LINEU ALBERTO CAVAZANI DE FREITAS

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

(*versão pré-defesa, compilada em 19 de outubro 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

DINF	Departamento de Informática
PPGINF	Programa de Pós-Graduação em Informática
UFPR	Universidade Federal do Paraná

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, tais parâmetros são quantidades desconhecidas estimadas por 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 destes parâmetros estimados se dá na chamada etapa de ajuste do modelo, e isto gera a equação da regressão ajustada.

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

Com o passar o tempo, o avanço computacional permitiu a proposição de modelos mais complexos, que necessitavam de processos iterativos para estimação dos parâmetros (Paula,

2004). A proposta de maior renome foram os modelos lineares generalizados (GLM) propostos por Nelder e Wedderburn (1972). Essa classe de modelos permitiu a flexibilização da distribuição da variável resposta de tal modo que esta pertença à família exponencial de distribuições. Em meio aos casos especiais de distribuições possíveis nesta classe de modelos estão a Bernoulli, binomial, Poisson, normal, gama, normal inversa, entre outras. Trata-se portanto, de uma classe de modelos de regressão univariados para dados de diferentes naturezas, tais como: dados contínuos simétricos e assimétricos, contagens e assim por diante. Tais características tornam esta classe uma flexível ferramenta de modelagem aplicável a diversos tipos de problema.

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

Quando trabalha-se com modelos de regressão, um interesse comum aos analistas é o de verificar se a retirada de determinada variável explicativa do modelo geraria uma perda no ajuste. Ou seja, uma conjectura de interesse é avaliar se há evidência suficiente nos dados para afirmar que determinada variável explicativa não possui efeito sobre a resposta. Isto é feito por dos chamados testes de hipóteses. Testes de hipóteses são ferramentas estatísticas que auxiliam no processo de tomada de decisão sobre valores desconhecidos (parâmetros) estimados por meio de uma amostra (estimativas). Tal procedimento permite verificar se existe evidência nos dados amostrais que apoiem ou não uma hipótese estatística formulada a respeito de um parâmetro. As suposições a respeito de um parâmetro desconhecido estimado com base nos dados são denominadas hipóteses estatísticas, estas hipóteses podem ser rejeitadas ou não rejeitadas com base nos dados. Segundo Lehmann (1993) podemos atribuir a teoria, formalização e filosofia dos testes de hipótese a Neyman e Pearson (1928a), Neyman e Pearson (1928b) e Fisher (1925). A teoria clássica de testes de hipóteses é apresentada formalmente em Lehmann e Romano (2006).

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

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

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

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

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

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

ACRESCENTAR PARÁGRAFO SOBRE TESTES DE COMPARAÇÕES MÚLTIPLAS
SEGUIR NA LINHA DE COMPLEMENTO À ANOVA E MANOVA

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 lineares funcionaria para os McGLMs. A construção do teste Wald em modelos tradicionais é baseada nas estimativas de máxima verossimilhança. Contudo a estatística de teste usada não depende da máxima verossimilhança, e sim de um vetor de estimativas dos parâmetros e uma matriz de variância e covariância destas estimativas. Assim, por mais que os McGLMs não sejam ajustados com base na maximização da função de verossimilhança para obtenção dos parâmetros do modelo, o método de estimação fornece os componentes necessários para a construção do teste. Neste sentido, das três opções clássicas de testes de hipóteses comumente aplicados a problemas de regressão (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 variável explicativa à variável resposta, por meio do estudo destes parâmetros é possível avaliar o efeito da variável explicativa sobre a resposta. Por meio do estudo dos parâmetros de dispersão pode-se avaliar o efeito da correlação entre unidades do estudo. E os parâmetros de potência nos fornecem um indicativo de qual distribuição de probabilidade melhor se adequa ao problema. O desenvolvimento de testes de hipóteses para fins de avaliação destas quantidades é de grande valia em problemas práticos.

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

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

1.6 ORGANIZAÇÃO DO DOCUMENTO

Esta dissertação está organizada em oito capítulos:

Capítulo 1: 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 modelos multivariados de covariância linear generalizada.

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.

Capítulo 3: referenciamos trabalhos correlatos.

Capítulo 4: é apresentada nossa proposta com os detalhes do teste Wald para avaliar suposições sobre parâmetros de um McGLM.

Capítulo 5: são apresentadas as descrições das funções implementadas no trabalho.

Capítulo 6: é dedicado aos resultados da avaliação de perfomance do teste proposto com base em um estudo de simulação.

Capítulo 7: no capítulo 7 aplicamos o método proposto a problemas práticos de análise de dados.

Capítulo 8: encerramos o trabalho com nossas considerações finais.

2 REFERENCIAL TEÓRICO

REFERENCIAL TEÓRICO, FUNDAMENTAÇÃO TEÓRICA, REVISÃO DE LITERATURA

2.1 MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA

2.1.1 Modelo linear generalizado

2.1.2 Modelo de covariância linear generalizada

2.1.3 Modelos multivariados de covariância linear generalizada

2.1.4 Estimação e inferência

2.2 TESTES DE HIPÓTESES

2.2.1 Elementos de um teste de hipóteses

2.2.2 Testes de hipóteses em modelos de regressão

2.2.3 ANOVA e MANOVA

2.3 TESTES DE COMPARAÇÕES MÚLTIPLAS

3 TRABALHOS RELACIONADOS

4 TESTE WALD EM MODELOS MULTIVARIADOS DE COVARIÂNCIA LINEAR GENERALIZADA

PROPOSTA, METODOLOGIA

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.

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

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

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

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

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

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

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

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.

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.

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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