LINEU ALBERTO CAVAZANI DE FREITAS

TESTE WALD PARA ESTUDO DE PARÂMETROS DE MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA

(*versão pré-defesa, compilada em 3 de novembro de 2021*)

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.

Coorientador: Prof. Dr. Marco Antônio Zanata Alves.

CURITIBA PR

2021

RESUMO

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

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

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

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

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

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

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

ABSTRACT

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

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

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

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

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

Keywords: Keyword 1. Keyword 2. Keyword 3.

LISTA DE FIGURAS

LISTA DE TABELAS

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

LISTA DE SÍMBOLOS

α	alfa, primeira letra do alfabeto grego
β	beta, segunda letra do alfabeto grego
γ	gama, terceira letra do alfabeto grego
ω	ômega, última letra do alfabeto grego
π	pi
τ	Tempo de resposta do sistema
θ	Ângulo de incidência do raio luminoso

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. Em Cao (2016) é dito que ciência de dados engloba técnicas de como: estatística, aprendizado de máquina, gerenciamento de *big data*, dentre outras.

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

Para entender minimamente um modelo de regressão, é necessário compreender o conceito de fenômeno aleatório, variável aleatória e distribuição de probabilidade. Um fenômeno aleatório é uma situação na qual diferentes observações podem fornecer diferentes desfechos. Estes fenômenos podem ser descritos por variáveis aleatórias que associam um valor numérico a cada desfecho possível do fenômeno. Os desfechos deste fenômeno podem ser descritos por uma escala que pode ser discreta ou contínua. Uma variável aleatória é considerada discreta quando os possíveis desfechos estão dentro de um conjunto enumerável de valores. Já uma variável aleatória contínua ocorre quando os possíveis resultados estão em um conjunto não enumerável de valores. Na prática existem probabilidades associadas aos valores de uma variável aleatória, e estas probabilidades podem ser descritas 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, esta função é chamada de função densidade de probabilidade.

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

que controlam aspectos da distribuição como escala e forma, 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 variávis explicativas é a média. Fazendo uso da equação resultante do processo de análise de regressão, é possível estudar a importância das variáveis explicativas sobre a resposta e realizar predições da variável resposta com base nos valores observados das variáveis explicativas.

Em contextos práticos o processo de análise via modelo de regressão parte de um conjunto de dados. Neste contexto, um conjunto de dados é uma representação tabular em que unidades amostrais são representadas nas linhas e seus atributos (variáveis) são representados nas colunas. Pode-se usar um modelo de regressão para, por exemplo, modelar a relação entre a média de uma variável aleatória e um conjunto de variáveis explicativas. Assume-se então que a variável aleatória segue uma distribuição de probabilidades e que o parâmetro de média desta distribuição pode ser descrito por uma combinação linear de parâmetros de regressão associados às variáveis explicativas. Sendo assim, o conhecimento a respeito da influência de uma variável explicativa sobre a resposta vem do estudo das estimativas dos parâmetros de regressão. A obtenção 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 as 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 o 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 com uma extensão multivariada dos GLMs que permite lidar com múltiplas respostas de diferentes naturezas e, de alguma forma, correlacionadas. Além disso, não há nesta classe suposições quanto à independência entre as observações, pois a correlação entre observações pode ser modelada por um preditor linear matricial que envolve matrizes conhecidas. Estas características tornam o McGLM uma classe flexível 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 estatísticasmais 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 estatísticamente significativa, significa que a retirada das variáveis do modelo completo prejudicam o ajuste. Caso não seja observada diferença entre o modelo completo e o restrito, significa que as variáveis retiradas não geram perda na qualidade e, por este motivo, tais variáveis podem ser descartadas.

Já o teste Wald, proposto por Wald (1943), requer apenas um modelo ajustado. A ideia consiste em verificar se existe evidência para afirmar que um ou mais parâmetros são iguais a valores postulados. O teste avalia quão longe o valor estimado está do valor postulado. Utilizando o teste Wald é possível formular hipóteses para múltiplos parâmetros, e costuma ser de especial interesse verificar se há evidência que permita afirmar que os parâmetros que associam determinada variável explicativa a variável resposta são iguais a zero. Caso tal hipótese não seja rejeitada, significa que 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, 1948), tal como o teste Wald, requer apenas um modelo ajustado. No caso do teste escore o modelo ajustado não possui o parâmetro de interesse e o que é feito é testar se adicionar esta variável omitida resultará em uma melhora significativa no modelo. Isto é feito com base na inclinação da função de verossimilhança, esta inclinação é usada para estimar a melhoria no modelo caso as variáveis omitidas fossem incluídas.

De certo modo, os três testes podem ser usados para verificar se a 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 assintóticamente equivalentes. Em amostras finitas estes testes podem apresentar resultados diferentes como discutido por Evans e Savin (1982).

Para o caso dos modelos lineares tradicionais existem técnicas como a análise de variância (ANOVA), proposta inicialmente por Fisher e Mackenzie (1923). Segundo St et al. (1989), a ANOVA é um dos métodos estatísticos mais amplamente usados para testar hipóteses e que está presente em praticamente todos os materiais introdutórios de estatística. O objetivo da técnica é a avaliação do efeito de cada uma das variáveis explicativas sobre a resposta. Isto é feito 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 forncece ferramentas para avaliar qualquer parâmetro de um McGLM.

Quando trabalhamos na classe dos McGLMs estimamos parâmetros de regressão, dispersão e potência. 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 repetivas. Já os parâmetros de potência nos fornecem um indicativo de qual distribuição de probabilidade melhor se adequa ao problema. 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 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? Existe efeito da estrutura de correlação entre indivíduos no estudo? Qual a distribuição de probabilidade que melhor se adequa ao problema? O efeito de determinada variável é o mesmo independente da resposta? Dentre outras.

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

1.6 ORGANIZAÇÃO DO DOCUMENTO

Esta dissertação está organizada em oito 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. As implementações computacionais do método proposto são apresentadas no Capítulo 5. O Capítulo 6 é dedicado aos resultados da avaliação de perfomance do teste proposto com base em um estudo de simulação. No capítulo 7 buscamos motivar o uso da proposta por meio da aplicação do método a problemas práticos e reais de análise de dados. Por fim, encerramos o trabalho com nossas considerações finais no Capítulo 8.

TODO

- **TALVEZ EXEMPLO DE PROBLEMA EM QUE SE APLIQUE UM MODELO DE REGRESSÃO MULTIVARIADO.**

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 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 à 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

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

em que $g(.)$ é a função de ligação, $\mathrm{V}\,(\boldsymbol{\mu}; p)$ 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 qualquer que reflita a relação entre os indivíduos da amostra teremos uma extensão do Modelo Linear Generalizado para observações dependentes. É justamente essa a ideia dos modelos de covariância linear generalizada, o cGLM, 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

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

em que $h(.)$ é a função de ligação de covariância, Z_d com $d = 0,\ldots, D$ são matrizes que representam a estrutura de covariância presente nos dados e $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_D)$ é um vetor $(D+1)\times 1$ de parâmetros de dispersão. Note que o número de matrizes usadas para especificar o preditor linear matricial, definido por D, 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, a especificação da função de ligação de covariância é discutida por Pinheiro e Bates (1996). É possível selecionar combinações de matrizes para se obter os mais conhecidos modelos da literatura para dados longitudinais, séries temporais, dados espaciais e espaço-temporais. 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, considre $\boldsymbol{Y}_{N\times R} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_R\}$ uma matriz de variáveis resposta e $\boldsymbol{M}_{N\times R} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_R\}$ uma matriz de valores esperados. Cada uma das variáveis resposta tem sua própria matriz de variância e covariância, responsável por modelar a covariância dentro de cada resposta, sendo expressa por

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

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

Finalmente, um McGLM é descrito como

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

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

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 superparametrizados, 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 $\boldsymbol{\mathcal{M}} = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_R^\top)^\top$ denota o vetor empilhado de ordem $NR \times 1$ da matriz de valores esperados $\boldsymbol{M}_{N\times R}$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2 TESTES DE HIPÓTESES

A palavra “inferir” significa tirar conclusão. O campo de estudo chamado de inferência estatística tem como objetivo o desenvolvimento e discussão de métodos e procedimentos que permitem, com certo grau de confiança, fazer afirmações sobre uma 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 assintóticamente 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:

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

em que $WT \sim \chi^2_s$. 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 da 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 numa 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 à 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 ao problema, contudo, cuidados devem ser tomados no que diz respeito à forma como o quadro foi elaborado. Como já mencionado, cada linha do quadro refere-se a uma hipótese e estas hipóteses podem ser formuladas de formas distintas. Formas conhecidas de se elaborar o quadro são as chamadas ANOVAs 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).

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 consiste basicamente de 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 nível de referência e, para os demais níveis, mede-se a mudança para a categoria de referência; este é o chamado constraste de tramento. 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 o contraste desejado na linha da matriz $\boldsymbol{K_1}$ e utilizar esta linha 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.

TODO

- **ESPECIFICAÇÃO DO MODELO, FUNÇÕES DE VARIÂNCIA E ALGUNS PONTOS DE ESTIMAÇÃO E INFERÊNCIA DIFEREM ENTRE OS ARTIGOS DO JRSS E JSS, QUAL USAR?**
- **O DANIEL ACHOU QUE A EXPLICAÇÃO SOBRE A SIGMA B NÃO ESTÁ CLARA. MAS NO ARTIGO DO MCGLM NO JRSS SO TEM ISSO "We introduce the R×R correlation matrix sigma b to model the correlation between response variables". O QUE ACRESCENTAR?**
- **MAIS DE UMA PESSOA DISSE QUE NÃO FICOU BOM MENCIONAR TRV E LMT SEM USARMOS, DEIXEI APENAS NA INTRODUÇÃO A EXPLICAÇÃO GENÉRICA DELES, REFERENCIEI NO REFERENCIAL E OPTEI POR TIRAR A EXPLICAÇÃO**
- **UMA IMAGEM COM OS ELEMENTOS DE UM TESTE DE HIPÓTESES CAIRIA BEM**
- **O TEXTO DE ANOVA PASSOU PELA QUALI SEM CRÍTICAS, MAS SINTO QUE ESTÁ FRACO**
- **UMA IMAGEM COM OS ELEMENTOS DE UM QUADRO DE ANOVA CAIRIA BEM**

3 TRABALHOS RELACIONADOS

TODO
IDEIA
No contexto da proposta dos McGLMs (contornar as principais restrições dos GLMs):

- Selecionar algumas classes de modelos univariados
- Selecionar algumas classes de modelos multivariados
- Referenciar e mencionar: ideia geral, restrições que contorna, limitações, como são testadas hipóteses

VER COM O WAGNER QUAIS ELE GOSTARIA DE INCLUIR
Univariados

- GLM (na chamada do capítulo)
- hGLM (Lee and Nelder, 1996)
- GEE (Liang and Zeger, 1989)
- GAMLSS (Rigny and Stasinopoulus)

Multivariados

- Multivariate generalized linear mixed models (Berridge and Crouchley, 2011)
- Multivariate dispersion models (Jorgensen and Lauritzen, 2000)
- Copula models (Joe, 2014)

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 e potência e que nossa proposta pode ser aplicada a qualquer parâmetro ou combinação de parâmetros. Além disso, 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 e 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.

4.1 HIPÓTESES E ESTATÍSTICA DE TESTE

4.1.1 Exemplo 1: hipótese para um único parâmetro

4.1.2 Exemplo 2: hipótese para múltiplos parâmetros

4.1.3 Exemplo 3: hipótese de igualdade de parâmetros

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

4.2 ANOVA E MANOVA VIA TESTE WALD

4.2.1 ANOVA e MANOVA tipo I

4.2.2 ANOVA e MANOVA tipo II

4.2.3 ANOVA e MANOVA tipo III

4.3 TESTE DE COMPARAÇÕES MÚLTIPLAS VIA TESTE WALD

5 IMPLEMENTAÇÃO COMPUTACIONAL

6 ESTUDO DE SIMULAÇÃO

ESTUDO DE SIMULAÇÃO, VALIDAÇÃO DA PROPOSTA, AVALIAÇÃO DE DESEMPENHO

7 ANÁLISE DE DADOS

8 CONSIDERAÇÕES FINAIS

8.1 CONCLUSÕES GERAIS

8.2 LIMITAÇÕES

8.3 TRABALHOS FUTUROS

REFERÊNCIAS

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Liang, K.-Y. e Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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