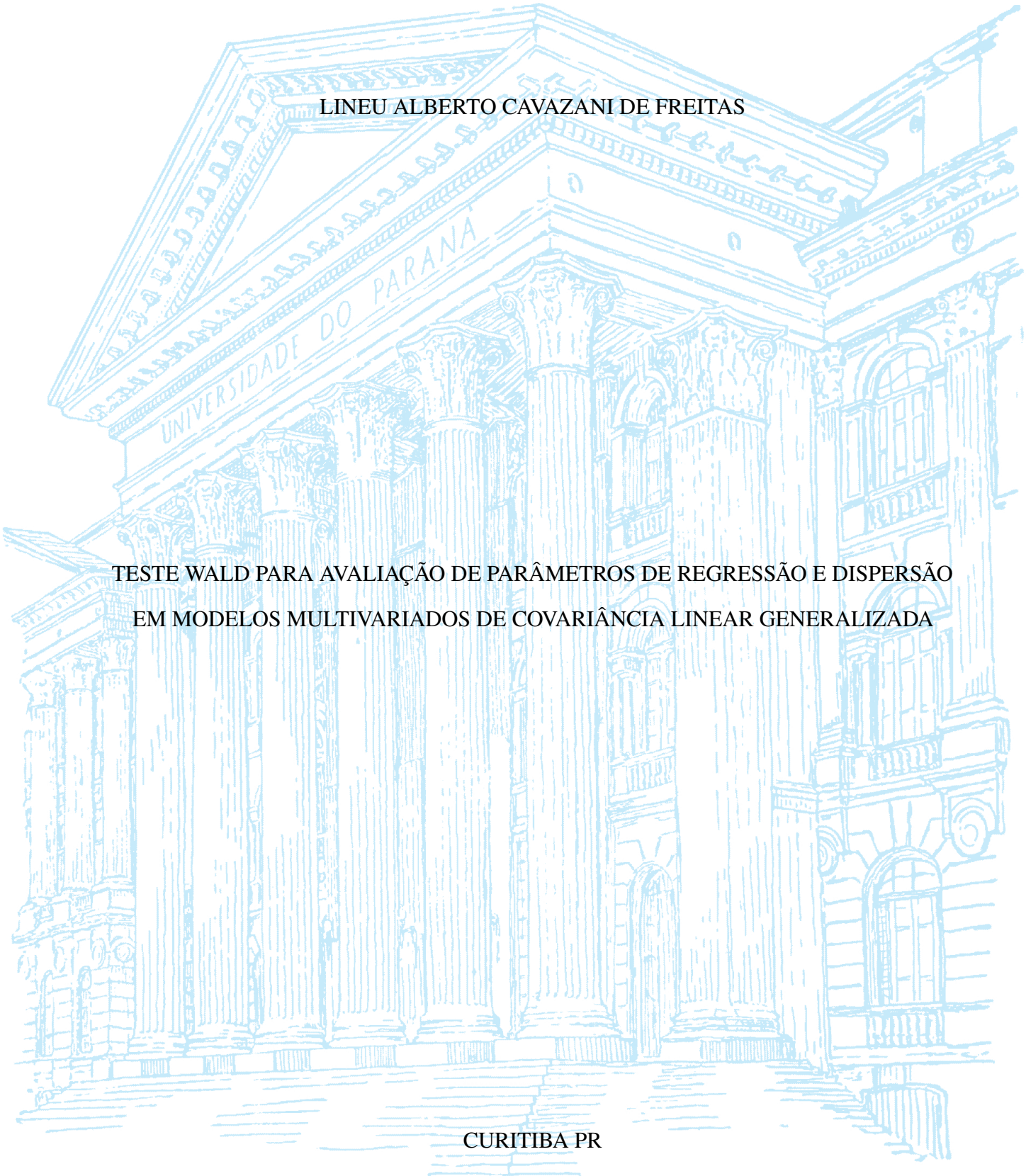
UNIVERSIDADE FEDERAL DO PARANÁ

LINEU ALBERTO CAVAZANI DE FREITAS



TESTE WALD PARA AVALIAÇÃO DE PARÂMETROS DE REGRESSÃO E DISPERSÃO EM MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA

CURITIBA PR

2022

LINEU ALBERTO CAVAZANI DE FREITAS

TESTE WALD PARA AVALIAÇÃO DE PARÂMETROS DE REGRESSÃO E DISPERSÃO EM MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática 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.

CURITIBA PR

2022

Ficha catalográfica

Substituir o arquivo `0-iniciais/catalografica.pdf` pela ficha catalográfica fornecida pela Biblioteca da UFPR (PDF em formato A4).

Instruções para obter a ficha catalográfica e fazer o depósito legal da tese/dissertação (contribuição de André Hochuli, abril 2019):

1. Estas instruções se aplicam a dissertações de mestrado e teses de doutorado. Trabalhos de conclusão de curso de graduação e textos de qualificação não precisam segui-las.
2. Verificar se está usando a versão mais recente do modelo do PPGInf e atualizar, se for necessário (`https://gitlab.c3sl.ufpr.br/maziero/tese`).
3. conferir o *checklist* de formato do Sistema de Bibliotecas da UFPR, em `https://portal.ufpr.br/teses_servicos.html`.
4. Enviar e-mail para "`referencia.bct@ufpr.br`" com o arquivo PDF da dissertação/tese, solicitando a respectiva ficha catalográfica.
5. Ao receber a ficha, inseri-la em seu documento (substituir o arquivo `0-iniciais/catalografica.pdf` do diretório do modelo).
6. Emitir a Certidão Negativa (CND) de débito junto a biblioteca (`https://www.portal.ufpr.br/cnd.html`).
7. Avisar a secretaria do PPGInf que você está pronto para o depósito. Eles irão mudar sua titulação no SIGA, o que irá liberar uma opção no SIGA pra você fazer o depósito legal.
8. Acesse o SIGA (`http://www.prppg.ufpr.br/siga`) e preencha com cuidado os dados solicitados para o depósito da tese.
9. Aguarde a confirmação da Biblioteca.
10. Após a aprovação do pedido, informe a secretaria do PPGInf que a dissertação/tese foi depositada pela biblioteca. Será então liberado no SIGA um link para a confirmação dos dados para a emissão do diploma.

Ficha de aprovação

Substituir o arquivo 0-iniciais/aprovacao.pdf pela ficha de aprovação fornecida pela secretaria do programa, em formato PDF A4.

"A tentação de formular teorias prematuras sobre dados insuficientes é a ruína da nossa profissão."Sherlock Holmes, de Sir Arthur Conan Doyle

AGRADECIMENTOS

À minha família, em especial a meus pais, Hamilton Alves de Freitas e Lúcia Elena Cavazani de Freitas, que mesmo atuando distante de sua realidade nunca deixaram de estar ao meu lado.

Ao meu orientador, Professor Doutor Wagner Hugo Bonat, que me acompanha nesta jornada desde muito antes do ingresso na pós graduação.

Ao professor Professor Doutor Walmes Marques Zeviani, apoiador e conselheiro deste e outros projetos desde a graduação.

Aos professores Wagner e Walmes também agradeço pela confiança depositada em todos os projetos que participei na Ômega - Escola de Data Science.

Aos professores Cesar Augusto Taconeli e José Luiz Padilha da Silva, meus primeiros orientadores em minha trajetória acadêmica e grandes apoiadores da sequência da minha carreira com foco em pesquisa e docência.

Ao professor doutor Paulo Justiniano Ribeiro Junior, pelas experiências ao seu lado na disciplina transversal de métodos estatísticos em pesquisa científica na Universidade Federal do Paraná - MEPC, pelas oportunidades, conversas, risadas e pela maneira gentil que sempre me tratou.

Ao professor doutor Marco Antonio Zanata Alves pelo acolhimento, paciência, por confiar no meu potencial para participar de projetos de linhas de pesquisa que não estou habituado, e pelos valiosos conselhos sobre como tratar do tema desta dissertação para com pessoas com formações diferentes das minhas.

Aos pesquisadores Ligia Oliveira Carlos, Marilia Ramos, Nathalia Wagner, Ingrid Felicidade e Antonio Carlos Campos, do grupo de pesquisa em obesidade, cirurgia bariátrica e microbioma do departamento de clínica cirurgica da Universidade Federal do Paraná, pela disponibilização do conjunto de dados usado para motivar as ideias deste trabalho; em especial à Doutora Ligia Carlos pelo suporte prestado ao longo da elaboração e revisão do trabalho.

A todos os colegas de mestrado, em especial àqueles do grupo de pesquisa HiPES, pelo acolhimento e paciência com um novo colega de formação diferente.

Aos professores dos departamentos de Estatística, Informática e Matemática que fizeram parte da minha trajetória na Universidade Federal do Paraná. Em especial àqueles do Laboratório de Estatística e Geoinformação.

À Universidade Federal do Paraná - UFPR e ao Programa de Pós Graduação em Informática - PPGINF, incluindo professores e equipe administrativa, pelo brilhante trabalho e resiliência apresentados mesmo em tempos de pandemia.

À Coordenação de Aperfeiçoamento Pessoal de Ensino Superior - CAPES, pelo suporte financeiro nestes anos de pesquisa.

Aos membros da banca de qualificação e defesa da dissertação pelas valiosas contribuições.

À todos que estiveram ao meu lado no decorrer destes anos e contribuíram direta ou indiretamente neste trabalho.

À todos vocês, meu sincero agradecimento.

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 têm 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). No contexto de modelos de regressão, é comum o interesse em avaliar os valores dos parâmetros por meio de testes de hipóteses e existem técnicas baseadas em tais testes, como as análises de variâncias univariadas, multivariadas e ainda os testes de comparações múltiplas. No entanto, considerando os McGLMs, não há discussão a respeito do uso destes testes para a classe. Nossa proposta é utilizar o teste Wald para a realização de testes de hipóteses gerais sobre parâmetros de regressão e dispersão de McGLMs. Por meio da avaliação dos parâmetros de regressão é possível identificar qual(is) variável(is) explicativa(s) apresentam efeito sobre a(s) resposta(s). Por meio do estudo dos parâmetros de dispersão pode-se avaliar o efeito da correlação entre unidades do estudo, como por exemplo em estudos longitudinais, temporais e de medidas repetidas. Apresentamos implementações em R de funções para efetuar tais testes, bem como funções para efetuar ANOVAs, MANOVAs e testes de comparações múltiplas. As propriedades e comportamento dos testes propostos foram verificados com base em estudos de simulação e o potencial de aplicação das metodologias discutidas foi motivado com base na aplicação a um conjunto de dados real. Os resultados mostraram que quanto mais distante a hipótese testada é dos valores verdadeiros dos parâmetros, maior é o percentual de rejeição da hipótese nula. Tal como esperado, os menores percentuais de rejeição foram observados quando a hipótese nula testada correspondia aos reais valores dos parâmetros. Também verificou-se que conforme aumenta-se o tamanho amostral, o percentual de rejeição aumenta para hipóteses nulas pouco diferentes dos valores simulados dos parâmetros. Logo, os resultados apontam que o teste Wald pode ser usado para avaliar hipóteses sobre parâmetros de regressão e dispersão de McGLMs, o que permite uma melhor interpretação do efeito das variáveis e estruturas de delineamento em contextos práticos.

Palavras-chave: McGLM. Testes de hipóteses. Teste Wald. ANOVA. MANOVA. Comparações múltiplas. Regressão.

ABSTRACT

Data science is an interdisciplinary field of study that comprises 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. Such 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 response(s). There are univariate and multivariate models: in univariate models there is only one response variable; in multivariate models there is more than one response. Among the classes of multivariate models are the multivariate covariance generalized linear models (McGLMs). In the context of regression models, there is a common interest in evaluating parameter values through hypothesis tests and there are techniques based on such tests, such as univariate and multivariate analyzes of variance and even multiple comparison tests. However, considering the McGLMs, there is no discussion regarding the use of these tests for the class. Our proposal is to use the Wald test to carry out tests of general hypotheses on regression and dispersion parameters of McGLMs. By evaluating the regression parameters, it is possible to identify which explanatory variable(s) have an effect on the response(s). Through the study of dispersion parameters, the effect of the correlation between study units can be evaluated, for example in longitudinal, temporal and repeated measures studies. We present R implementations of functions to perform such tests, as well as functions to perform ANOVAs, MANOVAs and multiple comparison tests. The properties and behavior of the proposed tests were verified based on simulation studies and the potential of application of the discussed methodologies was motivated based on the application to a real dataset. The results showed that the further the tested hypothesis is from the true values of the parameters, the greater the percentage of rejection of the null hypothesis. As expected, the lowest rejection percentages were observed when the null hypothesis tested corresponded to the real values of the parameters. It was also verified that as the sample size increases, the rejection percentage increases for null hypotheses that are little different from the simulated values of the parameters. Therefore, the results indicate that the Wald test can be used to evaluate hypotheses about regression and dispersion parameters of McGLMs, which allows a better interpretation of the effect of variables and design structures in practical contexts.

Keywords: McGLM. Hypothesis tests. Wald test. ANOVA. MANOVA. Multiple comparisons. Regression.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS

LM	Modelo linear
GLM	Modelo linear generalizado
cGLM	Modelo de covariância linear generalizada
McGLM	Modelo multivariado de covariância linear generalizada
ANOVA	Análise de variância
MANOVA	Análise de variância multivariada
hGLM	Modelo linear generalizado hierárquico
GEE	Equações de Estimação Generalizadas
GAMLSS	Modelos aditivos generalizados para locação, escala e forma
GLMM	Modelos lineares generalizados mistos
MGLMM	Modelos lineares generalizados multivariados mistos
MLM	Modelos lineares multivariados
MGLM	Modelos lineares generalizados multivariados
RYGB	Bypass Gástrico em Y de Roux
BES	Escala de compulsão alimentar
YFAS	Escala de vício alimentar

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

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. Já Cao (2016) menciona que ciência de dados engloba técnicas de áreas 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, ou seja, os resultados não podem ser previstos com exatidão. 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 por meio 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, a função chamada de função densidade de probabilidade é usada para determinar as probabilidades associadas a intervalos da variável aleatória.

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, sendo que estes parâmetros são quantidades desconhecidas estimadas por meio dos dados. Na análise de regressão busca-se modelar os parâmetros das distribuições de probabilidade como uma função de outras variáveis. Isto é feito por meio 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 variáveis 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 destas estimativas dos parâmetros 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. A literatura fornece inúmeras classes de modelos de regressão, mencionaremos neste trabalho três delas: os modelos lineares (LM), os lineares generalizados (GLM) e os multivariados de covariância linear generalizada (McGLM). No cenário univariado, durante muitos anos o LM normal (Galton, 1886) teve papel de destaque no contexto dos modelos de regressão devido principalmente às suas facilidades computacionais. Um dos pressupostos do LM 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 do tempo, o avanço computacional permitiu a proposição de modelos mais complexos, que necessitavam de processos iterativos para estimação dos parâmetros (Paula, 2004). A classe de maior renome foram os GLMs 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 univariados de regressão 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.

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 McGLMs propostos por Bonat e Jørgensen (2016). Essa classe pode ser vista como 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 que possibilita 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 ausência de determinada variável explicativa do modelo geraria uma perda no ajuste. Deste modo, 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 por meio dos chamados testes de hipóteses. Testes de hipóteses são ferramentas estatísticas mais gerais, aplicadas a contextos além de regressão, 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. Os modelos de regressão tradicionais buscam encontrar as estimativas dos valores dos parâmetros que associam variáveis explicativas às respostas que maximizam a função de verossimilhança, ou seja, buscam encontrar um conjunto de valores 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 por meio da diferença do logaritmo da função de verossimilhança. Caso essa diferença seja estatisticamente 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 se estas variáveis forem retiradas, não existirá perda de qualidade no modelo.

O teste do multiplicador de lagrange ou teste escore (Aitchison e Silvey, 1958), (Silvey, 1959), (Rao, 1948a), 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 ausência de determinada variável do modelo prejudica o ajuste. No caso do teste de razão de verossimilhanças, dois modelos precisam ser ajustados. Já os testes Wald e escore necessitam de apenas um modelo. Além disso, os testes são assintoticamente 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 por meio 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).

Complementar às ANOVAs e MANOVAs estão os 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 de um fator. Com isso, o teste de comparações múltiplas é mais um procedimento baseado em testes de hipóteses, 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. Para detectar tais diferenças empregam-se os testes de comparações múltiplas. Dentre os testes discutidos na literatura encontram-se o teste de Dunnett, Tukey, t de student (LSD), Scott-Knott, dentre outros. Hsu (1996) discute diversos procedimentos para fins de comparações múltiplas. Já Bretz et al. (2008) trata de procedimentos de comparações múltiplas em modelos lineares.

1.2 DESAFIO

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 McGLMs, não há discussão a respeito da construção destes testes para a classe.

1.3 HIPÓTESE

Apesar da falta de estudos que busquem propor testes de hipóteses para os McGLMs, não é difícil vislumbrar que existem argumentos a favor da hipótese de que o teste Wald clássico utilizado em modelos tradicionais funcionaria para os McGLMs. A construção do teste Wald em sua forma usual 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 (razão de verossimilhanças, Wald e escore), o teste Wald se torna o mais atrativo. Outra vantagem do teste Wald em relação a seus concorrentes é que existe a possibilidade não só de formular hipóteses sobre conjuntos de parâmetros como também é possível confrontar as estimativas com qualquer valor desejado. Quando se trata dos McGLMs, esta ideia se torna especialmente atrativa pois fornece ferramentas para avaliar qualquer parâmetro de um McGLM.

Quando trabalhamos na classe dos McGLMs estimamos parâmetros de regressão, dispersão, potência e correlação. Os parâmetros de regressão são aqueles que associam a(s) variável(is) explicativa(s) à(s) variável(is) resposta(s), por meio do estudo destes parâmetros é possível avaliar o efeito da(s) variável(is) explicativa(s) sobre a(s) resposta(s). Por meio do estudo dos parâmetros de dispersão pode-se avaliar o efeito da correlação entre unidades do estudo, muito útil em situações em que as observações do conjunto de dados são correlacionadas entre si, como por exemplo em estudos longitudinais, temporais e de medidas repetidas. Os parâmetros de potência nos fornecem um indicativo de qual distribuição de probabilidade melhor se adequa ao problema. E os parâmetros de correlação estimam a força da associação entre respostas em um problema multivariado. O desenvolvimento de testes de hipóteses para fins de avaliação destas quantidades é de grande valia em problemas práticos e leva a formas procedurais para avaliação das quantidades resultantes do modelo.

1.4 OBJETIVO

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. Temos os seguintes objetivos específicos: propor a utilização do teste Wald para realização de testes de hipóteses gerais sobre parâmetros de regressão e dispersão de McGLMs, implementar em R funções para efetuar tais testes, bem como funções para efetuar análises de variância, análises de variância multivariadas e testes de comparações múltiplas para os McGLMs. Outro objetivo é 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.

1.5 CONTRIBUIÇÃO

Nossa proposta visa uma maneira procedural e segura de responder questões comuns no contexto de modelagem que frequentemente surgem em projetos de ciência de dados, como: quais variáveis estão associadas ao desfecho do fenômeno de interesse, se existe efeito da estrutura de correlação entre indivíduos no estudo, se 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 por meio da classe.

1.6 ORGANIZAÇÃO DO DOCUMENTO

Esta dissertação está organizada em 8 capítulos. Na atual seção foi exposto o tema e a ideia do trabalho de forma a enfatizar as características dos modelos de regressão, utilidade dos testes de hipóteses neste contexto, os testes mais famosos utilizados, procedimentos baseados em testes de hipóteses e nosso objetivo de propor o teste Wald para avaliação dos parâmetros de McGLMs. O Capítulo 2 é dedicado ao referencial teórico do trabalho, trata-se de uma revisão bibliográfica da estrutura dos McGLMs, testes de hipótese, análises de variância e testes de comparações múltiplas. No Capítulo 3 referenciamos trabalhos correlatos. No Capítulo 4 é apresentada nossa proposta com os detalhes do teste Wald para avaliar suposições sobre parâmetros de um McGLM. O Capítulo 5 é dedicado aos resultados da avaliação de perfomance do teste proposto com base em um estudo de simulação. As implementações computacionais do método proposto são apresentadas no Capítulo 6. No Capítulo 7 buscamos motivar o uso da proposta por meio da aplicação do método a um problema prático real de análise de dados. Por fim, encerramos o trabalho com nossas considerações finais no Capítulo 8.

2 REFERENCIAL TEÓRICO

Nosso referencial teórico aborda predominantemente três temas. O primeiro deles é uma revisão da estrutura geral e estimação dos parâmetros de um McGLM, baseado nas ideias de Bonat e Jørgensen (2016). A segunda parte do referencial diz respeito ao procedimento dos chamados testes de hipóteses com o foco de tratar do objetivo, notação, componentes e aplicação deste tipo de procedimento no contexto de modelos de regressão. Por fim, a última parte do referencial diz respeito a procedimentos específicos baseados em testes de hipóteses para avaliar os parâmetros de um modelo de regressão: as análises de variância e os testes de comparações múltiplas.

2.1 MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA

Os GLMs, propostos por Nelder e Wedderburn (1972), são uma forma de modelagem que lida exclusivamente com uma resposta em que esta resposta pode ser contínua, binária ou até mesmo uma contagem. 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: i) um leque restrito de distribuições disponíveis para modelagem, ii) a incapacidade de lidar com observações dependentes e iii) 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 McGLMs. Tais modelos, levam em conta a não normalidade por meio de uma função de variância. Além disso, a estrutura de média é modelada por meio de uma função de ligação e um preditor linear. Os parâmetros dos modelos são obtidos por meio de funções de estimação baseadas em suposições de segundo momento.

Vamos discutir os McGLMs como uma extensão dos GLMs tal como apresentado em de Bonat e Jørgensen (2016). Vale ressaltar que é usada uma especificação menos usual de um GLM, porém trata-se de uma notação mais conveniente para chegar a uma especificação mais simples de um McGLM.

2.1.1 Modelo linear generalizado

Para definição da extensão de um 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 $\vartheta(\boldsymbol{\mu}; p) = \mu^p$, 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 $\vartheta(\boldsymbol{\mu}; p) = \mu + \tau\mu^p$, 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 $\boldsymbol{I}$ no centro Equação 2.1. Podemos imaginar que, substituindo esta matriz identidade por uma matriz simétrica adequada 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, também apresentados em Bonat e Jørgensen (2016).

2.1.2 Modelo de covariância linear generalizada

Os cGLMs 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

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

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. Note que o número D de matrizes usadas para especificar o preditor linear matricial, é indefinido, ou seja, podem ser usadas quantas matrizes forem necessárias para especificação no modelo da relação entre os indivíduos no conjunto de dados. Cada uma das matrizes é associada a um parâmetro de dispersão e podemos utilizar estes parâmetros para avaliar a existência de efeito da correlação entre indivíduos do conjunto de dados. Tal estrutura pode ser vista como um análogo ao preditor linear para a média e foi nomeado como preditor linear matricial. 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. Mais 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 McGLMs podem ser entendidos como uma extensão multivariada dos cGLMs e que portanto contornam as principais restrições presentes nos GLMs. Para definição de um McGLM, considere $\boldsymbol{Y}_{N\times R} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_R\}$ 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 de dimensão $N \times N$, responsável por modelar a covariância dentro de cada resposta, sendo expressa por

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

Além disso, é necessário estimar 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

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

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 $N \times N$.

Com isso, chega-se a uma classe de modelos com um leque maior de distribuições disponíveis, graças às funções de variância. Além disso, se torna possível a modelagem de dados com estrutura de covariância, por meio da especificação do preditor matricial. E ainda é possível a modelagem de múltiplas respostas. Vale ressaltar que os McGLMs são flexíveis ao ponto de que podemos considerar R preditores lineares diferentes, com R funções de ligação diferentes e R funções de variância diferentes. Esta flexibilidade torna os McGLMs uma classe muito atrativa para aplicação, contudo, dependendo da estrutura e complexidade do problema, existe a possibilidade de ajustar modelos super parametrizados, ou seja, chegar a um cenário com mais parâmetros do que observações.

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 $\mathcal{M} = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_R^\top)^\top$ 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

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \boldsymbol{D}^\top \boldsymbol{C}^{-1}(\mathcal{Y} - \mathcal{M}), \tag{2.3}$$

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

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

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

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

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

$$\psi_{\lambda_i}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathrm{tr}(W_{\lambda i}(\boldsymbol{r}^\top \boldsymbol{r} - \boldsymbol{C})), \;\; i = 1, .., Q, \tag{2.4}$$

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

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

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

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

em que $k_l^{(4)}$ denota a quarta cumulante de $\mathcal{Y}_l$. 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

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

Seja $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\intercal, \hat{\boldsymbol{\lambda}}^\intercal)^\top$ o estimador baseado na Equação 2.3 e Equação 2.4, a distribuição assintótica de $\hat{\boldsymbol{\theta}}$ é

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

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

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

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 população 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 por meio 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 valor-p.

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 $\hat{\pi}$ a estimativa de um parâmetro π da população. Neste contexto, uma hipótese estatística é uma afirmação a respeito do valor do parâmetro π que é estudado por meio da estimativa $\hat{\pi}$ 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, por meio do estudo da quantidade $\hat{\pi}$ verificamos a plausibilidade de se afirmar que π é igual a um valor π_0. Portanto, três tipos de hipóteses podem ser especificadas:

1. $H_0 : \pi = \pi_0 \; vs \; H_1 : \pi \neq \pi_0$.
2. $H_0 : \pi = \pi_0 \; vs \; H_1 : \pi > \pi_0$.
3. $H_0 : \pi = \pi_0 \; vs \; H_1 : \pi < \pi_0$.

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 usualmente 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 por meio 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, valor-p ou ainda α^*. 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 valor-p 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 valor-p 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 valor-p.
8. Concluir pela rejeição ou não rejeição da hipótese nula.

2.2.2 Testes de hipóteses em modelos de regressão

A ideia de modelos de regressão consiste em modelar uma variável em função de um conjunto de variáveis explicativas. Estes modelos contêm parâmetros que são quantidades desconhecidas que estabelecem a relação entre as variáveis sob o modelo. Basicamente, o parâmetro de interesse da distribuição de probabilidades utilizada 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. Estes testes são assintoticamente equivalentes; em amostras finitas podem apresentar resultados diferentes de tal modo que a estatística do teste Wald é maior que a estatística do teste da razão de verossimilhanças que, por sua vez, é maior que a estatística do teste escore (Evans e Savin, 1982). Engle (1984) descreve a formulação geral dos três testes. Dedicaremos parte deste referencial ao teste Wald.

2.2.2.1 Teste Wald

O teste Wald (Wald, 1943) 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 dos parâmetros. 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 são 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 definição de um teste Wald, considere um único modelo de regressão ajustado em que os parâmetros foram estimados por meio da maximização da função de verossimilhança. Neste contexto, considere $\boldsymbol{\beta}$ o vetor de parâmetros de regressão $k \times 1$ deste modelo, em que as estimativas são dadas por $\hat{\boldsymbol{\beta}}$.

Considere que há interesse em testar s restrições ao modelo original. As hipóteses são especificadas por meio de uma matriz $\boldsymbol{L}$ de dimensão $s \times k$ e um vetor $\boldsymbol{c}$ de valores postulados, de dimensão s. 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:

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

em que $W \sim \chi^2_s$. 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.

2.2.3 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 por meio 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 em uma tabela: o chamado quadro de análise de variância. Em geral, este quadro contém em cada linha: a variável, o valor de uma estatística de teste referente à hipótese de nulidade de todos os parâmetros associados a esta variável, os graus de liberdade desta hipótese, e um valor-p associado à hipótese testada naquela linha do quadro.

Trata-se de um interessante procedimento para avaliar a relevância de uma variável no 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 dos tipos 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 quadros de análise de variância, faz-se uso de uma sequência de testes da razão de verossimilhanças para avaliar o efeito de cada variável explicativa do modelo.

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. A MANOVA clássica (Smith et al., 1962) é um assunto com vasta discussão na literatura e possui diversas propostas com o objetivo de verificar o efeito de variáveis explicativas

sobre múltiplas respostas, 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).

Na prática, é possível gerar quadros de análise de variância por meio do teste Wald. 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.

2.2.4 Testes de comparações múltiplas

Quando a ANOVA aponta para efeito significativo de uma variável categórica, costuma ser de interesse do pesquisador avaliar quais dos níveis diferem entre si. Para isso são empregados os testes de comparações múltiplas. Na literatura existem diversos procedimentos para efetuar tais testes, muitos deles descritos em Hsu (1996).

No contexto de modelos de regressão costuma ser de interesse avaliar comparações aos pares a fim de detectar para quais níveis da variável categórica os valores da resposta se alteram. Tal tipo de situação pode ser avaliada utilizando o teste Wald. Através da correta especificação da matriz $\boldsymbol{L}$, é possível avaliar hipóteses sobre qualquer possível contraste entre os níveis de uma determinada variável categórica. Portanto, é possível usar a estatística de Wald para efetuar também testes de comparações múltiplas.

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

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

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

Nesta parametrização o primeiro nível da variável categórica é mantido como categoria de referência e, para os demais níveis, mede-se a mudança para a categoria de referência; este é o chamado contraste de tratamento. Neste contexto β_0 representa a média ajustada do nível A, enquanto que β_1 representa a diferença de A para B, β_2 representa a diferença de A para C e β_3 representa a diferença de A para D. Com esta parametrização é possível obter o valor predito para qualquer uma das categorias de tal modo que se o indivíduo pertencer à categoria A, β_0 representa o predito; se o indivíduo pertencer à categoria B, $\beta_0 + \beta_1$ representa o predito; para a categoria C, $\beta_0 + \beta_2$ representa o predito e, por fim, para a categoria D, $\beta_0 + \beta_3$ representa o predito.

Matricialmente, estes resultados podem ser descritos como

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

Note que o produto $\boldsymbol{K_0\beta}$ gera o vetor de preditos para cada nível de X. Por meio da subtração das linhas da matriz de combinações lineares $\boldsymbol{K_0}$ podemos gerar uma matriz de contrastes $\boldsymbol{K_1}$

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

Para proceder um teste de comparações múltiplas basta selecionar os contrastes desejados nas linhas da matriz $\boldsymbol{K_1}$ e utilizar estas linhas como matriz de especificação de hipóteses do teste Wald. Por fim, como usual em testes de comparações múltiplas, é recomendada a correção dos valores-p por meio da correção de Bonferroni. A correção de Bonferroni atualiza os valores-p levando em conta a quantidade de testes realizados, tornando o teste mais rigoroso. O novo valor-p é dado pelo valor original dividio pelo número de comparações realizadas.

3 TRABALHOS RELACIONADOS

Os McGLMs foram propostos com o objetivo de contornar as restrições dos GLMs no que diz respeito à quantidade de distribuições disponíveis, alternativas para modelar a correlação entre unidades e a capacidade de modelar múltiplas respostas de diferentes naturezas. Nesta seção de revisão de literatura serão brevemente citadas propostas que visam contornar restrições dos GLMs e como são efetuados testes de hipóteses para estas propostas. Os parâmetros dos GLMs são estimados por meio do método da máxima verossimilhança e por isso os testes de hipóteses clássicos são amplamente utilizados: teste da razão de verossimilhanças, teste Wald e teste escore.

No cenário univariado, existe um conjunto de propostas em que a ideia é utilizar efeitos aleatórios para acomodar a correlação entre observações quando há necessidade. A ideia dos modelos de efeitos aleatórios é que as medidas são correlacionadas pois compartilham de um mesmo efeito que não é observado. Dentre as propostas que envolvem efeitos aleatórios estão os modelos lineares generalizados mistos (GLMM), que estendem os modelos mistos permitindo que a resposta pertença à família exponencial. Algumas referências que discutem modelos de efeitos aleatórios são Laird e Ware (1982), Jiang e Nguyen (2007), Stroup (2012).

Em modelos que contam com efeitos aleatórios, a interpretação dos parâmetros de regressão, também chamados de componentes de efeitos fixos, dependem de manter o efeito aleatório fixado, pois o vetor de parâmetros de regressão tem interpretação condicional ao nível dos efeitos aleatórios. Por essa razão costumam ser chamados de coeficientes de regressão específicos do indivíduo. A estimação destes modelos não é simples; o ajuste envolve integrais complexas e é uma tarefa computacionalmente desafiadora, mas é possível usar máxima verossimilhança, o que torna o uso dos testes de hipóteses tradicionais uma alternativa para inferência após ajuste dos modelos. Detalhes sobre procedimentos inferenciais nos GLMMs podem ser consultados em Tuerlinckx et al. (2006).

Outra proposta é a dos modelos lineares generalizados hierarquicos (hGLMs) (Lee e Nelder, 1996). Nesta classe a estimação dos parâmetros é baseada na chamada h-verossimilhança e supera algumas das limitações dos GLMs mas, em geral, as implementações são limitadas a uma variável resposta, apesar de ser possível pensar na proposta para o caso multivariado. Os hGLMs não são estimados com base na máxima verossimilhança, contudo os autores sugerem um teste de hipóteses ao estilo do teste de razão de verossimilhanças para os parâmetros de regressão. Além disso, o teste Wald pode ser utilizado tanto para parâmetros de regressão quanto para parâmetros de dispersão do modelo.

Em uma linha diferente das propostas que envolvem efeitos aleatórios, uma alternativa para acomodar a correlação entre observações é por meio das equações de estimação generalizadas (Liang e Zeger, 1986), popularmente chamadas de GEE. Trata-se de uma abordagem em que a ideia consiste em incluir no processo de estimação uma matriz de correlação de trabalho. Na prática existem diversas implementações, contudo há um conjunto limitado de estruturas de covariância possíveis. O foco do método não é a modelagem da estrutura de correlação entre os indivíduos, mas sim a correção dos erros padrões das estimativas e por isso os testes de hipóteses tradicionais são utilizados sem qualquer complicação.

Dentre as propostas mais recentes que lidam com uma única resposta estão os modelos aditivos generalizados para locação, escala e forma (GAMLSS), propostos por Stasinopoulos e Rigby (2008). Esta é uma classe de modelos de regressão univariados com um número considerável de distribuições de probabilidade disponíveis para modelagem. Além disso é

possível modelar todos os parâmetros distribucionais em função de variáveis explicativas e ainda incluir aos preditores efeitos aleatórios e termos suavizadores. A estimação dos GAMLSS é feita com base na maximização de uma função de verossimilhança que é penalizada nos casos em que há a inclusão de termos não paramétricos e os testes de hipóteses tradicionais costumam ser empregados.

Já no cenário multivariado, uma das propostas mais populares é a dos modelos lineares generalizados multivariados mistos (MGLMM) (Berridge e Crouchley, 2019). O MGLMM é uma extensão do GLMM para lidar com múltiplas respostas e segue a mesma ideia no que diz respeito à inclusão de efeitos aleatórios para acomodar a dependência entre observações. Tal como no caso univariado, a estimação para esta classe em alguns casos se torna complexa pois a função de verossimilhança não apresenta forma fechada e a distribuição marginal de cada resposta não é conhecida, pois a especificação do modelo é feita de forma condicional aos efeitos aleatórios, o que torna a obtenção das marginais uma tarefa não trivial. Além disso, o algoritmo de estimação por si só não é simples e a interpretação dos coeficientes de regressão não é igual à forma como interpreta-se um GLM devido a presença de efeito aleatório. Contudo, a estimação pode ser feita maximizando uma função de verossimilhança o que torna possível utilizar os testes de hipóteses usuais.

Existe ainda na literatura uma grande variedade de modelos de regressão multivariados para fins específicos. Zhang et al. (2017) propõe um modelo geral para análise de contagens multivariadas em que os testes usuais se aplicam. Mardalena et al. (2020) mostra como estimar os parâmetros e ainda como efetuar testes de hipóteses ao estilo da razão de verossimilhanças para um modelo de regressão multivariado com distribuição Poisson inversa gaussiana. Já Sari et al. (2021) mostra como estimar os parâmetros e testar hipóteses por meio do testes de razão de verossimilhanças e teste Wald em um modelo de regressão Poisson zero inflacionado multivariado. Berliana et al. (2019) apresenta um modelo de regressão Poisson generalizado multivariado com exposição e correlação em função de covariáveis em que teste de hipóteses podem ser feitos com base na razão de verossimilhanças. Rahayu et al. (2020) propõe um modelo de regressão multivariado gamma em que as hipóteses sobre os parâmetros podem ser testadas por meio do teste Wald e um procedimento análogo ao teste de razão de verossimilhanças.

Existem ainda propostas para realizar testes de hipóteses multivariados e análises de variância multivariadas. Rao (1948b) em um artigo clássico discute e busca desenvolver uma abordagem unificada para o problema de testes de significância em análise multivariada. Em um trabalho mais atual, Smaga (2017) discute métodos bootstrap para testes de hipóteses multivariados e, para o problema de testar o vetor médio de uma distribuição multivariada, a validade assintótica dos métodos bootstrap é comprovada. Olson (1976) trata a respeito da escolha de uma estatística de teste na análise de variância multivariada. Referências como Hand e Taylor (1987) mostram como proceder análises de variância em um contexto de medidas repetidas. Adeleke et al. (2014), por meio de estudos de Monte-Carlo, avalia os comportamentos de algumas técnicas existentes para realização de análises de variância multivariadas quando a suposição de normalidade é violada.

Comparado às classes de modelos apresentadas, os McGLMs configuram uma classe mais geral pois acomoda múltiplas respostas, com uma grande variedade de distribuições disponíveis e ainda é possível não só especificar no modelo que há uma estrutura de correlação entre as observações como também modelar esta estrutura. Por isso, apesar de ainda não explorado, o estudo a respeito de testes de hipóteses para os McGLMs é um complemento necessário para tornar a classe ainda mais completa.

4 TESTE WALD EM MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA

Este capítulo é dedicado à apresentação de nossa proposta: o uso do teste Wald para avaliação dos parâmetros de McGLMs. Vale lembrar que nos McGLMs existem parâmetros de regressão, dispersão, potência e correlação e que, neste trabalho, estamos interessados na avaliação dos parâmetros de regressão e dispersão. Cada conjunto de parâmetros possui uma interpretação prática bastante relevante de tal modo que por meio dos parâmetros de regresão é possível identificar as explicativas relevantes, por meio dos parâmetros de dispersão é possível avaliar o impacto da correlação entre unidades do conjunto de dados, por meio dos parâmetros de potência é possível identificar qual distribuição de probabilidade melhor se adequa ao problema de acordo com a função de variância e por meio dos parâmetros de correlação é possível avaliar a associação entre respostas.

É importante notar que como temos que as estimativas dos parâmetros seguem distribuição normal, qualquer forma quadrática obtida por meio do estimador ou combinações lineares (tal como a estatística de Wald) é também uma forma quadrática que tem distribuição qui-quadrado. Portanto, nossa proposta não é uma simples adaptação, mas é um procedimento assintoticamente válido com respaldo na teoria da distribuição de formas quadráticas. Mais sobre a distribuição de formas quadráticas pode ser visto nos trabalhos de Graybill e Marsaglia (1957), Luther (1965) e Baldessari (1967).

4.1 HIPÓTESES E ESTATÍSTICA DE TESTE

Um McGLM possui vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\lambda}^\top)^\top$, em que $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_R^\top)^\top$ 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.

Considere $\boldsymbol{\theta}^*$ o vetor $h \times 1$ de parâmetros desconsiderando os parâmetros de correlação, ou seja, $\boldsymbol{\theta}^*$ refere-se apenas a parâmetros de regressão, dispersão ou potência. As estimativas dos parâmetros de $\boldsymbol{\theta}^*$ são dadas por $\hat{\boldsymbol{\theta}}^*$. De maneira similar, considere J^{*-1} a inversa da matriz de informação de Godambe desconsiderando os parâmetros de correlação, de dimensão $h \times h$.

Seja $\boldsymbol{L}$ uma matriz de especificação de hipóteses a serem testadas, de dimensão $s \times h$ e $\boldsymbol{c}$ um vetor de dimensão $s \times 1$ com os valores sob hipótese nula, em que s denota o número de restrições. As hipóteses a serem testadas podem ser escritas como:

$$H_0 : \boldsymbol{L\theta}^* = \boldsymbol{c} \; vs \; H_1 : \boldsymbol{L\theta}^* \neq \boldsymbol{c}. \tag{4.1}$$

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

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

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

Em geral, cada coluna da matriz $\boldsymbol{L}$ corresponde a um dos h parâmetros de $\boldsymbol{\theta}^*$ e cada linha a uma restrição. Sua construção consiste basicamente em preencher a matriz com 0, 1 e eventualmente -1 de tal modo que o produto $\boldsymbol{L\theta}^*$ represente corretamente as hipóteses de

interesse. A correta especificação de $\boldsymbol{L}$ permite testar qualquer parâmetro individualmente ou até mesmo formular hipóteses para diversos parâmetros.

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, isto é, se o efeito das variáveis sobre a resposta é o mesmo.

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 respostas, denotadas por Y_1 e Y_2. Para tal tarefa coletou-se uma amostra com N observações e para cada observação registrou-se os valores de X_1, Y_1 e Y_2. Com base nos dados coletados ajustou-se um McGLM bivariado, com preditor dado por:

$$g_r(\mu_r) = \beta_{r0} + \beta_{r1} X_1, r = 1, 2, \tag{4.2}$$

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: existe efeito da variável X_1 apenas sobre a primeira resposta? Ou apenas sobre a segunda resposta? É possível que a variável X_1 possua efeito sobre as duas respostas ao mesmo tempo? É possível que o efeito da variável seja o mesmo para ambas as respostas? Todas essas perguntas podem ser respondidas por meio de testes de hipóteses sobre os parâmetros do modelo e especificadas por meio da Equação 4.1. Nas subseções a seguir são apresentados os elementos para responder a cada uma destas perguntas.

4.1.1 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 sobre a primeira resposta. A hipótese pode ser escrita da seguinte forma:

$$H_0 : \beta_{11} = 0 \; vs \; H_1 : \beta_{11} \neq 0. \tag{4.3}$$

Esta mesma hipótese pode ser reescrita na notação mais conveniente para aplicação da estatística do teste Wald, tal como na Equação 4.1 em que:

- $\boldsymbol{\theta}^{*T} = \begin{bmatrix} \beta_{10} & \beta_{11} & \beta_{20} & \beta_{21} & \tau_{11} & \tau_{21} \end{bmatrix}$.
- $\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$.
- $\boldsymbol{c} = \begin{bmatrix} 0 \end{bmatrix}$.

Note que o vetor $\boldsymbol{\theta}^*$ possui seis elementos, consequentemente a matriz $\boldsymbol{L}$ contém seis colunas (uma para cada elemento) e 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 $\boldsymbol{L\theta}^*$ representa a hipótese de interesse inicialmente postulada na Equação 4.3. Com isso, a distribuição assintótica do teste é χ^2_1.

4.1.2 Exemplo 2: hipótese para múltiplos parâmetros

Suponha agora que o interesse neste problema genérico não seja mais testar o efeito da variável explicativa apenas em uma resposta. Suponha que o interesse seja avaliar se existe evidência suficiente para afirmar que há efeito da variável explicativa X_1 em ambas as respostas simultaneamente. Neste caso teremos que testar 2 parâmetros: 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:

$$H_0 : \beta_{r1} = 0 \; vs \; H_1 : \beta_{r1} \neq 0, \tag{4.4}$$

ou, de forma equivalente:

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

As hipóteses na forma da Equação 4.1 possui os seguintes elementos:

- $\boldsymbol{\theta^{*T}} = \begin{bmatrix} \beta_{10} & \beta_{11} & \beta_{20} & \beta_{21} & \tau_{11} & \tau_{21} \end{bmatrix}$.
- $\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$.
- $\boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

O vetor $\boldsymbol{\theta^*}$ 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 $\boldsymbol{L\theta^*}$ representa a hipótese de interesse inicialmente postulada na Equação 4.4. Com isso, a distribuição assintótica do teste é χ^2_2.

4.1.3 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 formaríamos uma hipótese de igualdade entre os parâmetros ou, em outros termos, se a diferença dos efeitos é nula:

$$H_0 : \beta_{11} - \beta_{21} = 0 \; vs \; H_1 : \beta_{11} - \beta_{21} \neq 0, \tag{4.5}$$

na notação da Equação 4.1 os elementos das hipóteses são:

- $\boldsymbol{\theta^{*T}} = \begin{bmatrix} \beta_{10} & \beta_{11} & \beta_{20} & \beta_{21} & \tau_{11} & \tau_{21} \end{bmatrix}$.
- $\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 & 0 \end{bmatrix}$.
- $\boldsymbol{c} = \begin{bmatrix} 0 \end{bmatrix}$.

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 é necessário colocar 1 na coluna referente a um parâmetro, e -1 na coluna referente ao outro parâmetro, de tal modo que o produto $\boldsymbol{L\theta^*}$ represente a hipótese de interesse inicialmente postulada, neste caso o produto $\boldsymbol{L\theta^*}$ gera exatamente a mesma hipótese especificada na Equação 4.5 e a distribuição assintótica do teste é χ^2_1.

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

A Equação 4.2 descreve um modelo bivariado genérico. É importante notar que neste exemplo ambas as respostas estão sujeitas ao mesmo preditor. Na prática, quando se trata dos McGLMs, preditores diferentes podem ser especificados entre variáveis respostas. Deste modo, o que foi exposto na Subseção 4.1.2 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, independente dos preditores entre respostas.

Contudo, nos casos em que as respostas estão sujeitas a preditores idênticos e as hipóteses sobre os parâmetros não se alteram de resposta para resposta, uma especificação alternativa do procedimento é utilizando o produto Kronecker para testar uma mesma hipótese sobre múltiplas respostas tal como utilizado em Bonat et al. (2020).

Suponha que, neste exemplo, as hipóteses de interesse seguem sendo escritas tal como na Equação 4.4. Contudo, como se trata de um modelo bivariado com mesmo preditor para as duas respostas, a hipótese de interesse é igual entre respostas e envolve apenas parâmetros de regressão, torna-se conveniente escrever a matriz $\boldsymbol{L}$ como o produto Kronecker de duas matrizes: uma matriz $\boldsymbol{G}$ e uma $\boldsymbol{F}$, ou seja, $\boldsymbol{L} = \boldsymbol{G} \otimes \boldsymbol{F}$. Desta forma, 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 $(s'R \times h)$.

Em geral, 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 por meio 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 uma matrix $\boldsymbol{L}$ 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.
- $\boldsymbol{c} = \begin{bmatrix}0 \\ 0\end{bmatrix}$: vetor de valores da hipótese nula.

Deste modo, 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 hipótese sobre os parâmetros de dispersão e esta especificação é bastante conveniente para a geração de quadros de análise de variância.

4.2 ANOVA E MANOVA VIA TESTE WALD

Com base na proposta de utilização do teste Wald para McGLMs, buscamos propor neste trabalho três diferentes procedimentos para geração de quadros de ANOVA e MANOVA para parâmetros de regressão, seguimos a nomenclatura tipos I, II e III. Além disso, buscamos propor também um procedimento análogo à uma ANOVA e MANOVA para avaliação dos parâmetros de dispersão de um dado modelo. No caso das ANOVAs gera-se um quadro para cada variável resposta. Para as MANOVAs apenas um quadro é gerado, por isso, para que seja possível realizar as MANOVAs, as respostas do modelo devem estar sujeitas ao mesmo preditor.

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

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

em que o índice r denota a variável resposta, r = 1,2; β_{r0} representa o intercepto; β_{r1} um parâmetro de regressão associado à variável X_1, β_{r2} um parâmetro de regressão associado à variável X_2 e β_{r3} um parâmetro de regressão associado à interação entre X_1 e X_2. Considere que as unidades em estudo são independentes, portanto cada resposta possui apenas um parâmetro de dispersão τ_{r0} associado a uma matriz $Z_0 = I$. Além disso considere que os parâmetros de potência foram fixados.

4.2.1 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 mais interesse.

4.2.2 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.

4.2.3 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 não avalia parâmetros de interação juntamente com parâmetros de efeito fixo. 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 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.

4.3 TESTE DE COMPARAÇÕES MÚLTIPLAS VIA TESTE WALD

Outra contribuição deste trabalho é a proposta de um procedimento para comparações aos pares a fim de detectar diferenças entre níveis de variáveis explicativas categóricas em McGLMs. Trata-se de uma generalização para os McGLMs do procedimento genérico descrito na Subseção 2.2.4, em que utiliza-se o teste Wald para comparação de cada par de níveis de uma variável categórica por meio da especificação de um conjunto de matrizes $\boldsymbol{L}$ do teste Wald.

O procedimento é idêntico ao descrito na Subseção 2.2.4 e consiste em:

1. Obter a matriz de combinações lineares dos parâmetros do modelo que resultam nas médias ajustadas.

2. Gerar a matriz de contrastes (subtração duas a duas das linhas da matriz de combinações lineares).

3. Selecionar as linhas de interesse desta matriz e usá-las como matriz de especificação de hipóteses do teste Wald, no lugar da matriz $\boldsymbol{L}$.

4. Devido ao grande número de testes feitos, os valores-p são corrigidos por meio da correção de Bonferroni.

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

Na prática, se o interesse for um teste de comparações múltiplas multivariado, existe a necessidade de todas as respostas estarem sujeitas a um mesmo preditor, e basta expandir a matriz de contrastes utilizando o produto Kronecker, seguindo uma ideia muito similar ao exposto na Subseção 4.1.4. No caso de um teste de comparações múltiplas para cada resposta, basta selecionar o vetor de estimativas e a partição correspondente ao vetor da matriz $J_{\boldsymbol{\theta}}^{-1}$ para a resposta específica e proceder com o teste. Desta forma é possível chegar a um simples e útil procedimento de comparações múltiplas para quando há um McGLM com variáveis explicativas categóricas e há interesse em determinar quais níveis diferem entre si.

5 ESTUDO DE SIMULAÇÃO

Com o objetivo de avaliar o poder do teste Wald em testes de hipóteses sobre parâmetros de McGLMs, foram executados estudos de simulação. Nestas simulações avaliamos o comportamento da proposta para três distribuições de probabilidade: Normal, Poisson e Bernoulli. Simulamos cenários univariados e também trivariados com diferentes tamanhos amostrais para verificar as propriedades dos testes sobre parâmetros de regressão e dispersão.

Para simular conjuntos de dados univariados foram usadas bibliotecas padrões do R. Para simular conjuntos de dados com múltiplas respostas seguindo distribuição Normal, foi usada a biblioteca R *mvtnorm* (Genz et al., 2021), (Genz e Bretz, 2009). Para as outras distribuições foi utilizado o método NORTA (Cario e Nelson, 1997) implementado na biblioteca R *NORTARA* (Su, 2014).

5.1 PARÂMETROS DE REGRESSÃO

Para avaliação de hipóteses sobre parâmetros de regressão foram considerados tamanhos amostrais de 50, 100, 250, 500 e 1000. Foram gerados 500 conjuntos de dados para cada tamanho amostral simulando uma situação com uma variável explicativa categórica de 4 níveis. Para distribuição Normal os parâmetros de regressão usados foram: $\beta_0 = 5, \beta_1 = 0, \beta_2 = 0, \beta_3 = 0$. Para a distribuição Poisson os parâmetros de regressão usados foram: $\beta_0 = 2,3, \beta_1 = 0, \beta_2 = 0, \beta_3 = 0$. E para a distribuição Bernoulli os parâmetros de regressão usados foram: $\beta_0 = 0,5, \beta_1 = 0, \beta_2 = 0, \beta_3 = 0$. Os valores foram escolhidos de tal modo que o coeficiente de variação para distribuição Normal fosse de 20%, as contagens para Poisson fossem próximas de 10 e a probabilidade de sucesso da Bernoulli fosse aproximadamente 0,6. Foram avaliados cenários univariados e trivariados com estas características. Para os cenários trivariados, existem 4 parâmetros por resposta que seguem as configurações descritas. Para cada amostra gerada foi ajustado um McGLM nos quais as funções de ligação e variância para cada distribuição são apresentadas na Tabela 5.1.

Distribuição	Função de ligação	Função de variância
Normal	Identidade	Constante
Poisson	Logarítmica	Tweedie
Bernoulli	Logito	Binomial

Tabela 5.1: Funções de ligação e variância utilizadas nos modelos para cada distribuição simulada.

Em todos os casos o preditor matricial para a matriz de variância e covariância foi especificado de forma a explicitar que as observações são independentes dentro de cada resposta. A correlação entre respostas no caso trivariado é dada pela matriz Σ_b descrita na Equação 5.1.

$$\Sigma_b = \begin{bmatrix} 1 & 0,75 & 0,5 \\ 0,75 & 1 & 0,25 \\ 0,5 & 0,25 & 1 \end{bmatrix} \tag{5.1}$$

Com os modelos ajustados, o procedimento consistiu em variar as hipóteses testadas sobre os parâmetros simulados. Consideramos 20 diferentes hipóteses baseadas em um decréscimo em β_0 e distribuição deste decréscimo nos demais βs da hipótese nula. O decréscimo para respostas

seguindo distribuição Normal foi de 0,15; para distribuição Poisson o decréscimo foi de 0,05; e para distribuição Bernoulli o decréscimo foi de 0,25. Estes valores foram escolhidos levando em conta o afastamento desejado das hipóteses testadas na escala da resposta. É importante notar que estes valores são diferentes devido ao impacto da função de ligação usada em cada configuração de modelo, e também devido às propriedades dos parâmetros das distribuições.

Para cada ponto avaliamos o percentual de rejeição da hipótese nula. A ideia é verificar o que ocorre quando afastamos as hipóteses nulas dos reais valores dos parâmetros. Espera-se que no primeiro ponto haja um percentual de rejeição baixo, pois a hipótese nula corresponde aos reais valores dos parâmetros. Para os demais pontos espera-se que o percentual de rejeição aumente gradativamente, pois as hipóteses afastam-se cada vez mais dos valores originalmente simulados. As hipóteses testadas em cada cenário estão disponíveis no apêndice desta dissertação.

Para representar graficamente os resultados tomamos a distância euclidiana de cada vetor de hipóteses com relação ao vetor usado para simular os dados. Adicionalmente dividimos o vetor de distâncias pela maior distância para obter distâncias padronizadas entre 0 e 1, independente dos parâmetros de regressão. Os resultados são apresentados na Figura 5.1.

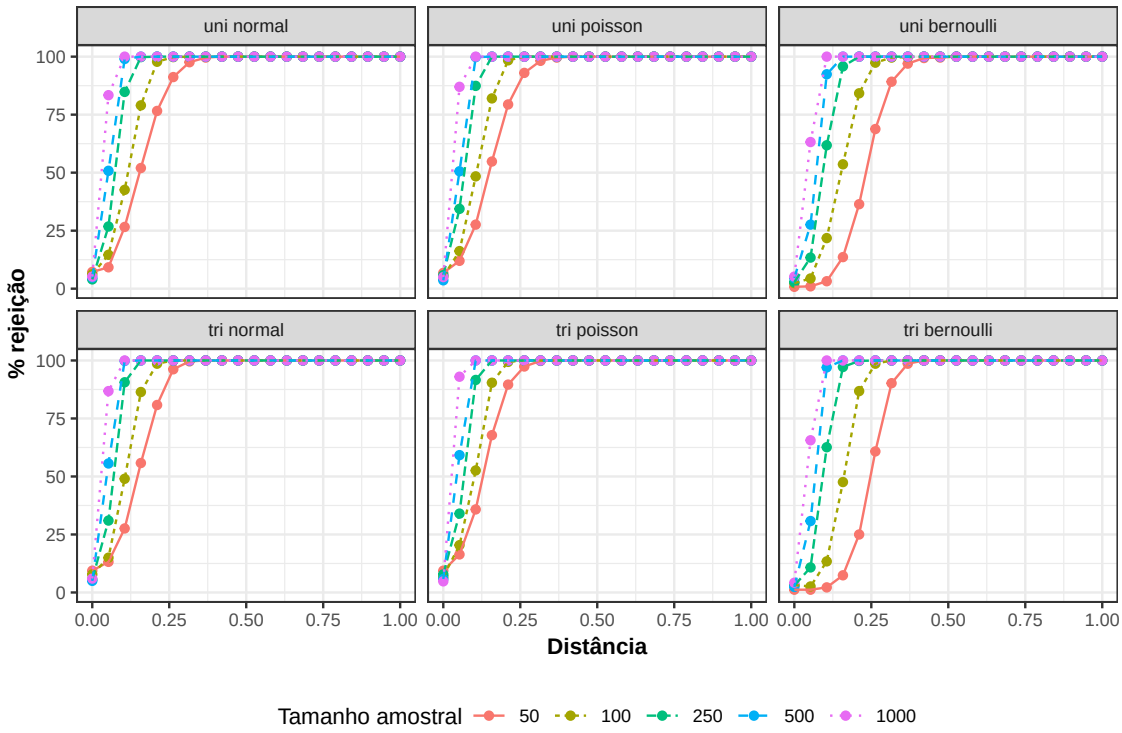


Figura 5.1: Resultados do estudo de simulação para os parâmetros de regressão.

De modo geral, quanto mais distante a hipótese é dos valores inicialmente simulados, maior é o percentual de rejeição. Como esperado, os menores percentuais foram observados na hipótese igual aos valores simulados. Nos cenários univariados o percentual de rejeição foi próximo de 5% quando a hipótese era igual aos valores simulados mesmo com tamanhos amostrais reduzidos. Para os cenários trivariados, no menor tamanho amostral avaliado, o percentual de rejeição não excedeu 10% e em tamanhos amostrais iguais a 500 o percentual de rejeição foi próximo de 5%. Também como esperado, foi possível verificar que conforme aumenta-se o tamanho amostral, o percentual de rejeição aumenta para hipóteses pouco diferentes dos valores simulados dos parâmetros.

5.2 PARÂMETROS DE DISPERSÃO

Para avaliação de hipóteses sobre parâmetros de dispersão foram considerados os mesmos tamanhos amostrais: 50, 100, 250, 500 e 1000. Contudo, os conjuntos de dados simulam uma situação em que cada unidade amostral fornece 5 medidas ao conjunto de dados. Foram gerados 500 conjuntos de dados para cada tamanho amostral e distribuição. Para distribuição Normal foram simulados vetores com média 5 e desvio padrão igual a 1. Para distribuição Poisson foram simuladas contagens com taxa igual a 10. Para distribuição Bernoulli foram simulados vetores de uma variável dicotômica com probabilidade de sucesso igual a 0,6.

Em todos os casos, os parâmetros de dispersão para gerar os conjuntos de dados foram fixados em $\tau_0 = 1$, $\tau_1 = 0$ e não foi incluído efeito de variáveis explicativas. Foram avaliados cenários univariados e trivariados com estas características. Para cada amostra gerada foi ajustado um McGLM com funções de ligação e variância tal como descrito na Tabela 5.1. Nos cenários trivariados a correlação entre respostas é dada pela Equação 5.1.

Neste caso, como o objetivo é avaliar a correlação dentro das respostas, é necessário especificar um preditor matricial. O objetivo é testar hipóteses sobre os parâmetros de dispersão associados a este preditor matricial.

Com os modelos ajustados, o procedimento consistiu em variar as hipóteses testadas sobre os parâmetros simulados. Consideramos 20 diferentes hipóteses baseadas em um decréscimo sucessivo de 0,02 em τ_0 e acréscimo de 0,02 em τ_1 para cada hipótese nula testada. Para cada ponto avaliamos o percentual de rejeição da hipótese nula. A ideia é afastar sucessivamente a hipótese dos valores simulados e avaliar se conforme afastamos a hipótese dos valores verdadeiros, o percentual de rejeição aumenta. As hipóteses testadas estão disponíveis no apêndice.

Do mesmo modo que foi feito para os parâmetros de regressão, foi tomada a distância euclidiana de cada vetor de hipóteses com relação ao vetor usado para simular os dados; e o vetor de distâncias foi padronizado para obter distâncias entre 0 e 1. Os resultados são apresentados na Figura 5.2.

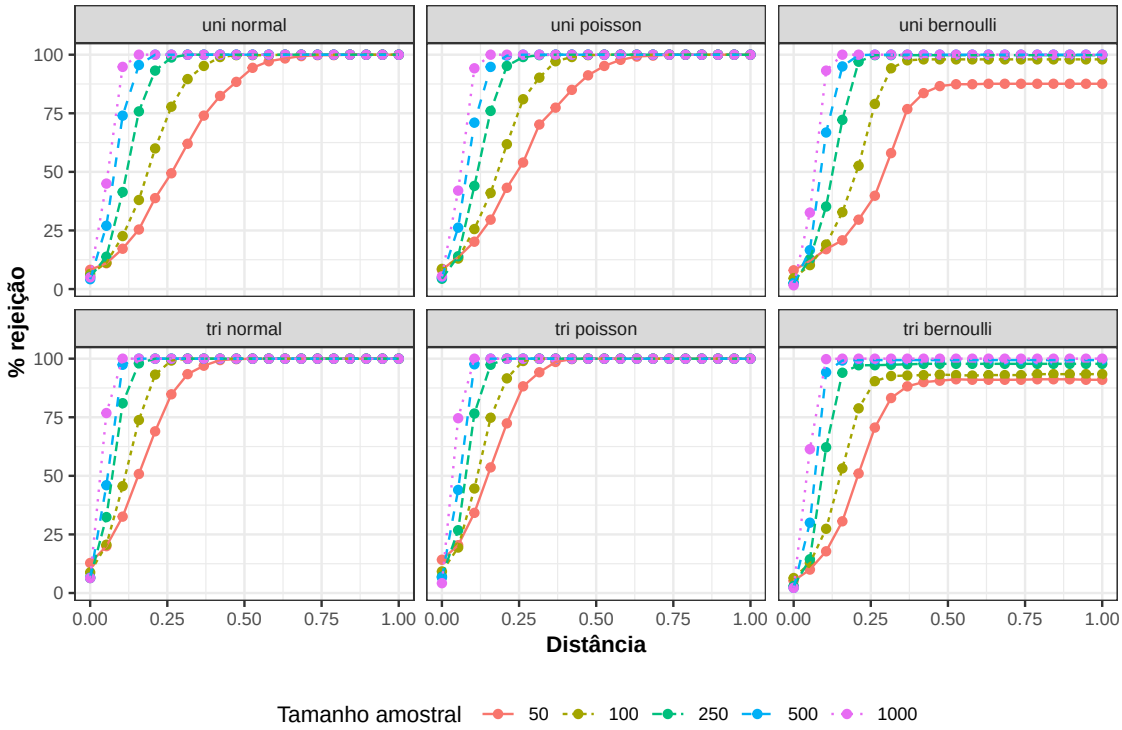


Figura 5.2: Resultados do estudo de simulação para os parâmetros de dispersão.

Assim como observado para os parâmetros de regressão, o comportamento dos gráficos mostra que, quanto mais distante a hipótese é dos valores inicialmente simulados, maior é o percentual de rejeição, e os menores percentuais são observados em hipóteses próximas aos valores simulados. Na primeira hipótese testada, para os cenários univariados, percentuais de rejeição próximos a 8% foram observados no menor tamanho amostral avaliado. A partir de tamanhos amostrais iguais a 250 o percentual de rejeição foi próximo de 5%. Já para os casos trivariados, no menor tamanho amostral, o percentual de rejeição excedia 10% no menor tamanho amostral; para tamanhos amostrais maiores, os percentuais ficaram em torno de 7%. Também verificou-se para os parâmetros de dispersão que conforme aumenta-se o tamanho amostral, o percentual de rejeição aumenta para hipóteses pouco diferentes dos valores simulados dos parâmetros.

6 IMPLEMENTAÇÃO COMPUTACIONAL

Um dos objetivos deste trabalho consiste em implementar e disponibilizar publicamente os testes apresentados no Capítulo 4, com o intuito de complementar as já possíveis análises permitidas pelo pacote *mcglm* (Bonat, 2018). Este capítulo é destinado à apresentação das implementações computacionais em R das funções que usam o teste Wald para avaliar parâmetros de McGLMs.

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 Wald para comparar modelos lineares e lineares generalizados aninhados. Já o pacote *survey* (Lumley, 2020); (Lumley, 2004);(Lumley, 2010) possui uma função que efetua testes Wald que, por padrão, testa se todos os coeficientes associados a um determinado termo de regressão são zero, mas é possível especificar hipóteses com outros valores.

O pacote *car* (Fox 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 aos quadros 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. Para comparações múltiplas, um dos principais pacotes disponíveis é o *multcomp* (Hothorn et al., 2008) que fornece uma interface para testes de comparações múltiplas para modelos paramétricos.

Contudo, quando se trata de McGLMs ajustados no pacote *mcglm* existe apenas um tipo de análise de variância univariada implementada na biblioteca e não existem opções para realização de testes de hipóteses lineares gerais, nem testes de comparações múltiplas.

Com base na proposta de uso do teste de Wald para McGLMs, desenvolvemos o pacote *htmcglm* com procedimentos para testar hipóteses lineares gerais, gerar quadros de ANOVA e MANOVA para parâmetros de regressão e dispersão e também testes de comparações múltiplas. Todos esses procedimentos foram implementados na linguagem R e complementam as funcionalidades existentes na biblioteca *mcglm*.

6.1 FUNÇÕES R

Todas as funções implementadas geram resultados mostrando graus de liberdade e valores-p baseados no teste Wald aplicado a um McGLM. A Tabela 6.1 mostra os nomes e uma breve descrição das funções implementadas.

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 univariadas visam avaliar o efeito das variáveis para cada resposta, as multivariadas visam avaliar o efeito das variáveis explicativas em todas as variáveis resposta simultaneamente. As nomenclaturas seguem o que foi exposto no Capítulo 4 e as funções recebem como argumento apenas o objeto que armazena o modelo devidamente ajustado.

Função	Descrição
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_dispersion()	ANOVA tipo III para dispersão
mc_manova_dispersion()	MANOVA tipo III para dispersão
mc_multcomp()	Testes de comparações múltiplas por resposta
mc_mult_multcomp()	Testes de comparações múltiplas multivariado
mc_linear_hypothesis()	Hipóteses lineares gerais especificadas pelo usuário

Tabela 6.1: Funções implementadas

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

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

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

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

Por fim, a função *mc_linear_hypothesis()* é a implementação computacional em R que permite a execução de qualquer um dos testes apresentados no Capítulo 4. É a função mais flexível que temos no conjunto de implementações. Com ela é possível especificar qualquer tipo de hipótese sobre parâmetros de regressão, dispersão ou potência de um modelo *mcglm*. Também

é possível especificar hipóteses sobre múltiplos parâmetros e o vetor de valores da hipótese nula é definido pelo usuário. Esta função recebe como argumentos o modelo e um vetor contendo os parâmetros que devem ser testados e os valores sob hipótese nula. Com algum trabalho, por meio da função de hipóteses lineares gerais, é possível replicar os resultados obtidos pelas funções de análise de variância.

6.2 INSTALAÇÃO

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

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

As implementações referentes a este trabalho estão disponíveis publicamente na plataforma github em https://github.com/lineu96/htmcglm e podem ser instaladas por meio da função *install_github()* do pacote *devtools*.

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

6.3 EXEMPLOS

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

6.3.1 Exemplo 1: soya

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

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

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

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

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

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

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

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

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

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

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

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

6.3.1.1 Quadros de análise de variância para parâmetros de regressão

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

ANOVA tipo I

A ANOVA tipo I avalia a inclusão de variáveis ao modelo. Na primeira linha é testada se todos os parâmetros são iguais a zero. Na segunda linha é testada a hipótese de que todos os parâmetros exceto intercepto são iguais a zero, ou seja, inclui-se o intercepto ao modelo.

Na terceira, é testada a hipótese de que todos os parâmetros exceto intercepto e os parâmetros associados a bloco são iguais a zero, ou seja, avalia-se a inclusão de intercepto e bloco ao modelo. E assim sucessivamente.

```
mc_anova_I(fit_joint)
```

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

ANOVA tipo II

Já a ANOVA tipo II avalia a exclusão de variáveis no modelo. Deste modo, na primeira linha testa-se se o intercepto é igual a zero. Na segunda, testa-se se o efeito de bloco é igual a zero, ou seja, é avaliado o impacto da retirada da variável bloco do problema. No caso das variáveis referentes aos níveis de água e potássio, como estamos avaliando também a interação entre estes fatores, o teste feito avalia a exclusão destas variáveis; logo, o teste de nulidade avalia tanto os parâmetros de efeitos principais quanto os de interação. E por fim, a retirada do efeito de interação é testada.

```
mc_anova_II(fit_joint)
```

```
## ANOVA type II using Wald statistic for fixed effects
##
## Call: grain ~ block + water * pot
```

```
##
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept  1 102.2961   0.0000
## 2     block  4  14.3051   0.0064
## 3     water 10  84.6677   0.0000
## 4       pot 12 350.9316   0.0000
## 5 water:pot  8  30.4494   0.0002
##
## Call: seeds ~ block + water * pot
##
##   Covariate Df       Chi Pr(>Chi)
## 1 Intercept  1 3993.9442   0.0000
## 2     block  4   11.6363   0.0203
## 3     water 10   70.8041   0.0000
## 4       pot 12  130.2022   0.0000
## 5 water:pot  8   12.7366   0.1212
##
## Call: viablepeasP ~ block + water * pot
##
##   Covariate Df      Chi Pr(>Chi)
## 1 Intercept  1  13.4353   0.0002
## 2     block  4   4.4305   0.3509
## 3     water 10  33.9928   0.0002
## 4       pot 12 295.2420   0.0000
## 5 water:pot  8  20.0549   0.0101
```

ANOVA tipo III

No caso da ANOVA tipo III os efeitos são testados independentemente da inclusão de interação no modelo. Diferente da ANOVA tipo II, nas linhas referentes aos níveis de água e potássio, apenas os parâmetros referentes à água e potássio nos efeitos principais serão testados, mantendo-se as interações.

```
mc_anova_III(fit_joint)
```

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

```
## 2     block  4   11.6363   0.0203
## 3     water  2    3.9399   0.1395
## 4       pot  4   19.1997   0.0007
## 5 water:pot  8   12.7366   0.1212
##
## Call: viablepeasP ~ block + water * pot
##
##   Covariate Df     Chi Pr(>Chi)
## 1 Intercept  1 13.4353   0.0002
## 2     block  4  4.4305   0.3509
## 3     water  2  5.2513   0.0724
## 4       pot  4 71.1026   0.0000
## 5 water:pot  8 20.0549   0.0101
```

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

MANOVA tipo I

```
mc_manova_I(fit_joint)
```

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

MANOVA tipo II

```
mc_manova_II(fit_joint)
```

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

MANOVA tipo III

```
mc_manova_III(fit_joint)
```

```
## MANOVA type III using Wald statistic for fixed effects
##
## Call: ~ block+water*pot
##   Covariate Df       Chi Pr(>Chi)
## 1 Intercept  3 5553.7954   0.0000
## 2     block 12   23.7478   0.0220
## 3     water  6    9.0173   0.1726
## 4       pot 12  149.0321   0.0000
## 5 water:pot 24   68.7879   0.0000
```

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

6.3.1.2 Teste sobre um único parâmetro de regressão

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

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

6.3.1.3 Teste sobre mais de um parâmetro de regressão

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

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

6.3.1.4 Teste de igualdade de efeitos entre parâmetros de regressão

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

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

6.3.2 Exemplo 2: Hunting

O conjunto de dados Hunting, apresentados em Bonat et al. (2017), também está disponível no pacote *mcglm*. Os dados tratam de um problema em que as respostas são contagens bivariadas longitudinais sobre animais caçados na vila de Basile Fang, Bioko Norte Province, Bioko Island, Equatorial Guinea. As variáveis respostas são: números mensais de blue duikers (BD) e outros pequenos animais (OT) baleados ou capturados em uma amostra aleatória de 52 caçadores comerciais de agosto de 2010 a setembro de 2013. Consideremos que o interesse é avaliar o efeito de um fator com 2 níveis que indica se o animal foi caçado por meio de arma de fogo ou armadilha (*METHOD*) e um fator com 2 níveis que indica o sexo do animal (*SEX*).

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

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

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

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

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

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

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

6.3.2.1 Teste sobre um único parâmetro de dispersão

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

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

6.3.2.2 Teste sobre mais de um parâmetro de dispersão

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

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

6.3.2.3 Teste de igualdade de efeitos entre parâmetros de dispersão

```
mc_linear_hypothesis(object =  fit,
                     hypothesis = c('tau12 = tau22'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## 1 tau12 = tau22
##
## Results:
##   Df    Chi Pr(>Chi)
## 1  1 5.8183   0.0159
```

6.3.2.4 *Quadro de análise de variância para parâmetros de dispersão*

As funções para avaliar os parâmetros de dispersão por meio de um procedimento análogo à análise de variância para parâmetros de regressão, requerem a especificação de mais argumentos: um deles que determina a relação entre parâmetros de dispersão e o outro que especifica os nomes que aparecerão na saída final. O resultado final é um quadro que testa a hipótese de nulidade de cada parâmetro ou conjunto de parâmetros de dispersão.

ANOVA tipo III para dispersão

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

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

MANOVA tipo III para dispersão

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

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

6.3.2.5 *Comparações múltiplas*

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

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

```
## Multiple comparisons test for each outcome using Wald statistic
##
## Call: BD ~ METHOD * SEX
##
##                          Contrast Df      Chi Pr(>Chi)
## 1 Escopeta:Female-Escopeta:Male  1 175.7657        0
## 2 Escopeta:Female-Trampa:Female  1  20.1379        0
## 3   Escopeta:Female-Trampa:Male  1  35.6372        0
## 4     Escopeta:Male-Trampa:Male  1  24.3946        0
## 5   Trampa:Female-Escopeta:Male  1 217.7398        0
## 6     Trampa:Female-Trampa:Male  1 132.6125        0
##
## Call: OT ~ METHOD * SEX
##
##                          Contrast Df     Chi Pr(>Chi)
## 1 Escopeta:Female-Escopeta:Male  1 14.3969   0.0009
## 2 Escopeta:Female-Trampa:Female  1  6.5843   0.0617
## 3   Escopeta:Female-Trampa:Male  1  5.6455   0.1050
## 4     Escopeta:Male-Trampa:Male  1  0.7480   1.0000
## 5   Trampa:Female-Escopeta:Male  1 31.3069   0.0000
## 6     Trampa:Female-Trampa:Male  1 25.3203   0.0000
```

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

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

```
## Multivariate multiple comparisons test using Wald statistic
##
## Call: ~ METHOD*SEX
##                          Contrast Df      Chi Pr(>Chi)
## 1 Escopeta:Female-Escopeta:Male  2 215.0490        0
## 2 Escopeta:Female-Trampa:Female  2  31.8503        0
## 3   Escopeta:Female-Trampa:Male  2  47.8804        0
## 4     Escopeta:Male-Trampa:Male  2  27.5459        0
## 5   Trampa:Female-Escopeta:Male  2 287.6161        0
## 6     Trampa:Female-Trampa:Male  2 184.8844        0
```

Com isso, descrevemos a implementação de procedimentos em R para realizar testes de hipóteses em parâmetros de McGLMs com base na estatística Wald. Os exemplos discutidos ilustram como avaliar as hipóteses mais comuns que surgem em problemas de regressão: avaliar parâmetros individualmente e avaliar conjuntos de parâmetros. Concentramos nossos esforços em ferramentas para avaliar parâmetros de regressão e dispersão, pois estudando os parâmetros de regressão é possível identificar as variáveis que têm efeito significativo na resposta; por outro lado, os parâmetros de dispersão permitem avaliar se há efeito de observações correlacionadas. Dessa forma, o estudo dessas quantidades fornece informações valiosas sobre a importância dos elementos de um problema de regressão multivariada.

7 ANÁLISE DE DADOS

Este capítulo é destinado à análise de dados reais fazendo uso do McGLM e dos testes de hipóteses propostos e implementados neste trabalho. O conjunto de dados utilizado diz respeito a um estudo que visa avaliar se o uso de probióticos é capaz de controlar vício e transtorno da compulsão alimentar em pacientes submetidos à cirurgia bariátrica.

7.1 CONTEXTO

A cirurgia bariátrica é conhecida como o método mais efetivo no tratamento de obesidade severa pois resulta em considerável perda de peso, remissão de comorbidades e melhora da qualidade de vida dos indivíduos (Mechanick et al., 2020).

Os transtornos alimentares são considerados doenças psiquiátricas graves, caracterizadas por comportamento alimentar anormal (Treasure et al., 2020), dentre os quais transtornos comuns são transtorno da compulsão alimentar e vício alimentar.

O transtorno da compulsão alimentar é reconhecido como um transtorno caracterizado pelo consumo de grandes quantidades de alimentos em um curto período de tempo e uma sensação de perda de controle sobre a alimentação durante esses episódios, associada a angústia e arrependimento ao indivíduo (Sarmiento e Lau, 2020), (Wilfley et al., 2016).

Já o vício alimentar está associado ao consumo de alimentos que, em geral são altamente palatáveis (densos em energia, ricos em açúcar, gordura e/ou sal), e por este motivo, excessivamente estimulantes para as vias de recompensa do cérebro, que podem promover o desejo incontrolável e insaciável de continuar comendo e desencadear uma série de sintomas (Avena et al., 2012), (Najem et al., 2020).

Uma nova abordagem que surge para o tratamento de transtornos psiquiátricos é o uso de probióticos e prebióticos como moduladores do eixo microbiota-intestino-cérebro, também conhecidos como psicobióticos (Dinan et al., 2013), (Mason, 2017), (Misra e Mohanty, 2019).

No entanto, ainda faltam estudos que avaliem a influência da suplementação de probióticos em fatores psicológicos ou comportamentais em indivíduos submetidos à cirurgia bariátrica. Portanto, o objetivo deste estudo foi analisar a influência da suplementação de probióticos no transtorno da compulsão alimentar e no vício alimentar em indivíduos submetidos ao Bypass Gástrico em Y de Roux (RYGB).

7.2 DESENHO EXPERIMENTAL E COLETA DE DADOS

Trata-se de um ensaio clínico randomizado, duplo-cego, controlado por placebo, realizado com pacientes submetidos ao RYGB no período de abril de 2018 a dezembro de 2019. O estudo foi aprovado pelo Comitê de Ética em Pesquisa da Pontifícia Universidade Católica do Paraná (PUCPR) (nº 4.252.808) e registrado pelo Registro Brasileiro de Ensaios Clínicos - REBEC (nºRBR-4x3gqp). A pesquisa foi explicada a cada participante antes de sua participação e, daqueles que concordaram, foi obtido o consentimento informado por escrito.

A divisão dos grupos (placebo ou probiótico) foi feita de forma aleatória. Os critérios de inclusão dos indivíduos no estudo foram: adultos (18-59 anos) que fariam RYGB, com índice de massa corporal (IMC) $\geq$ 35 kg/m2 e que não usaram antibióticos antes da cirurgia.

Foram retirados do estudo pacientes que foram submetidos a outras técnicas cirúrgicas ou reoperação, tiveram complicações pós-cirúrgicas, fizeram antibioticoterapia concomitante ao

uso de probiótico/placebo ou não usaram os comprimidos de probiótico/placebo por mais de nove dias consecutivos.

Ambos os grupos receberam as mesmas orientações alimentares após a cirurgia, foram acompanhados pela mesma equipe cirúrgica (médico, nutricionista e psicólogo) e tiveram o mesmo número de consultas pré-agendadas antes e após a cirurgia, seguindo o protocolo estabelecido pela instituição onde o estudo foi realizado.

No sétimo dia de pós-operatório, os participantes foram orientados a ingerir dois comprimidos mastigáveis/dia de placebo ou comprimido probiótico Flora Vantage, 5 bilhões de Lactobacillus acidophilus NCFM ®Strain e 5 bilhões de Bifidobacterium lactis Bi-07 ®) da Bariatric Advantage (Aliso Viejo, CA, EUA) por 90 dias.

Os indivíduos foram avaliados em 3 momentos. A primeira avaliação (T0) foi realizada aproximadamente 10 dias antes da cirurgia. As avaliações de acompanhamento foram realizadas aproximadamente três meses (T1) e um ano de pós-operatório (T2). Nestes momentos, foram realizadas avaliações clínicas e antropométricas, bem como os questionários autoaplicáveis foram entregues aos participantes a cada encontro.

A avaliação do transtorno da compulsão alimentar foi feita com base na escala de compulsão alimentar (BES), uma das ferramentas mais usadas para medir a compulsão alimentar em pesquisa, com inúmeros trabalhos comprovando sua eficácia, traduzido para o português e validado para indivíduos com obesidade e submetidos a cirurgia bariátrica.

Trata-se de um questionário em formato de escala likert de 16 itens, elaborado de acordo com o Manual Diagnóstico e Estatístico de Transtornos Mentais (3ª edição) (Spitzer et al., 1980) por Gormally et al. (1982). Os indivíduos foram orientados a selecionar a opção que melhor representasse sua resposta e o escore final foi a soma dos pontos de cada item, este escore varia de 0 a 46.

Para avaliação de vício alimentar foi utilizada a escala de vício alimentar (YFAS), um questionário que busca detectar sintomas de comportamentos alimentares aditivos. O YFAS foi baseado nos critérios de dependência de substâncias do Manual Diagnóstico e Estatístico IV – Revisão de Texto (DSM-IV-TR) (Segal, 2010) e endossado para alimentos altamente processados. Este questionário foi desenvolvido por Gearhardt et al. (2009). O questionário é uma combinação de 25 opções em escala Likert e a opção de avaliação utilizada foi o número de sintomas de vício.

7.3 CONJUNTO DE DADOS

A amostra final é formada por 71 indivíduos, dos quais 33 pertencem ao grupo placebo e 38 ao grupo tratamento. Se todos estes indivíduos fossem avaliados nos 3 momentos definidos no estudo, o conjunto de dados teria 213 observações. Contudo, ao longo do estudo, diversos indivíduos não compareceram às consultas, o que faz com que haja dados faltantes no conjunto de dados. Após tratamento dos dados e exclusão de observações faltantes restaram 184 observações.

É importante notar que trata-se de um problema com duas variáveis resposta: um escore que caracteriza compulsão e o número de sintomas apresentados que caracterizam vício. Além disso, as observações não são independentes, tendo em vista que medidas tomadas em um mesmo indivíduo são correlacionadas. Portanto trata-se de um problema que técnicas de modelagem tradicionais seriam de difícil aplicação mas é um cenário ideal para resolução via McGLM e no qual testes de hipóteses são de extrema importância para avaliar o efeito da interação entre momento e uso do probiótico sobre vício e compulsão alimentar.

Para fins de análise, o escore que caracteriza compulsão e o número de sintomas apresentados que caracterizam vício foram transformados para a escala unitária, considerando

que tratam-se de variáveis restritas. O objetivo da análise é avaliar o efeito de momento e grupo nas métricas de vício e compulsão. O conjunto de dados contém as seguintes variáveis:

- id: variável identificadora de indivíduo.
- momento: variável identificadora de momento (T0, T1, T2).
- grupo: variável identificadora de grupo (placebo, probiótico)
- YFAS: proporção de sintomas que caracterizam vício.
- BES: proporção de escore de compulsão.

7.4 ANÁLISE EXPLORATÓRIA

A análise gráfica apresentada na Figura 7.1 mostra, em (a) e (d) que ambas as variáveis de interesse apresentam considerável assimetria à direita. Os boxplots das métricas em função de grupo apresentados em (b) e (e) mostram sensíveis diferenças entre o grupo placebo e probiótico para ambas as respostas. Já os boxplots das métricas em função dos momentos de avaliação, apresentados em (c) e (f), evidenciam que para ambas as métricas os valores eram mais altos no momento T0, havendo considerável redução no momento T1. Quando comparamos T1 e T2, para YFAS parece que há um leve aumento na última avaliação; já para BES, T1 e T2 não parecem diferir.

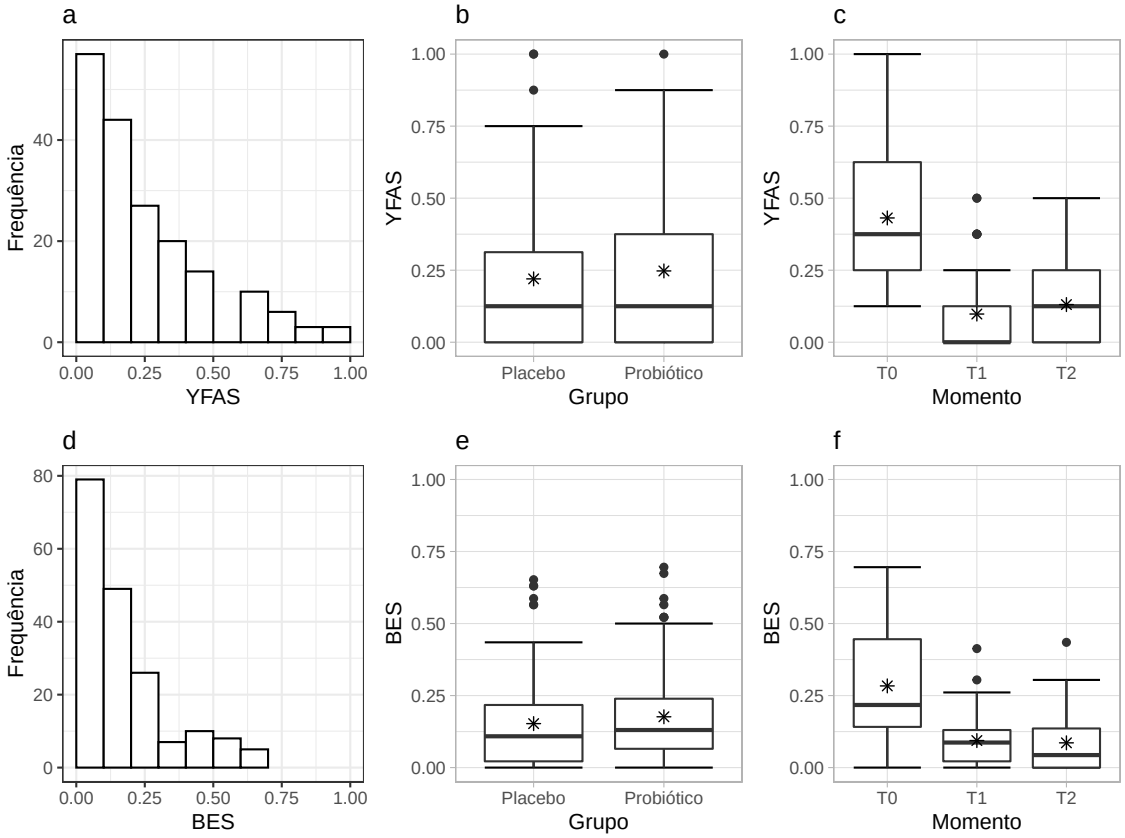


Figura 7.1: Análise exploratória gráfica: (a) histograma YFAS, (b) boxplots YFAS em função de grupo, (c) boxplots YFAS em função de momento, (d) histograma BES, (b) boxplots BES em função de grupo, (c) boxplots BES em função de momento. O asterísco nos boxplots indica a média.

Ainda de forma exploratória podemos avaliar o comportamento das métricas de vício e compulsão por meio da avaliação das medidas descritivas por momento e por grupo, apresentadas na Tabela 7.1. É possível verificar a redução de indivíduos ao longo do tempo, algo comum em estudos prospectivos. Quanto às medidas, nota-se que ambos os grupos (placebo ou probiótico) apresentam médias mais altas no momento T0 do que nos demais momentos. Portanto, existe uma clara redução das métricas quando comparado ao pré operatório. Quando comparamos os momentos pós-operatório (T1 e T2) verificamos que as medidas de YFAS, tanto para o grupo placebo quanto para o grupo probiótico apresentam, em média, um aumento das métricas no último momento de avaliação; o mesmo pode ser observado para BES no grupo placebo. Já as medidas de BES no grupo probiótico apresentam queda.

Grupo	Momento	n	YFAS	BES
			Média (desvio padrão)	Média (desvio padrão)
Placebo	T0	33	0,37 (0,26)	0,24 (0,20)
Placebo	T1	32	0,11 (0,15)	0,09 (0,10)
Placebo	T2	22	0,16 (0,15)	0,10 (0,12)
Probiótico	T0	38	0,49 (0,24)	0,32 (0,18)
Probiótico	T1	37	0,09 (0,12)	0,10 (0,08)
Probiótico	T2	22	0,10 (0,14)	0,07 (0,09)

Tabela 7.1: Número de indivíduos, média e desvio padrão para YFAS e BES para cada combinação de grupo e momento.

7.5 ESPECIFICAÇÃO DO MODELO

Para análise dos dados foi ajustado um modelo multivariado com os efeitos fixos das variáveis momento e grupo. Adicionalmente, foi incluído no modelo o efeito da interação entre estas duas variáveis explicativas.

Como já mencionado, trata-se de um experimento em que as observações não são independentes pois medidas tomadas em um mesmo indivíduo são correlacionadas e esta correlação deve ser especificada no modelo.

Como mencionado na Seção 7.3, ambas as respostas foram transformadas para a escala unitária, tendo em vista que se tratam originalmente de números inteiros restritos ao intervalo. Por este motivo foi utilizada a função de ligação logito com função de variância binomial. Adicionalmente, estimou-se o parâmetro de potência para todas as respostas em análise. Os preditores lineares são dados por

$$g_r(\mu_r) = \beta_{0r} + \beta_{1r}T1 + \beta_{2r}T2 + \beta_{3r}Probiotico + \beta_{4r}T1 * Probiotico + \beta_{5r}T2 * Probiotico,$$

em que o índice r refere-se às variáveis respostas do estudo (1 para YFAS, 2 para BES). Foram consideradas categorias de referência o grupo placebo e o momento T0. β_{0r} representa o intercepto, β_{1r} o efeito do momento T1, β_{2r} o efeito do momento T2, β_{3r} o efeito de probiótico. Os parâmetros β_{4r} e β_{5r} referem-se à interação entre momento e grupo, de tal forma que β_{4r} representa o efeito da interação entre T1 e probiótico, e β_{5r} representa o efeito da interação entre T2 e probiótico.

Os preditores matriciais, iguais para ambas as respostas, são dados por $h\{\boldsymbol{\Omega}(\boldsymbol{\tau})\} = \tau_0 Z_0 + \tau_1 Z_1$. A função $h(.)$ utilizada foi a identidade, τ_0 e τ_1 representam os parâmetros de dispersão, Z_0 representa uma matriz identidade de ordem, 184×184 e Z_1 representa uma matriz

de dimensão 184×184 especificada de forma a explicitar que as medidas provenientes do mesmo indivíduo são correlacionadas.

Para exemplificar a forma do preditor matricial, vamos considerar 3 indivíduos: A, B e C. Suponha que o indivíduo A compareceu às 3 consultas, portanto temos informações deste indivíduo em T0, T1 e T2. O indivíduo B compareceu em T0 e T1. Já o indivíduo C compareceu apenas em T0. Deste modo temos 3 indivíduos e 6 observações. Logo Z_0 é uma matriz identidade 6×6 e Z_1 é uma espécie de matriz bloco diagonal em que o tamanho dos blocos varia de acordo com o número de medidas para cada indivíduo. Neste cenário o preditor matricial tem a forma

$$h\left\{\boldsymbol{\Omega}(\boldsymbol{\tau})\right\} = \tau_0 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{7.1}$$

7.6 RESULTADOS DO AJUSTE

Com o propósito de verificar a qualidade do ajuste do modelo, foi feita a análise de resíduos do modelo. A análise mostra que os resíduos de Pearson para YFAS e BES apresentam média 0 e desvio padrão próximo de 1. Na Figura 7.2 são exibidos os histogramas dos resíduos de Pearson por resposta, a distribuição dos resíduos é aproximadamente simétrica com a maior parte dos dados entre -2 e 2.

Na Figura 7.3 são exibidos os resíduos versos preditos do modelo. Os resultados mostram que não parece haver qualquer tipo de relação entre resíduos e preditos. De forma geral, o modelo parece estar razoavelmente bem ajustado aos dados.

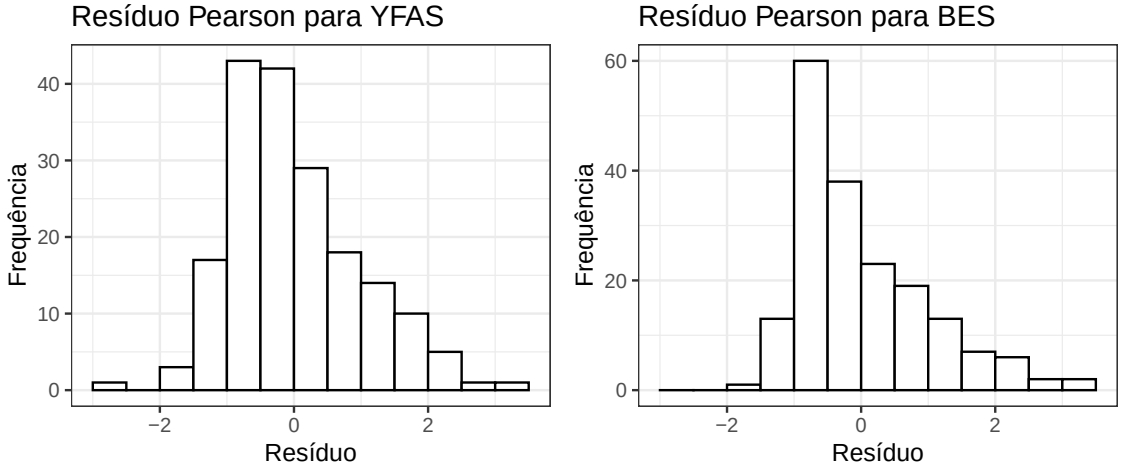


Figura 7.2: Histograma dos resíduos de Pearson por resposta.

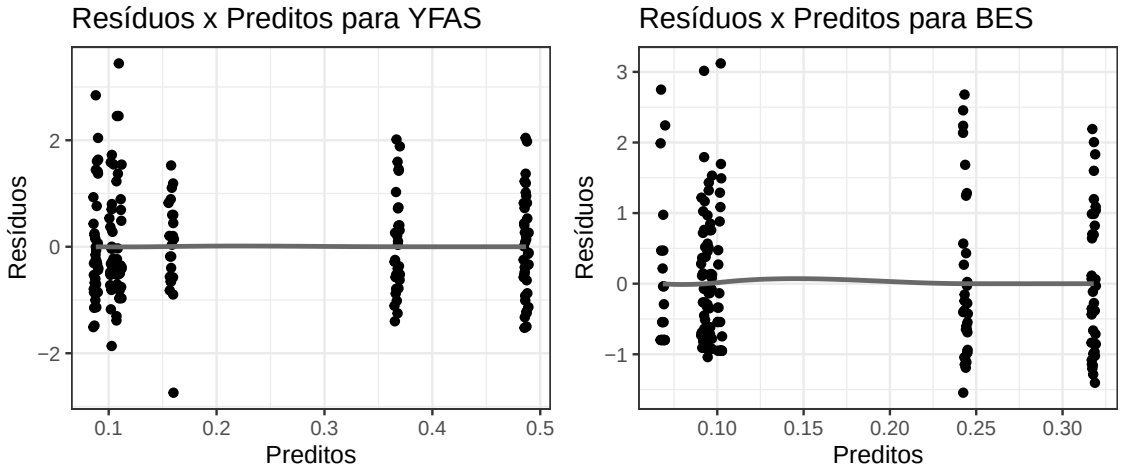


Figura 7.3: Gráfico de resíduos Pearson versus preditos com linha de tendência suave para cada resposta.

As estimativas dos parâmetros, intervalos de confiança assintóticos com 95% de confiança e valores-p da hipótese de nulidade dos parâmetros são mostrados na Tabela 7.2. Adicionalmente, a Figura 7.4 mostra os valores preditos para cada combinação dos fatores para uma melhor interpretação dos resultados.

Parâmetro	YFAS			BES		
	Estimativa	Intervalo de confiança	Valor-p	Estimativa	Intervalo de confiança	Valor-p
β_0	-0,54	(-0,87;-0,22)	<0,01	-1,13	(-1,44;-0,83)	<0,01
β_1	-1,55	(-2,17;-0,94)	<0,01	-1,16	(-1,62;-0,69)	<0,01
β_2	-1,13	(-1,75;-0,51)	<0,01	-1,05	(-1,58;-0,52)	<0,01
β_3	0,49	(0,05;0,93)	0,0284	0,37	(-0,03;0,77)	0,0733
β_4	-0,73	(-1,60;0,14)	0,0	-0,33	(-0,96;0,30)	0,3081
β_5	-0,98	(-1,93;-0,03)	0,0429	-0,80	(-1,58;-0,02)	0,0449
τ_0	0,18	(0,01;0,35)	0,0411	0,17	(0,00;0,34)	0,0458
$\tau 1$	0,01	(-0,02;0,04)	0,5718	0,04	(-0,01;0,10)	0,1357
p	0,91	(0,47;1,34)	<0,05	1,23	(0,77;1,68)	<0,05

Tabela 7.2: Estimativas dos parâmetros, intervalos com 95% de confiança e valores-p do modelo.

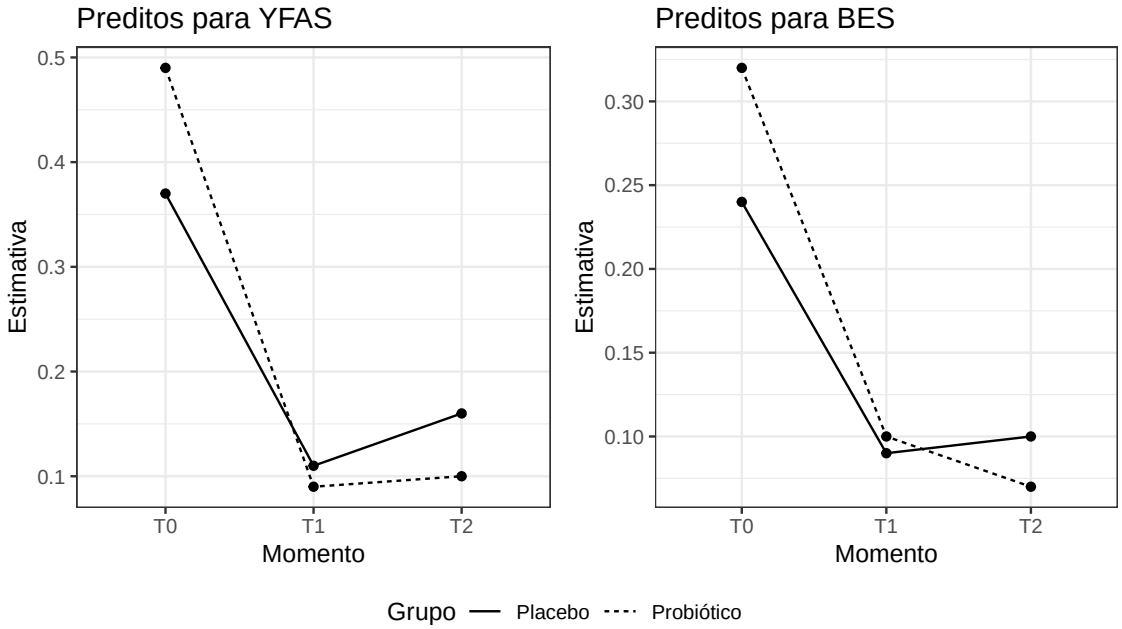


Figura 7.4: Gráfico de preditos pelo modelo para cada combinação entre momento e grupo.

7.7 TESTES DE HIPÓTESES

Até este ponto foi apresentada uma análise padrão, com os resultados usuais da análise de um modelo de regressão. Indo mais além nesta análise, podemos fazer uso dos resultados e implementações deste trabalho.

Podemos optar por uma análise de variância multivariada do tipo II para avaliar a importância das variáveis no problema. O resultado, apresentado na tabela Tabela 7.3 aponta para existência significativa do efeito de momento e ausência de efeito de grupo, indicando que para ambas as respostas as métricas se alteram ao longo do tempo mas sem alteração entre grupos.

Variável	Graus de liberdade	Qui-quadrado	Valor-p
Intercepto	2	53,1581	<0,01
Momento	8	139,0161	<0,01
Grupo	6	8,4928	0,2042
Momento*Grupo	4	6,9923	0,1363

Tabela 7.3: Análise de variância multivariada tipo II.

A fim de avaliar os resultados por resposta, podemos utilizar uma análise de variância univariada do tipo II. Os resultados são apresentados na Tabela 7.4. Considerando um nível de significância de $0,01$, existe evidência que aponta para efeito de momento.

		YFAS		BES	
Variável	Graus de liberdade	Qui-quadrado	Valor-p	Qui-quadrado	Valor-p
Intercepto	1	10,6128	<0,01	53,1473	<0,01
Momento	4	102,9875	<0,01	99,5681	<0,01
Grupo	3	6,6837	0,0827	5,3083	0,1506
Momento*Grupo	2	5,5984	0,0609	4,2477	0,1196

Tabela 7.4: Análise de variância univariada do tipo II.

Como as análises de variâncias apontaram para efeito significativo de variáveis categóricas, podemos explorar quais níveis diferem entre si. A Tabela 7.5 apresenta as comparações duas a duas entre momentos. Os resultados mostram que, para ambas as respostas existem diferenças entre o primeiro versus segundo e primeiro versus terceiro momento, mas os dois últimos momentos não diferem entre si.

Contraste	Graus de liberdade	Qui-quadrado	Valor-p
T0-T1	2	97,9874	<0,01
T0-T2	2	67,2462	<0,01
T1-T2	2	2,4730	0,8712

Tabela 7.5: Comparações duas a duas entre momentos para ambas as respostas.

A tabela Tabela 7.6 apresenta as comparações entre grupos para cada momento para ambas as respostas. Os resultados apontam para a ausência de diferença entre grupos em cada momento.

Contraste	Graus de liberdade	Qui-quadrado	Valor-p
T0:Placebo-T0:Probiótico	2	5,5819	0,9204
T1:Placebo-T1:Probiótico	2	0,6096	1
T2:Placebo-T2:Probiótico	2	1,7645	1

Tabela 7.6: Comparações duas a duas entre grupos para cada momento para ambas as respostas.

No modelo, incluímos a informação de que existem medidas que foram tomadas em um mesmo indivíduo. Esta informação é declarada no preditor matricial que estima um parâmetro de dispersão associado à matriz que aponta a relação entre os indivíduos. Uma hipótese de interesse pode ser avaliar se existe evidência para crer que, neste problema, as medidas tomadas em um mesmo indivíduo são de fato correlacionadas. Para isso podemos postular hipóteses sobre os parâmetros de dispersão. Tal como nas análises de variância, isso pode ser feito por resposta ou para ambas as respostas simultaneamente.

A Tabela 7.7 apresenta os resultados do teste multivariado, ou seja, avalia a hipótese de que em ambas as respostas as medidas sejam correlacionadas. Os resultados apontam que não há evidência para crer que as medidas tomadas em um mesmo indivíduo apresentam correlação.

Variável	Graus de liberdade	Qui-quadrado	Valor-p
τ_0	2	7,1936	0,0274
τ_1	2	2,3201	0,3135

Tabela 7.7: Análise de variância multivariada do tipo III para parâmetros de dispersão.

Com isso, apresentamos uma análise de dados real, em que o uso de testes de hipóteses para avaliação de parâmetros de um modelo multivariado se mostrou de grande valia para extrair o máximo de informações a respeito do problema. Exploramos testes sobre parâmetros de regressão e dispersão a fim de concluir investigar os elementos associados ao desfecho dos fenômenos sob análise.

8 CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi desenvolver procedimentos para realizar testes de hipóteses sobre parâmetros de McGLMs baseados na estatística de Wald. McGLMs contam com parâmetros de regressão, dispersão, potência e correlação; cada conjunto de parâmetros possui uma interpretação prática bastante relevante no contexto de análise de problemas com potenciais múltiplas respostas em função de um conjunto de variáveis explicativas.

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

As propriedades dos testes foram avaliadas com base em estudos de simulação. Foram considerados cenários univariados e trivariados com diferentes distribuições de probabilidade para as respostas e diferentes tamanhos amostrais. A ideia desta etapa do trabalho foi gerar conjuntos de dados com parâmetros de regressão e dispersão fixados e testar hipóteses sobre parâmetros de modelos ajustados com estes dados.

Em um primeiro momento, testamos a hipótese de que os parâmetros eram realmente iguais aos fixados na simulação. Em seguida afastamos gradativamente as hipóteses dos valores simulados a fim de verificar se, à medida que afasta-se a hipótese dos verdadeiros valores, o percentual de rejeição aumenta.

8.1 CONCLUSÕES GERAIS

De modo geral, os estudos de simulação mostraram que quanto mais distante a hipótese é dos valores inicialmente simulados, maior é o percentual de rejeição. Tal como esperado, os menores percentuais foram observados na hipótese igual aos valores simulados e também foi possível verificar que conforme aumenta-se o tamanho amostral, o percentual de rejeição aumenta para hipóteses pouco diferentes dos valores simulados dos parâmetros, indicando que o poder do teste cresce à medida que a amostra aumenta.

Sendo assim, os resultados das simulações mostraram que o teste Wald pode ser usado para avaliar hipóteses sobre parâmetros de regressão e dispersão de McGLMs, o que permite uma melhor interpretação do efeito das variáveis e estruturas de delineamento em contextos práticos.

Adicionalmente, fizemos a aplicação das metodologias propostas a um conjunto de dados real em que o objetivo é avaliar o efeito do uso de probióticos no controle de vícios e compulsões alimentares. Trata-se de um problema com duas variáveis resposta: um escore que caracteriza compulsão e o número de sintomas apresentados que caracterizam vício. Para fins de análise, como ambas as respostas são restritas a um intervalo, foram transformadas para a escala unitária.

Neste estudo, um conjunto de indivíduos foi dividido em dois grupos: um deles recebeu um placebo e o outro recebeu o tratamento. Além disso os indivíduos foram avaliados ao longo tempo; deste modo o delineamento gera observações que não são independentes, já que medidas tomadas em um mesmo indivíduo tendem a ser correlacionadas.

Este é um problema em que técnicas de modelagem tradicionais seriam de difícil aplicação. Contudo trata-se de um problema de possível análise via McGLM e testes de hipóteses

podem ser empregados para avaliar o efeito da interação entre momento e uso do probiótico sobre vício e compulsão alimentar.

Os resultados, baseados nos testes propostos neste trabalho indicam que existe evidência que aponta para efeito de momento, ou seja, vício e compulsão alimentar alteram-se ao longo do tempo. Os testes de comparações múltiplas indicam que, para ambas as respostas, existem diferenças entre o primeiro versus segundo e primeiro versus terceiro momento, mas os dois últimos momentos não diferem entre si. Os resultados também apontam para a ausência de diferença entre grupos em cada momento. Uma avaliação dos parâmetros de dispersão mostra que não há evidência para crer que as medidas tomadas em um mesmo indivíduo apresentam correlação.

8.2 LIMITAÇÕES

Algumas limitações deste trabalho dizem respeito a casos não explorados nos estudos de simulação, tais como: avaliação do desempenho dos testes ao definir hipóteses lineares que combinem parâmetros de diferentes tipos, impacto de um número diferente de observações por indivíduos em problemas longitudinais ou de medidas repetidas, impacto no poder do teste conforme o número de parâmetros testados aumenta e o comportamento do teste em problemas multivariados com distribuições de probabilidade diferentes das exploradas.

8.3 TRABALHOS FUTUROS

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

REFERÊNCIAS

Adeleke, B. L., Yahaya, W. e Usman, A. (2014). A comparison of some test statistics for multivariate analysis of variance model with non-normal responses.

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.

Avena, N. M., Gold, J. A., Kroll, C. e Gold, M. S. (2012). Further developments in the neurobiology of food and addiction: update on the state of the science. *Nutrition*, 28(4):341–343.

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

Baldessari, B. (1967). The distribution of a quadratic form of normal random variables. *The Annals of Mathematical Statistics*, 38(6):1700–1704.

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

Berliana, S. M., Purhadi, Sutikno e Rahayu, S. P. (2019). Multivariate generalized poisson regression model with exposure and correlation as a function of covariates: Parameter estimation and hypothesis testing. Em *AIP Conference Proceedings*, volume 2192, página 090001. AIP Publishing LLC.

Berridge, D. M. e Crouchley, R. (2019). *Multivariate generalized linear mixed models using R*. CRC Press.

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

Bonat, 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.

Bretz, F., Hothorn, T. e Westfall, P. (2008). Multiple comparison procedures in linear models. Em *COMPSTAT 2008*, páginas 423–431. Springer.

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

Cario, M. C. e Nelson, B. L. (1997). Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Relatório técnico, Citeseer.

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

Dinan, T. G., Stanton, C. e Cryan, J. F. (2013). Psychobiotics: a novel class of psychotropic. *Biological psychiatry*, 74(10):720–726.

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.

Gearhardt, A. N., Corbin, W. R. e Brownell, K. D. (2009). Preliminary validation of the yale food addiction scale. *Appetite*, 52(2):430–436.

Genz, A. e Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. e Hothorn, T. (2021). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.1-3.

Gormally, J., Black, S., Daston, S. e Rardin, D. (1982). The assessment of binge eating severity among obese persons. *Addictive behaviors*, 7(1):47–55.

Graybill, F. A. e Marsaglia, G. (1957). Idempotent matrices and quadratic forms in the general linear hypothesis. *The Annals of Mathematical Statistics*, 28(3):678–686.

Hand, D. J. e Taylor, C. C. (1987). *Multivariate analysis of variance and repeated measures: a practical approach for behavioural scientists*, volume 5. CRC press.

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

Hothorn, T., Bretz, F. e Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363.

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

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

Jiang, J. e Nguyen, T. (2007). *Linear and generalized linear mixed models and their applications*, volume 1. Springer.

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.

Laird, N. M. e Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, páginas 963–974.

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

Lee, Y. e Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(4):619–656.

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.

Luther, N. Y. (1965). Decomposition of symmetric matrices and distributions of quadratic forms. *The Annals of Mathematical Statistics*, páginas 683–690.

Mardalena, S., Purhadi, P., Purnomo, J. D. T. e Prastyo, D. D. (2020). Parameter estimation and hypothesis testing of multivariate poisson inverse gaussian regression. *Symmetry*, 12(10):1738.

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

Mason, B. L. (2017). Feeding systems and the gut microbiome: gut-brain interactions with relevance to psychiatric conditions. *Psychosomatics*, 58(6):574–580.

Mechanick, J. I., Apovian, C., Brethauer, S., Garvey, W. T., Joffe, A. M., Kim, J., Kushner, R. F., Lindquist, R., Pessah-Pollack, R., Seger, J. et al. (2020). Clinical practice guidelines for the perioperative nutrition, metabolic, and nonsurgical support of patients undergoing bariatric procedures–2019 update: cosponsored by american association of clinical endocrinologists/american college of endocrinology, the obesity society, american society for metabolic & bariatric surgery, obesity medicine association, and american society of anesthesiologists. *Surgery for Obesity and Related Diseases*, 16(2):175–247.

Misra, S. e Mohanty, D. (2019). Psychobiotics: A new approach for treating mental illness? *Critical reviews in food science and nutrition*, 59(8):1230–1236.

Najem, J., Saber, M., Aoun, C., El Osta, N., Papazian, T. e Khabbaz, L. R. (2020). Prevalence of food addiction and association with stress, sleep quality and chronotype: A cross-sectional survey among university students. *Clinical nutrition*, 39(2):533–539.

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.

Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological bulletin*, 83(4):579.

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

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

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.

Rahayu, A., Prastyo, D. D. et al. (2020). Multivariate gamma regression: Parameter estimation, hypothesis testing, and its application. *Symmetry*, 12(5):813.

Rao, C. R. (1948a). 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.

Rao, C. R. (1948b). Tests of significance in multivariate analysis. *Biometrika*, 35(1/2):58–79.

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.

Sari, D. N., Purhadi, P., Rahayu, S. P. e Irhamah, I. (2021). Estimation and hypothesis testing for the parameters of multivariate zero inflated generalized poisson regression model. *Symmetry*, 13(10):1876.

Sarmiento, C. e Lau, C. (2020). Diagnostic and statistical manual of mental disorders: Dsm-5. *The Wiley Encyclopedia of Personality and Individual Differences: Personality Processes and Individual Differences*, páginas 125–129.

Segal, D. L. (2010). Diagnostic and statistical manual of mental disorders (dsm-iv-tr). *The Corsini Encyclopedia of Psychology*, páginas 1–3.

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

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

Smaga, Ł. (2017). Bootstrap methods for multivariate hypothesis testing. *Communications in Statistics-Simulation and Computation*, 46(10):7654–7667.

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

Spitzer, R. L., Md, K. K. e Williams, J. B. (1980). Diagnostic and statistical manual of mental disorders. Em *American psychiatric association*. Citeseer.

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

Stasinopoulos, D. M. e Rigby, R. A. (2008). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23:1–46.

Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. CRC press.

Su, P. (2014). *NORTARA: Generation of Multivariate Data with Arbitrary Marginals*. R package version 1.0.0.

Treasure, J., Duarte, T. e Schmidt, U. (2020). Eating disorders. the lancet.

Tuerlinckx, F., Rijmen, F., Verbeke, G. e De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59(2):225–255.

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.

Wilfley, D. E., Citrome, L. e Herman, B. K. (2016). Characteristics of binge eating disorder in relation to diagnostic criteria. *Neuropsychiatric disease and treatment*.

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.

Zhang, Y., Zhou, H., Zhou, J. e Sun, W. (2017). Regression models for multivariate count data. *Journal of Computational and Graphical Statistics*, 26(1):1–13.

APÊNDICE A – HIPÓTESES TESTADAS NO ESTUDO DE SIMULAÇÃO

Hipótese Nula	β_0	β_1	β_2	β_3
H_{01}	5	0	0	0
H_{02}	4,85	0,05	0,05	0,05
H_{03}	4,7	0,1	0,1	0,1
H_{04}	4,55	0,15	0,15	0,15
H_{05}	4,4	0,2	0,2	0,2
H_{06}	4,25	0,25	0,25	0,25
H_{07}	4,1	0,3	0,3	0,3
H_{08}	3,95	0,35	0,35	0,35
H_{09}	3,8	0,4	0,4	0,4
H_{10}	3,65	0,45	0,45	0,45
H_{11}	3,5	0,5	0,5	0,5
H_{12}	3,35	0,55	0,55	0,55
H_{13}	3,2	0,6	0,6	0,6
H_{14}	3,05	0,65	0,65	0,65
H_{15}	2,9	0,7	0,7	0,7
H_{16}	2,75	0,75	0,75	0,75
H_{17}	2,6	0,8	0,8	0,8
H_{18}	2,45	0,85	0,85	0,85
H_{19}	2,3	0,9	0,9	0,9
H_{20}	2,15	0,95	0,95	0,95

Tabela A.1: Hipóteses testadas para parâmetros de regressão nos modelos com resposta seguindo distribuição Normal.

Hipótese Nula	β_0	β_1	β_2	β_3
H_{01}	2,3	0	0	0
H_{02}	2,25	0,017	0,017	0,017
H_{03}	2,2	0,033	0,033	0,033
H_{04}	2,15	0,05	0,05	0,05
H_{05}	2,10	0,067	0,067	0,067
H_{06}	2,05	0,083	0,083	0,083
H_{07}	2	0,1	0,1	0,1
H_{08}	1,95	0,117	0,117	0,117
H_{09}	1,9	0,133	0,133	0,133
H_{10}	1,85	0,15	0,15	0,15
H_{11}	1,8	0,167	0,167	0,167
H_{12}	1,75	0,167	0,167	0,167
H_{13}	1,7	0,2	0,2	0,2
H_{14}	1,65	0,217	0,217	0,217
H_{15}	1,6	0,233	0,233	0,233
H_{16}	1,55	0,25	0,25	0,25
H_{17}	1,5	0,267	0,267	0,267
H_{18}	1,45	0,283	0,283	0,283
H_{19}	1,4	0,3	0,3	0,3
H_{20}	1,35	0,317	0,317	0,317

Tabela A.2: Hipóteses testadas para parâmetros de regressão nos modelos com resposta seguindo distribuição Poisson.

Hipótese Nula	β_0	β_1	β_2	β_3
H_{01}	0,5	0	0	0
H_{02}	0,250	0,083	0,083	0,083
H_{03}	0	0,167	0,167	0,167
H_{04}	-0,25	0,25	0,25	0,25
H_{05}	-0,500	0,333	0,333	0,333
H_{06}	-0,750	0,417	0,417	0,417
H_{07}	-1,0	0,5	0,5	0,5
H_{08}	-1,250	0,583	0,583	0,583
H_{09}	-1,500	0,667	0,667	0,667
H_{10}	-1,75	0,75	0,75	0,75
H_{11}	-2,000	0,833	0,833	0,833
H_{12}	-2,250	0,917	0,917	0,917
H_{13}	-2,5	1,0	1,0	1,0
H_{14}	-2,750	1,083	1,083	1,083
H_{15}	-3,000	1,167	1,167	1,167
H_{16}	-3,25	1,25	1,25	1,25
H_{17}	-3,500	1,333	1,333	1,333
H_{18}	-3,750	1,417	1,417	1,417
H_{19}	-4,0	1,5	1,5	1,5
H_{20}	-4,250	1,583	1,583	1,583

Tabela A.3: Hipóteses testadas para parâmetros de regressão nos modelos com resposta seguindo distribuição Bernoulli.

Hipótese Nula	β_0	β_1
H_{01}	1	0
H_{02}	0,98	0,02
H_{03}	0,96	0,04
H_{04}	0,94	0,06
H_{05}	0,92	0,08
H_{06}	0,9	0,1
H_{07}	0,88	0,12
H_{08}	0,86	0,14
H_{09}	0,84	0,16
H_{10}	0,82	0,18
H_{11}	0,8	0,2
H_{12}	0,78	0,22
H_{13}	0,76	0,24
H_{14}	0,74	0,26
H_{15}	0,72	0,28
H_{16}	0,7	0,3
H_{17}	0,68	0,32
H_{18}	0,66	0,34
H_{19}	0,64	0,36
H_{20}	0,62	0,38

Tabela A.4: Hipóteses testadas para parâmetros de dispersão.