOVA tipo I

Nossa proposta de análise de variância do tipo I para os McGLMs realiza testes sobre os parâmetros de regressão de forma sequencial. Neste cenário, os seguintes testes seriam efetuados:

1. Testa se todos os parâmetros são iguais a 0.
2. Testa se todos os parâmetros, exceto intercepto, são iguais a 0.
3. Testa se todos os parâmetros, exceto intercepto e os parâmetros referentes a x_1, são iguais a 0.
4. Testa se todos os parâmetros, exceto intercepto e os parâmetros referentes a x_1 e x_2, são iguais a 0.

Cada um destes testes seria uma linha do quadro de análise de variância. No caso da ANOVA seria gerado um quadro por resposta, no caso da MANOVA um quadro em que as hipóteses são testadas para ambas as respostas. Este procedimento pode ser chamado de sequencial pois a cada linha é acrescentada uma variável. Em geral, justamente por esta sequencialidade, se torna difícil interpretar os efeitos das variáveis pela análise de variância do tipo I. Em contrapartida, as análises do tipo II e III testam hipóteses que são, geralmente de maior interesse.

3.2.2 ANOVA e MANOVA tipo II

Nossa análise de variância do tipo II efetua testes similares ao último teste da análise de variância sequencial. Em um modelo sem interação o que é feito é, em cada linha, testar o modelo completo contra o modelo sem uma 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.

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. 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.

Note que nas linhas em que se busca entender o efeito de x_1 e x_2 a interação também é avaliada, pois retira-se do modelo todos os parâmetros que envolvem aquela variável.

3.2.3 ANOVA e MANOVA tipo III

Na análise de variância do tipo II são feitos testes comparando o modelo completo contra o modelo sem todos os parâmetros que envolvem determinada variável (sejam efeitos principais ou interações). Já nossa análise de variância do tipo III considera o modelo completo contra o modelo sem determinada variável, seja ela efeito principal ou de interação. Deste modo, cuidados devem ser tomados nas conclusões pois uma variável não ter efeito constatado como efeito principal não quer dizer que não haverá efeito de interação. Considerando o preditor exemplo, a análise de variância do tipo III faria os seguintes testes:

1. Testa se o intercepto é igual a 0.
2. Testa se os parâmetros de efeito principal referentes a x_1 são iguais a 0. Ou seja, é avaliado o impacto da retirada de x_1 nos efeitos principais do modelo. Neste caso, diferente do tipo II, nada se supõe a respeito do parâmetro de interação, por mais que envolva x_1.
3. Testa se os parâmetros de efeito principal referentes a x_2 são iguais a 0. Ou seja, é avaliado o impacto da retirada de x_2 nos efeitos principais do modelo. Novamente, diferente do tipo II, nada se supõe a respeito do parâmetro de interação, por mais que envolva x_2.
4. Testa se o efeito de interação é 0.

Note que nas linhas em que se testa o efeito de x_1 e x_2 mantém-se o efeito da interação, diferentemente do que é feito na análise de variância do tipo II. É importante notar que que as análises de variância do tipo II e III tal como foram propostas nesse trabalho geram os mesmos resultados quando aplicadas a modelos sem efeitos de interação. Além disso, generalizamos o procedimento tipo III para lidar com parâmetros de dispersão.

4 RESULTADOS PRELIMINARES, PENDÊNCIAS E CRONOGRAMA

Este capítulo é destinado à apresentação das funções já implementadas, apresentação das tarefas a serem cumpridas até o fim do mestrado a cronograma alvo a ser seguido até a defesa.

4.1 FUNÇÕES IMPLEMENTADAS

No capítulo Capítulo 3 vimos como chegar a um procedimento para realização de testes de hipóteses sobre qualquer parâmetro ou combinação de parâmetros de um McGLM. Deste modo um dos objetivos deste trabalho consiste em implementar tais testes no software R (R Core Team, 2020) com o intuito de complementar as já possíveis análises permitidas pelo pacote *mcglm* (Bonat, 2018).

No que diz respeito à implementações do teste Wald em outros contextos no R, o pacote *lmtest* (Zeileis e Hothorn, 2002) possui uma função genérica para realizar testes de Wald para comparar modelos lineares e lineares generalizados aninhados. Já o pacote *survey* (Lumley, 2020); (Lumley, 2004);(Lumley, 2010) possui uma função que realiza teste de 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 e 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 às tabelas de análise de variância, o R possui a função anova no pacote padrão *stats* (R Core Team, 2020) aplicável a modelos lineares e lineares generalizados. Já o pacote *car* (Fox e Weisberg, 2019) possui uma função que retorna quadros de análise de variância dos tipos II e III para diversos modelos.

Contudo, quando se trata de modelos multivariados de covariância linear generalizada ajustados no pacote *mcglm*, não existem opções para realização de testes de hipóteses lineares gerais nem de análises de variância utilizando a estatística de Wald.

Deste modo, baseando-nos nas funcionalidades do pacote *car* (Fox e Weisberg, 2019), implementamos funções que permitem a realização de análises de variância por variável resposta (ANOVA), bem como análises de variância multivariadas (MANOVA). Estas funções recebem como argumento apenas o objeto que armazena o modelo devidamente ajustado através da função *mcglm()* do pacote *mcglm*. Foram implementadas também funções que geram quadros como os de análise de variância focados no preditor linear matricial, ou seja, quadros cujo objetivo é verificar a significância dos parâmetros de dispersão.

Por fim, foi implementada uma função para hipóteses lineares gerais especificadas pelo usuário, na qual é possível testar hipóteses sobre parâmetros de regressão, dispersão ou potência.

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

Todas as funções geram resultados mostrando graus de liberdade e p-valores baseados no teste Wald aplicado aos modelos multivariados de covariância linear generalizada (McGLM). A Tabela 4.1 mostra os nomes e descrições das funções implementadas.

Função	Descrição
mc_linear_hypothesis()	Hipóteses lineares gerais especificadas pelo usuário
mc_anova_I()	ANOVA tipo I
mc_anova_II()	ANOVA tipo II
mc_anova_III()	ANOVA tipo III
mc_manova_I()	MANOVA tipo I
mc_manova_II()	MANOVA tipo II
mc_manova_III()	MANOVA tipo III
mc_anova_disp()	ANOVA tipo III para dispersão
mc_manova_disp()	MANOVA tipo III para dispersão

Tabela 4.1: Funções implementadas

A função *mc_linear_hypothesis()* é a implementação computacional em R que permite a execução de qualquer um dos testes apresentados no Capítulo 3. 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*.

As funções *mc_anova_I()*, *mc_anova_II()* e *mc_anova_III()* são funções destinadas à avaliação 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 simples 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 Capítulo 3.

Tal como descrito no Capítulo 2, a matriz $\boldsymbol{\Omega(\tau)}$ tem como objetivo modelar a correlação existente entre linhas do conjunto de dados através do chamado preditor linear matricial. Na prática temos, para cada matriz do preditor matricial, um parâmetro de dispersão τ_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_disp()* e *mc_manova_disp()*.

A função *mc_anova_disp()* 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 repetidas tal como especificado no preditor matricial para aquela resposta. Já a função *mc_manova_disp()* 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 correlocionadas é o mesmo para todas as respostas.

Por fim, ressaltamos que as todas as funções de prefixo *mc_anova* e *mc_manova* foram implementadas no sentido de facilitar o procedimento de análise da importâncias das variáveis. Contudo, dentre as funções implementadas, a mais flexível é a função *mc_linear_hypothesis()* que implementa e da liberdade ao usuário de efetuar qualquer teste utilizando a estatística de Wald no contexto dos McGLMs. A partir desta função é possível replicar os resultados de qualquer uma das funções de análise de variância e testar hipóteses mais gerais como igualdade de efeitos, formular hipóteses com testes usando valores diferentes de zero e até mesmo formular hipóteses que combinem parâmetros de regressão, dispersão e potência quando houver alguma necessidade prática.

4.2 PENDÊNCIAS

Através da adaptação do teste Wald para o contexto dos McGLMs gostaríamos ainda de propor e implementar neste trabalho procedimentos para realização de testes de comparações múltiplas. Tais procedimentos são utilizados quando a análise de variância aponta como conclusão a existência de efeito significativo dos parâmetros associados a uma variável categórica, ou seja, há ao menos uma diferença significativa entre os níveis do fator. Com isso, o teste de comparações múltiplas é utilizado para determinar onde estão estas diferenças. Por exemplo, suponha que há no modelo uma variável categórica x de três níveis: A, B e C. A análise de variância mostrará se há efeito da variável x no modelo, isto é, se os valores da resposta estão associados aos níveis de x, contudo este resultado não nos mostrará se os valores da resposta diferem de A para B, ou de A para C, ou ainda se B difere de C. Com isso, testes de comparações múltiplas se encaixam no escopo deste trabalho no sentido de fornecer ferramentas para melhor compreender o efeito das variáveis explicativas sobre as respostas no contexto dos McGLMs.

Ainda nesta linha, outra tarefa a ser cumprida é a adequação dos testes para que sejam válidos para diferentes contrastes. Um contraste é uma combinação linear de variáveis em que a soma dos coeficientes é igual a zero, o que permite a comparação entre os níveis de uma variável categórica. Através da definição dos contrastes é possível estabelecer diferentes comparações entre os níveis de variáveis categóricas que entram no modelo como variáveis explicativas. Logo, os contrastes definem como as variáveis categóricas são tratadas nos modelos. O contraste mais comum utilizado é chamado constraste de tramento em que o primeiro nível da variável categórica é mantido como valor de referência e, para os demais níveis, mede-se a mudança para a categoria de referência; nossa adaptação e funções funcionam para este caso. Para outros tipos

de contrastes as funções necessitam de modificações. Tais modificações se encaixam no escopo do trabalho pois a definição dos contrastes define quais as comparações são de interesse em um contexto prático e nem sempre o esquema de constrastes tradicionais se encaixam ao problema.

Com estas modificações feitas, pretendemos avaliar as propriedades e comportamento dos testes propostos com base em estudos de simulação. O objetivo consiste em verificar o poder dos testes sob diferentes configurações de modelos, isto é, variando o número de respostas do modelo, distribuições da(s) resposta(s), tamanhos amostrais, correlação entre respostas, correlação entre unidades do conjunto de dados, dentre outras características a fim de verificar para quais cenários nossa proposta apresenta resultados satisfatórios, para quais cenários os resultados não são satisfatórios de que forma pode-se melhorar o desempenho da proposta nestes casos.

Por fim, visamos finalizar uma dissertação de mestrado que apresente a adaptação do teste, implementação das funções, resultados de estudos de simulação que comprovem o desempenho destas funções e ainda motivar o potencial de aplicação das metodologias discutidas com base na aplicação a conjuntos de dados reais. Considerando o grupo de pesquisa (Data Science & Big Data) o trabalho teria as seguintes contribuições: propor um teste para uma classe de modelos não usual mas com alto potencial de aplicação em contextos práticos, fornecer ferramentas para cientistas de dados que necessitem efetuar testes de hipóteses sobre parâmetros de modelos de regressão multivariados com o objetivo de compreender o impacto de variáveis explicativas sobre variáveis respostas nos mais diversos contextos, comprovar o funcionamento da proposta através de um estudo de simulação e ainda motivar o uso da proposta através de análises de conjuntos de dados reais de diferentes áreas de aplicação.

4.3 CRONOGRAMA

A Tabela 4.2 mostra o cronograma alvo a ser seguido até a defesa da dissertação para obtenção da titulação.

Tarefa	**Data de início**	**Data de finalização**
Implementação dos testes de comparações múltiplas	17/08	17/09
Adaptação das funções para diferentes contrastes	17/09	17/10
Desenho e execução do estudo de simulação	17/10	17/11
Análise de dados	17/11	17/12
Sumarização dos resultados	17/12	17/01
Entrega e defesa da dissertação	17/01	17/02

Tabela 4.2: Cronograma para cumprimento das pendências para titulação.

REFERÊNCIAS

Aitchison, J. e Silvey, S. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The annals of mathematical Statistics*, páginas 813–828.

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

Azzalini, A. (2017). *Statistical inference: Based on the likelihood*. Routledge.

Barndorff-Nielsen, O. E. e Cox, D. R. (2017). *Inference and asymptotics*. Routledge.

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

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

Bonat, W. H., Petterle, R. R., Balbinot, P., Mansur, A. e Graf, R. (2020). Modelling multiple outcomes in repeated measures studies: Comparing aesthetic eyelid surgery techniques. *Statistical Modelling*, página 1471082X20943312.

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

Cao, L. (2016). Data science and analytics: a new era.

Cordeiro, G. M. e Demétrio, C. G. (2008). Modelos lineares generalizados e extensões. *Piracicaba: USP*.

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

Engle, R. F. (1984). Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of econometrics*, 2:775–826.

Evans, G. e Savin, N. E. (1982). Conflict among the criteria revisited; the w, lr and lm tests. *Econometrica: Journal of the Econometric Society*, páginas 737–748.

Fisher, R. A. (1925). Statistical methods for research workers. oliver and boyd. *Edinburgh, Scotland*, 6.

Fisher, R. A. (1929). The statistical method in psychical research. Em *Proceedings of the Society for Psychical Research*, volume 39, páginas 189–192.

Fisher, R. A. (1992). The arrangement of field experiments. Em *Breakthroughs in statistics*, páginas 82–91. Springer.

Fisher, R. A. e Mackenzie, W. A. (1923). Studies in crop variation. ii. the manurial response of different potato varieties. *The Journal of Agricultural Science*, 13(3):311–320.

Fox, J. e Weisberg, S. (2019). *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

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

Institute, S. (1985). *SAS user's guide: Statistics*, volume 2. Sas Inst.

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

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

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

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

Larsen, S., Andersen, T. e Hessen, D. O. (2011). The pco2 in boreal lakes: Organic carbon as a universal predictor? *Global Biogeochemical Cycles*, 25(2).

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

Lehmann, E. L. (1993). The fisher, neyman-pearson theories of testing hypotheses: one theory or two? *Journal of the American statistical Association*, 88(424):1242–1249.

Lehmann, E. L. e Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Ley, C. e Bordas, S. P. (2018). What makes data science different? a discussion involving statistics2. 0 and computational sciences. *International Journal of Data Science and Analytics*, 6(3):167–175.

Liang, K.-Y. e Zeger, S. L. (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, M. A. (2013). A general modelling framework for multivariate disease mapping. *Biometrika*, 100(3):539–553.

Michelon, T. B., Taconeli, C. A., Vieira, E. S. N. e Panobianco, M. (2019). Dados de contagem em sementes de eucalyptus cloeziana: uma análise comparativa entre modelos estatísticos. *Ciência e Agrotecnologia*, 43.

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

Neyman, J. e Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, páginas 175–240.

Neyman, J. e Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part ii. *Biometrika*, páginas 263–294.

Neyman, J. e Pearson, E. S. (1933). Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337.

Neyman, J. e Pearson, E. S. (2020a). *On the use and interpretation of certain test criteria for purposes of statistical inference. Part I.* University of California Press.

Neyman, J. e Pearson, E. S. (2020b). *On the use and interpretation of certain test criteria for purposes of statistical inference. Part II.* University of California Press.

Paula, G. A. (2004). *Modelos de regressão: com apoio computacional*. IME-USP São Paulo.

Petterle, R. R., de Freitas, C. A., Furtado, A. M., de Carvalho, F. H. e Bonat, W. H. (2017). Comparação e aplicação de modelos de regressão binária na retenção de capacetes de motociclistas. *Revista Brasileira de Biometria*, 35(2):266–282.

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

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

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

Press, G. (2013). A very short history of data science. `https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/?sh=1c01914855cf`. Acessado em 14/04/2021.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. Em *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 44, páginas 50–57. Cambridge University Press.

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

Silvey, S. D. (1959). The lagrangian multiplier test. *The Annals of Mathematical Statistics*, 30(2):389–407.

Silvey, S. D. (2017). *Statistical inference*. Routledge.

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

St, L., Wold, S. et al. (1989). Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Weihs, C. e Ickstadt, K. (2018). Data science: the impact of statistics. *International Journal of Data Science and Analytics*, 6(3):189–194.

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

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62.

Zeileis, A. e Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, 2(3):7–10.