LINEU ALBERTO CAVAZANI DE FREITAS

TESTES DE HIPÓTESES EM MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA

(*versão pré-defesa, compilada em 26 de abril 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

O resumo deve conter no máximo 500 palavras, devendo ser justificado na largura da página e escrito em um único parágrafo[1] com um afastamento de 1,27 cm na primeira linha. O espaçamento entre linhas deve ser de 1,5 linhas. O resumo deve ser informativo, ou seja, é a condensação do conteúdo e expõe finalidades, metodologia, resultados e conclusões.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

Palavras-chave: Palavra-chave 1. Palavra-chave 2. Palavra-chave 3.

[1] E também não deve ter notas de rodapé; em outras palavras, não siga este exemplo... ;-)

ABSTRACT

The abstract should be the English translation of the “resumo”, no more, no less.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

Keywords: Keyword 1. Keyword 2. Keyword 3.

LISTA DE FIGURAS

LISTA DE TABELAS

4.1 Funções implementadas 24

SUMÁRIO

1 INTRODUÇÃO 7
2 REVISÃO DE LITARATURA 12
2.1 MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA 12
2.1.1 GLM 12
2.1.2 cGLM 13
2.1.3 McGLM 14
2.1.4 Estimação e inferência 14
2.2 TESTES DE HIPÓTESE 16
3 PROPOSTA 17
3.1 TESTE WALD NO CONTEXTO DOS MCGLM 17
3.1.1 O teste Wald 17
3.1.2 Adaptação do teste para os McGLM 17
3.1.3 Exemplos de hipóteses 18
3.1.4 Exemplo 1 19
3.1.5 Exemplo 2 20
3.1.6 Exemplo 3 21
3.1.7 ANOVA e MANOVA via teste Wald 21
4 RESULTADOS PRELIMINARES 23
4.1 FUNÇÕES IMPLEMENTADAS 23
5 CRONOGRAMA 28
REFERÊNCIAS 29

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).

O objetivo destes modelos consiste em obter uma equação que explique a relação entre as variáveis. Fazendo uso desta equação é possível ainda realizar predições da variável resposta com base nos valores observados das variáveis explicativas. De forma geral, um modelo de regressão é uma expressão matemática que relaciona a média da variável resposta às variáveis preditoras, em que a variável resposta segue uma distribuição de probabilidade condicional às covariáveis e a média é descrita por um preditor linear.

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. Deste modo pode-se usar um modelo de regressão para, por exemplo, modelar a relação entre uma variável aleatória y e um conjunto de variáveis explicativas x_1, x_2, ... , x_p. As variáveis explicativas são incorporadas ao modelo e um conjunto de parâmetros β_0, β_1, β_2, ... , β_p desconhecidos são estimados com base nos dados disponíveis em que as estimativas são denotadas por $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, ... , $\hat{\beta}_p$. Estes parâmetros determinam a relação entre as variáveis explicativas e a resposta. Sendo assim, o conhecimento a respeito da influência de uma variável explicativa x_i sobre a resposta y vem do estudo destes parâmetros. 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. 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, 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 as variáveis retiradas 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 score (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 extende-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 hipótese e também de procedimentos baeados 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.

ESPECIFICAR QUE ESTAMOS TRABALHANDO COM O TESTE WALD E POR QUAL RAZAO

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 projeto de qualificação 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 McGLM 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

2.1 MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA

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

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

Com o objetivo de contornar estas restrições, foi proposta por Bonat e Jørgensen (2016), uma estrutura geral para análise de dados não gaussianos com múltiplas respostas em que não se faz suposições quanto à independência das observações: os chamados Modelos Multivariados de Covariância Linear Generalizada (McGLM).

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

2.1.1 GLM

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

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

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

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

1. A função de variância potência, que caracteriza a família Tweedie de distribuições, em que a função de variância é dada por $\vartheta\left(\boldsymbol{\mu}; p\right) = \mu^p$, na qual destacam-se a distribuições: normal ($p = 0$), Poisson ($p = 1$), gama ($p = 2$) e normal inversa ($p = 3$). 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 $\vartheta(\boldsymbol{\mu}; p) = \mu + \mu^p$ e tem como casos particulares os mais famosos modelos para dados de contagem: Hermite ($p = 0$), Neyman tipo A ($p = 1$), binomial negativa ($p = 2$) e Poisson–inversa gaussiana (p = 3) (Jørgensen e Kokonendji, 2015).

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

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

2.1.2 cGLM

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

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

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

Com isso, substituindo a matriz identidade pela equação do preditor linear matricial, temos uma classe com toda a flexibilidade dos GLM, porém contornando a restrição da independência entre as observações desde que o preditor linear matricial seja adequadamente especificado.

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

2.1.3 McGLM

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

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

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

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 McGLM permitem a divisão dos parâmetros em dois conjuntos: $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\lambda}^\top)^\top$. Desta forma, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_R^\top)^\top$ 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, $\mathcal{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_R^\top)^\top$ denota o vetor empilhado de ordem $NR \times 1$ da matriz de variáveis resposta $\boldsymbol{Y}_{N\times R}$ e $\mathcal{M} = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_R^\top)^\top$ denota o vetor empilhado de ordem $NR \times 1$ da matriz de valores esperados $\boldsymbol{M}_{N\times R}$.

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

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

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

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

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

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

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

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

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

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

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 $\mathcal{Y}_l$. No processo de estimação dos McGLM são usadas as versões empíricas.

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

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

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

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

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

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

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

2.2 TESTES DE HIPÓTESE

3 PROPOSTA

3.1 TESTE WALD NO CONTEXTO DOS MCGLM

3.1.1 O teste Wald

O teste Wald é um teste de hipóteses amplamente difundido para análises de Modelos Lineares e Modelos Lineares Generalizados para verificar suposições sobre os parâmetros do modelo, isto é, verifcar se a estimativa do parâmetro é ou não estatísticamente igual a um valor qualquer.

A grosso modo, é um teste que avalia a distância entre a estimativa do parâmetro e o valor postulado sob a hipótese nula. Esta diferença é ainda ponderada por uma medida de precisão da estimativa do parâmetro e, quanto mais distante de 0 for o valor da distância ponderada, menor é a chance da hipótese de igualdade ser verdadeira, ou seja, do valor postulado ser igual ao valor estimado.

Além destes elementos o teste pressupõe que os estimadores dos parâmetros do modelo sigam distribuição assintótica Normal. Para avaliação da estatística de teste e verificação de significância estatística utiliza-se distribuição assintótica Qui-quadrado (χ^2).

Quando trabalhamos com modelos de regressão, estes tipos de teste são extremente úteis quando usados para avaliar o efeito das variáveis explicativas sobre a(s) variável(is) resposta do modelo. Por exemplo: se ajustarmos um modelo com uma variável resposta e uma variável explicativa numérica, vamos estimar um único parâmetro de regressão; este parâmetro associa a variável explicativa à variável resposta. Através de um teste de hipótese podemos avaliar o efeito desta variável explicativa, basta verificar se existe evidência que permita afirmar que o valor que associa as variáveis é igual a 0.

Existe também a possibilidade de formular hipóteses para mais de um parâmetro de regressão e ainda testar valores diferentes de 0, tudo depende do objetivo do estudo e do interesse do pesquisador.

3.1.2 Adaptação do teste para os McGLM

Quando trabalhamos na classe dos McGLM estimamos parâmetros de regressão, dispersão e potência. Os parâmetros de regressão são aqueles que associam a variável explicativa à variável resposta. Os parâmetros de dispersão estão associados ao preditor matricial e, em geral, cada matriz do preditor matricial diz respeito a uma estrutura de correlação existente entre as unidades amostrais do conjunto de dados, deste modo, os parâmetros de dispersão podem ser usados para avaliar se existe efeito da relação entre as unidades amostrais tal como foi especificado pelo preditor matricial. Já os parâmetros de potência nos fornecem um indicativo de qual distribuição de probabilidade melhor se adequa ao problema.

Nossa adaptação do teste Wald tradicional visa uma forma de formular e testar hipóteses para todos esses parâmetros estimados na classe dos McGLM para responder questões comuns de analistas no contexto de modelagem, como: quais variáveis influenciam a resposta? Existe efeito da estrutura de correlação entre indivíduos no meu estudo? Qual a distribuição de probabilidade que melhor se adequa ao meu problema? Dentre outras.

Vale ressaltar que por si só, o McGLM já contorna 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.

As hipóteses a serem testadas podem ser escritas como:

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

Em que $\boldsymbol{L}$ é a matriz de especificação das hipóteses a serem testadas, tem dimensão $s \times h$, $\boldsymbol{\theta_{\beta,\tau,p}}$ é o vetor de dimensão $h \times 1$ de parâmetros de regressão, dispersão e potência do modelo, $\boldsymbol{c}$ é um vetor de dimensão $s \times 1$ com os valores sob hipótese nula.

A generalização da estatística de teste do teste Wald para verificar a validade de uma hipótese sobre parâmetros de um McGLM é 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 $\boldsymbol{L}$ é a mesma matriz da especificação das hipóteses a serem testadas, tem dimensão $s \times h$; $\boldsymbol{\hat{\theta}_{\beta,\tau,p}}$ é o vetor de dimensão $h \times 1$ com todas as estimativas dos parâmetros de regressão, dispersão e potência do modelo; $\boldsymbol{c}$ é um vetor de dimensão $s \times 1$ com os valores sob hipótese nula; e $J^{-1}_{\boldsymbol{\beta,\tau,p}}$ é a inversa da matriz de informação de Godambe desconsiderando os parâmetros de correlação, de dimensão $h \times h$.

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 da matriz permite testar qualquer parâmetro individualmente ou até mesmo formular hipóteses para diversos parâmetros simultaneamente, sejam eles de regressão, dispersão ou potência. 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.

3.1.3 Exemplos de hipóteses

Em um contexto prático, um analista após a obtenção dos parâmetros do modelo pode estar interessado em 3 tipos de hipótese: 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 mais de um parâmetro simultâneamente são iguais a um vetor de valores postulado; e 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 hipótese mencionadas, considere um problema qualquer 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_{r1} 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. Como existe apenas um parâmetro de dispersão isso quer dizer que nossas unidades amostrais são independentes.

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 um teste de hipóteses sobre os parâmetros do modelo.

3.1.4 Exemplo 1

Considere o primeiro tipo de hipótese: o analista deseja saber 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 da hipótese nula.

Note que o vetor $\boldsymbol{\theta_{\beta,\tau,p}}$ possui 6 elementos, consequentemente a matriz $\boldsymbol{L}$ contém 6 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.

3.1.5 Exemplo 2

Imagine agora que o interesse neste problema genérico não é mais testar o efeito da variável explicativa apenas em uma resposta. Imagine que o analista tem interesse em 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 da hipótese nula.

O vetor $\boldsymbol{\theta_{\beta,\tau,p}}$ mantém 6 elementos e a matriz $\boldsymbol{L}$ 6 colunas. Neste caso estamos testando 2 parâmetros, portanto a matriz $\boldsymbol{L}$ possui 2 linhas. Novamente, essas linhas são composta 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.

3.1.6 Exemplo 3

Imagine agora 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:

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

É possível testar qualquer parâmetro individualmente, formular hipóteses para diversos parâmetros simultaneamente (sejam eles de regressão, dispersão, potência), formular hipóteses para combinações entre estes parâmetros e 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 $\boldsymbol{L}$, denotado por s.

3.1.7 ANOVA e MANOVA via teste Wald

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). Este método consiste em efetuar testes sucessivos impondo restrições ao modelo original. O objetivo é testar se a ausência de determinada variável gera perda ao modelo. Os resultados destes sucessivos testes são sumarizados numa tabela, o chamado quadro de análise de variância, que 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). Em geral, recomenda-se ao usuário estar seguro de qual tipo de análise está sendo utilizada pois, caso contrário, interpretações equivocadas podem ser feitas.

Testar se a ausência de determinada variável gera perda ao modelo quer dizer, em outros termos, realizar um teste para verificar a nulidade dos parâmetros que associam esta variável à resposta. Isto geralmente é feito através de uma sequência de testes de Razão de Verossimilhança, contudo é possível gerar quadros de Análise de Variância utilizando o teste Wald pois sempre estarão sendo comparados o modelo completo e o modelo sem determinada ou determinadas variáveis. Ou seja, no contexto dos McGLM basta então, 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.

Do mesmo modo que é feito para um modelo univariado, podemos chegar também a uma Análise de Variância Multivariada (MANOVA) realizando sucessivos testes do tipo Wald em que estamos interessados 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 lambda de Wilk's (Wilks, 1932), traço de Hotelling-Lawley (Lawley, 1938); (Hotelling, 1951), traço de Pillai (Pillai et al., 1955) e maior raiz de Roy (Roy, 1953). 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.

4 RESULTADOS PRELIMINARES

4.1 FUNÇÕES IMPLEMENTADAS

No capítulo anterior vimos que podemos chegar a um teste de hipóteses sobre qualquer um dos parâmetros de um McGLM (Bonat e Jørgensen, 2016). Ou seja, somos capazes de gerar conhecimento sobre problemas práticos através do estudo das estimativas dos parâmetros de modelos de uma classe em que podemos lidar com múltiplas respostas, de diferentes naturezas, modelando também a correlação entre indivíduos da amostra. Deste modo um dos objetivos deste trabalho consiste em implementar tais testes no software R (R Core Team, 2020) com o objetivo 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 já mencionado 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, etc; 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 variância dos tipos II e III para diversos modelos.

Contudo, quando se trata de Modelos Multivariados de Covariância Linear Generalizada ajustados no pacote *mcglm* (Bonat, 2018), 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). Note que no caso da MANOVA os preditores devem ser iguais para todas as respostas sob análise. 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. Estas funções recebem como argumento apenas o objeto que armazena o modelo devidamente ajustado através da função *mcglm()* do pacote *mcglm*.

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). Todas as funções implementadas podem ser acessadas em *https://github.com/lineu96/msc*. 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 do que foi exposto no **??**. É a função mais flexível que temos no conjunto de implementações. Com ela é possível especificar qualquer tipo de hipótese sobre parâmetros de regressão, dispersão ou potência de um modelo *mcglm*.

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*. Implementamos 3 tipos diferentes de análises de variância mas não necessariamente essas implementações apresentarão os mesmos resultados que as versões com nomenclatura similar destinadas a modelos univariados disponíveis em outras bibliotecas.

Para fins de ilustração dos testes feitos por cada tipo das análise de variância implementada, considere um modelo bivariado com preditor dado por:

$$g_r(\mu_r) = \beta_{r0} + \beta_{r1}x_1 + \beta_{r2}x_2 + \beta_{r3}x_1x_2. \tag{4.1}$$

Em que o índice r denota a variável resposta, r = 1,2. Temos deste modo um intercepto para cada resposta: β_{10} para a primeira e β_{20} para a segunda; temos também três parâmetros de regressão para cada resposta: β_{11} é o efeito de x_1 sobre a resposta 1, β_{21} é o efeito de x_1 sobre a resposta 2; β_{12} é o efeito de x_2 sobre a resposta 1, β_{22} é o efeito de x_2 sobre a resposta 2; por fim, β_{13} representa o efeito da interação entre as variáveis x_1 e x_2 sobre a resposta 1 e β_{23} representa o efeito da interação entre as variáveis x_1 e x_2 sobre a resposta 2. Todas as funções de análise de variância neste contexto retornariam dois quadros, um para cada resposta.

Nossa implementação de análise de variância do tipo I (*mc_anova_I()*) realiza testes sobre os parâmetros de regressão de forma sequencial. Neste cenário, nossa função faria os seguintes testes para cada resposta:

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, e pode ser chamada 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 ao analista.

Nossa análise de variância do tipo II (*mc_anova_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 para cada resposta:

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.

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 (*mc_anova_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 para cada resposta:

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 implementadas nesse trabalho geram os mesmos resultados quando aplicadas a modelos sem efeitos de interação. Além disso, o *mcglm* ajusta modelos com múltiplas respostas; deste modo, para cada resposta seria gerado um quadro de análise de variância.

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*.

Estas funções são generalizações das funções *mc_anova_I()*, *mc_anova_II()* e *mc_anova_III()*. 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.

Deste modo, em nosso exemplo, as funções de análise de variância univariadas retornariam um quadro para cada uma das respostas avaliando o efeito das variáveis para cada uma delas. Já as funções de análise de variância multivariadas retornariam um único quadro, em que avalia-se o efeito das variáveis em todas as respotas ao mesmo tempo. A sequência de testes feitos a cada linha do quadro são os mesmos mostrados para as funções de análise de variância univariadas.

Na prática, utilizando o *mcglm*, podemos ajustar modelos com diferentes preditores para as respostas, nestes casos as funções *mc_anova_I()*, *mc_anova_II()* e *mc_anova_III()* funcionam

sem problema algum. Contudo as funções *mc_manova_I()*, *mc_manova_II()* e *mc_manova_III()* necessitam que os preditores sejam iguais para todas as respostas.

Tal como descrito no **??**, a matriz $\boldsymbol{\Omega(\tau)}$ tem como objetivo modelar a correlação existente entre linhas do conjunto de dados. A matriz é descrita como uma combinação de matrizes conhecidas e tal estrutura foi batizada como preditor linear matricial. Com isso é possível especificar extensões multivariadas para diversos modelos famosos da literatura que lidam com dados em que haja alguma relação ente as unidades amostrais, como estudos de medidas repetidas, dados longitudinais, séries temporais, dados espaciais e espaço-temporais.

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 correlocionadas 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 McGLM. 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.

5 CRONOGRAMA

TO DO

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.

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

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

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

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. 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.

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.

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.

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.