

Testes de hipóteses em Modelos Multivariados de Covariância Linear Generalizada

Qualificação

Lineu Alberto Cavazani de Freitas

Orientador: Prof. Dr. Wagner Hugo Bonat

Co-orientador: Prof. Dr. Marco Antonio Zanata Alves

PPG Informática
Data Science & Big Data
Universidade Federal do Paraná

[https://lineu96.github.io/st/
lineuacf@gmail.com](https://lineu96.github.io/st/lineuacf@gmail.com)

Conteúdo

1 Introdução

2 Revisão de literatura

- McGLM
- Testes de hipóteses

3 Proposta

- Adaptação do teste Wald para os McGLM
- Exemplos de hipóteses

1

Introdução

UNIVERSIDADE DO PARANÁ

Ciência de dados

- ▶ 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; BORDAS, 2018).
- ▶ Tem diversos campos de interesse.
- ▶ Os métodos estatísticos são de fundamental importância em grande parte das etapas da ciência de dados (WEIHS; ICKSTADT, 2018).
- ▶ Neste sentido, os modelos de regressão tem papel importante.

Modelos de regressão

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 situação na qual diferentes observações podem fornecer diferentes desfechos.
- ▶ **Variáveis aleatórias** associam um valor numérico a cada desfecho possível do fenômeno. Podem ser discretas ou contínuas.
- ▶ Existem probabilidades associadas aos valores de uma variável aleatória. Estas probabilidades podem ser descritas por funções:
 - ▶ Função de probabilidade, para variáveis aleatórias discretas.
 - ▶ Função densidade de probabilidade, para variáveis aleatórias contínuas.

Modelos de regressão

- ▶ Modelos probabilísticos que buscam descrever as probabilidades de variáveis aleatórias, as chamadas **distribuições de probabilidade**.
- ▶ Em problemas práticos, podemos buscar uma distribuição de probabilidades que melhor descreva o fenômeno de interesse.
- ▶ Estas distribuições são descritas por funções.
- ▶ Estas funções possuem parâmetros que controlam aspectos da distribuição.
- ▶ Os parâmetros são quantidades desconhecidas estimadas através dos dados.

Modelos de regressão

- ▶ Na análise de regressão busca-se modelar os parâmetros das distribuições de probabilidade como uma função de outras variáveis.
- ▶ Isto é feito através da decomposição do parâmetro da distribuição em outros parâmetros, chamados de parâmetros de regressão.
- ▶ Assim, o objetivo dos modelos de regressão consiste em obter uma equação que explique a relação entre as variáveis explicativas e o parâmetro de interesse da distribuição de probabilidades selecionada para modelar a variável aleatória.
- ▶ Em geral, o parâmetro de interesse da distribuição de probabilidades modelado em função das variáveis explicativas é a média.

Modelos de regressão

- ▶ O processo de análise via modelo de regressão parte de um conjunto de dados.
- ▶ Pode-se usar um modelo para modelar a relação entre a média de uma variável aleatória e um conjunto de variáveis explicativas.
- ▶ Assume-se 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.
- ▶ A obtenção destes parâmetros estimados se dá na chamada etapa de ajuste do modelo.
- ▶ Fazendo uso da equação resultante do processo é possível estudar a importância das variáveis explicativas sobre a resposta e realizar previsões da variável resposta com base nos valores observados das variáveis explicativas.

Modelos de regressão

- ▶ Existem 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:
 - ▶ Modelos lineares.
 - ▶ Modelos lineares generalizados.
 - ▶ Modelos multivariados de covariância linear generalizada.

Modelo linear normal

- ▶ No cenário univariado, durante muitos anos o modelo linear normal (GALTON, 1886) teve papel de destaque.
- ▶ Muito usado principalmente por suas facilidades computacionais.
- ▶ Um dos pressupostos do modelo linear normal é de que a variável resposta, condicional às variáveis explicativas, segue a distribuição normal.
- ▶ Quando tal pressuposto não era atendido, uma alternativa, por muito tempo adotada, foi buscar uma transformação da variável resposta, tal como a família de transformações Box-Cox (BOX; COX, 1964).

Modelos lineares generalizados

- ▶ 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) (NELDER; 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.

Modelos multivariados de covariância linear generalizada

- ▶ 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) (BONAT; JØRGENSEN, 2016).
- ▶ Esta classe pode ser vista com uma extensão multivariada dos GLMs que permite lidar com múltiplas respostas de diferentes naturezas e, de alguma forma, correlacionadas.
- ▶ O McGLM é 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.

Testes de hipóteses

- ▶ Em regressão, um interesse comum é o de verificar se a retirada de determinada variável explicativa do modelo geraria uma perda no ajuste.
- ▶ Isto é feito através dos chamados testes de hipóteses.
- ▶ Testes de hipóteses são ferramentas estatísticas que auxiliam no processo de tomada de decisão sobre valores desconhecidos (parâmetros) estimados por meio de uma amostra (estimativas).
- ▶ Podemos atribuir a teoria, formalização e filosofia dos testes de hipótese a Neyman, Pearson e Fisher.

Testes de hipóteses

No contexto de modelos de regressão, três testes de hipóteses são comuns, todos baseados na função de verossimilhança:

- ▶ O teste da razão de verossimilhanças (WILKS, 1938).
- ▶ O teste Wald (WALD, 1943).
- ▶ O teste do multiplicador de lagrange, também conhecido como teste escore (AITCHISON; SILVEY, 1958), (SILVEY, 1959), (RAO, 1948).

Técnicas baseadas em testes de hipóteses

- ▶ Existem técnicas como a análise de variância (ANOVA) (FISHER; MACKENZIE, 1923).
- ▶ O objetivo da técnica é a avaliação do efeito de cada uma das variáveis explicativas sobre a resposta.
- ▶ Isto é feito através da comparação via testes de hipóteses entre modelos com e sem cada uma das variáveis explicativas.
- ▶ 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 para a análise de variância multivariada (SMITH; GNANADESIKAN; HUGHES, 1962), a MANOVA.

Proposta

Objetivos gerais:

- ▶ Considerando os McGLMs, não há discussão a respeito da construção de testes de hipóteses.
- ▶ Nosso objetivo geral é o desenvolvimento de testes de hipóteses para os McGLMs.
- ▶ Buscamos propor uma adaptação do teste de Wald clássico utilizado em modelos lineares para os McGLMs.

Proposta

Objetivos específicos:

- ▶ Adaptar o teste Wald para realização de testes de hipóteses gerais sobre parâmetros de McGLMs.
- ▶ Implementar funções para efetuar tais testes, bem como funções para efetuar ANOVAs e MANOVAs para os McGLMs.
- ▶ Avaliar as propriedades e comportamento dos testes propostos com base em estudos de simulação.
- ▶ Avaliar o potencial de aplicação das metodologias discutidas com base na aplicação a conjuntos de dados reais.

2

Revisão de literatura

Revisão de literatura

A revisão de literatura compreende 2 temas:

- ▶ Modelos multivariados de covariância linear generalizada.
- ▶ Testes de hipóteses.

Modelos Multivariados de Covariância Linear Generalizada

Modelos multivariados de covariância linear generalizada

- ▶ Os GLM são uma forma de modelagem para lidar com apenas uma resposta para dados de diferentes naturezas.
- ▶ É uma classe de modelos flexível e aplicável a diversos tipos de problema.
- ▶ Apresenta três importantes restrições:
 - ▶ A incapacidade de lidar com observações dependentes.
 - ▶ A incapacidade de lidar com múltiplas respostas simultaneamente.
 - ▶ Leque reduzido de distribuições disponíveis.
- ▶ Com o objetivo de contornar estas restrições, foram propostos os chamados Modelos Multivariados de Covariância Linear Generalizada (McGLM).
- ▶ Vamos discutir os McGLM como uma extensão dos GLM, seguindo as ideias de (BONAT; JØRGENSEN, 2016) .

GLM

Considere:

- ▶ Y um vetor $N \times 1$ de valores observados da variável resposta.
- ▶ X uma matriz de delineamento $N \times k$
- ▶ β um vetor de parâmetros de regressão $k \times 1$.

GLM

um GLM pode ser descrito da forma

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

Em que:

- ▶ $g(\cdot)$ é a função de ligação.
- ▶ $V(\mu; p)$ é uma matriz diagonal em que as entradas principais são dadas pela função de variância aplicada ao vetor μ .
- ▶ p é o parâmetro de potência.
- ▶ τ_0 o parâmetro de dispersão.
- ▶ I é a matriz identidade de ordem $N \times N$.

GLM

1. Função de variância potência (JØRGENSEN, 1987) e (JØRGENSEN, 1997).
 - ▶ caracteriza a família Tweedie de distribuições.
 - ▶ função de variância é dada por $\vartheta(\mu; p) = \mu^p$
 - ▶ casos particulares: normal ($p = 0$), Poisson ($p = 1$), gama ($p = 2$) e normal inversa ($p = 3$).
2. Função de dispersão Poisson–Tweedie (JØRGENSEN; KOKONENDJI, 2015).
 - ▶ caracteriza a família Poisson-Tweedie de distribuições
 - ▶ visa contornar a inflexibilidade da utilização da função de variância potência para respostas discretas.
 - ▶ função de dispersão dada por $\vartheta(\mu; p) = \mu + \mu^p$
 - ▶ 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 ($p = 3$)
3. Função de variância binomial.
 - ▶ dada por $\vartheta(\mu) = \mu(1 - \mu)$
 - ▶ utilizada quando a variável resposta é binária, restrita a um intervalo ou quando tem-se o número de sucessos em um número de tentativas.

cGLM

- ▶ Alternativa para problemas em que a suposição de independência entre as observações não é atendida.
- ▶ A solução proposta é substituir a matriz identidade I da equação que descreve a matriz de variância e covariância por uma matriz não diagonal $\Omega(\tau)$.
- ▶ A matriz $\Omega(\tau)$ é descrita como uma combinação de matrizes conhecidas (ANDERSON et al., 1973) (POURAHMADI, 2000).

cGLM

A matriz $\mathbf{\Omega}(\boldsymbol{\tau})$ pode ser escrita como:

$$h\{\mathbf{\Omega}(\boldsymbol{\tau})\} = \tau_0 \mathbf{Z}_0 + \dots + \tau_D \mathbf{Z}_D, \quad (2)$$

em que

- ▶ $h(\cdot)$ é a função de ligação de covariância.
- ▶ \mathbf{Z}_d com $d = 0, \dots, D$ são matrizes que representam a estrutura de covariância presente nos dados.
- ▶ $\boldsymbol{\tau} = (\tau_0, \dots, \tau_D)$ é um vetor $(D + 1) \times 1$ de parâmetros de dispersão.
- ▶ Tal estrutura pode ser vista como um análogo ao preditor linear para a média e foi nomeado como preditor linear matricial.

McGLM

- ▶ Pode ser entendido como uma extensão multivariada do cGLM.
- ▶ Contorna as principais restrições presentes nos GLM.

Considere

- ▶ $\mathbf{Y}_{N \times R} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_R\}$ uma matriz de variáveis resposta
- ▶ $\mathbf{M}_{N \times R} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_R\}$ uma matriz de valores esperados.
- ▶ $\boldsymbol{\Sigma}_b$, uma matriz de ordem $R \times R$, que descreve a correlação entre as variáveis resposta

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

$$\boldsymbol{\Sigma}_r = \mathbf{V}_r(\boldsymbol{\mu}_r; \mathbf{p})^{1/2} \boldsymbol{\Omega}_r(\boldsymbol{\tau}) \mathbf{V}_r(\boldsymbol{\mu}_r; \mathbf{p})^{1/2}. \quad (3)$$

McGLM

Um MCGLM é descrito como

$$\begin{aligned} E(\mathbf{Y}) &= \mathbf{M} = \{g_1^{-1}(\mathbf{X}_1\boldsymbol{\beta}_1), \dots, g_R^{-1}(\mathbf{X}_R\boldsymbol{\beta}_R)\} \\ \text{Var}(\mathbf{Y}) &= \mathbf{C} = \boldsymbol{\Sigma}_R \overset{\text{G}}{\otimes} \boldsymbol{\Sigma}_b, \end{aligned} \quad (4)$$

em que

- ▶ $\boldsymbol{\Sigma}_R \overset{\text{G}}{\otimes} \boldsymbol{\Sigma}_b = \text{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \dots, \tilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b \otimes \mathbf{I})\text{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^\top, \dots, \tilde{\boldsymbol{\Sigma}}_R^\top)$ 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.
- ▶ \mathbf{I} uma matriz identidade $N \times N$.
- ▶ Toda metodologia do McGLM está implementada no pacote *mcglm* (BONAT, 2018) do software estatístico R.

Estimação e inferência

UNIVERSIDADE DO PARANÁ

Funções de estimação

As funções de estimação para os parâmetros de regressão (função quasi-score) e de dispersão (função de estimação de Pearson) são dadas por:

$$\begin{aligned}\psi_{\beta}(\boldsymbol{\beta}, \boldsymbol{\lambda}) &= \mathbf{D}^{\top} \mathbf{C}^{-1}(\mathbf{y} - \mathcal{M}) \\ \psi_{\lambda_i}(\boldsymbol{\beta}, \boldsymbol{\lambda}) &= \text{tr}(\mathbf{W}_{\lambda_i}(\mathbf{r}^{\top} \mathbf{r} - \mathbf{C})), i = 1, \dots, Q\end{aligned}$$

Em que:

- ▶ $\boldsymbol{\beta}_r$ denota um vetor $k_r \times 1$ de parâmetros de regressão.
- ▶ $\boldsymbol{\lambda}$ é um vetor $Q \times 1$ de parâmetros de dispersão.
- ▶ \mathbf{y} é um vetor $NR \times 1$ com os valores da matriz de variáveis respostas $\mathbf{Y}_{N \times R}$ empilhados.
- ▶ \mathcal{M} é um vetor $NR \times 1$ com os valores da matriz de valores esperados $\mathbf{M}_{N \times R}$ empilhados.
- ▶ $\mathbf{D} = \nabla_{\boldsymbol{\beta}} \mathcal{M}$ é uma matriz $NR \times K$, e $\nabla_{\boldsymbol{\beta}}$ denota o operador gradiente.
- ▶ $\mathbf{W}_{\lambda_i} = -\frac{\partial \mathbf{C}^{-1}}{\partial \lambda_i}$
- ▶ $\mathbf{r} = (\mathbf{y} - \mathcal{M})$

Distribuição assintótica e algoritmo de estimação

- ▶ Para resolver o sistema de equações $\psi_{\beta} = 0$ e $\psi_{\lambda} = 0$ faz-se uso do algoritmo Chaser modificado:

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

- ▶ Seja $\hat{\theta} = (\hat{\beta}^{\top}, \hat{\lambda}^{\top})^{\top}$ o estimador baseado em funções de estimação de θ .
- ▶ A distribuição assintótica de $\hat{\theta}$ é:

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

J_{θ}^{-1} é a inversa da matriz de informação de Godambe, dada por

$$J_{\theta}^{-1} = S_{\theta}^{-1} V_{\theta} S_{\theta}^{-\top},$$

em que $S_{\theta}^{-\top} = (S_{\theta}^{-1})^{\top}$.

Testes de hipóteses

UNIVERSIDADE DO PARANÁ

Testes de hipóteses

- ▶ Um dos objetivos principais da análise estatística é inferir conclusões válidas a respeito de uma população por meio do estudo de uma amostra.
- ▶ Problemas de inferência estatística são:
 1. Estimação de parâmetros com base em informação amostral.
 2. Testes de hipóteses:
 - ▶ Com base na evidência amostral, podemos considerar que dado parâmetro tem determinado valor?
- ▶ Uma hipótese estatística é uma afirmação ou conjectura sobre parâmetros.

Testes de hipóteses

- ▶ São postuladas 2 hipóteses, chamadas de nula e alternativa.
- ▶ Avalia-se uma estatística de teste.
- ▶ Com base no valor da estatística e de acordo com sua distribuição de probabilidade, toma-se a decisão de rejeitar ou não rejeitar a hipótese nula.

Notação:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

Testes de hipóteses

Desfechos possíveis:

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

Tabela 1. Desfechos possíveis em um teste de hipóteses

- ▶ A probabilidade do erro do tipo I recebe o nome de nível de significância.
- ▶ 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.
- ▶ P-valor: probabilidade de a estatística de teste tomar um valor igual ou mais extremo do que aquele que foi observado. Quanto menor for o P, maior será o grau com que os dados amostrais contrariam a hipótese nula.

Testes de hipóteses

- ▶ Em modelos de regressão, testes de hipóteses são usados para verificar se a retirada de determinada variável explicativa do modelo geraria uma perda no ajuste.
- ▶ Os três testes mais usados para este fim são:
 - ▶ O teste da razão de verossimilhanças (WILKS, 1938).
 - ▶ O teste Wald (WALD, 1943).
 - ▶ O teste do multiplicador de lagrange, também conhecido como teste escore (AITCHISON; SILVEY, 1958), (SILVEY, 1959), (RAO, 1948).
- ▶ Todos eles são baseados na função de verossimilhança dos modelos.
- ▶ São assintoticamente equivalentes. Em amostras finitas estes testes podem apresentar resultados diferentes (EVANS; SAVIN, 1982).

Teste da razão de verossimilhanças

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

Teste da razão de verossimilhanças

- ▶ Considere 2 modelos de regressão encaixados em que a diferença entre o número de parâmetros entre os modelos é igual a q .
- ▶ Considere β um vetor de parâmetros de regressão e $\mathbf{0}$ o vetor nulo.
- ▶ As hipóteses podem ser descritas como

$$\begin{cases} H_0 : \beta = \mathbf{0} \\ H_1 : \beta \neq \mathbf{0} \end{cases}$$

- ▶ A estatística de teste é dada por:

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

em que L_0 representa a verossimilhança do modelo restrito e L_1 a verossimilhança do irrestrito.

- ▶ $\text{LRT} \sim \chi^2_q$.

Teste Wald

- ▶ O teste Wald, 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.

Teste Wald

- ▶ Considere um modelo de regressão ajustado com p parâmetros.
- ▶ Considere β um vetor de parâmetros de regressão, em que as estimativas são dadas por $\hat{\beta}$ e c um vetor de valores postulados de dimensão q .
- ▶ Considere L uma matriz de especificação das hipóteses, de dimensão $q \times p$.
- ▶ As hipóteses podem ser descritas como

$$\begin{cases} H_0 : L\beta = c \\ H_1 : L\beta \neq c \end{cases}$$

- ▶ A estatística de teste é dada por:

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

- ▶ $WT \sim \chi^2_q$.

Teste Escore

- ▶ O teste do multiplicador de lagrange ou teste score, 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.

Teste Escore

- ▶ Considere um modelo de regressão ajustado com p parâmetros.
- ▶ Considere β um vetor de parâmetros de regressão, em que as estimativas são dadas por $\hat{\beta}$ e $\mathbf{0}$ o vetor nulo.
- ▶ As hipóteses podem ser descritas como

$$\begin{cases} H_0 : \beta = \mathbf{0} \\ H_1 : \beta \neq \mathbf{0} \end{cases}$$

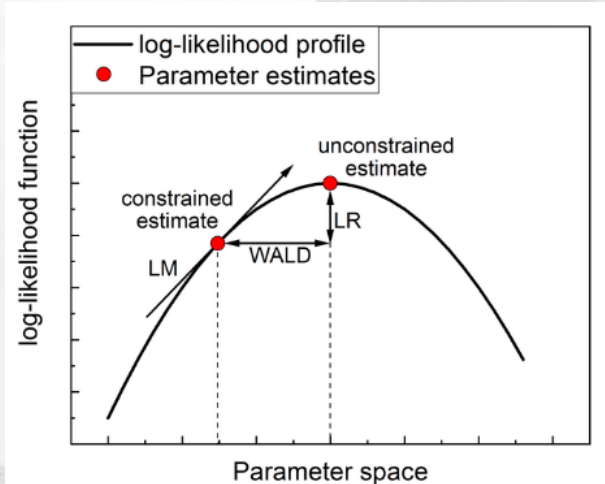
- ▶ A estatística de teste é dada por:

$$LMT = \mathbf{S}'(\hat{\beta}) \mathbf{Var}(\hat{\beta}) \mathbf{S}(\hat{\beta}).$$

Em que $\mathbf{S}(\hat{\beta})$ representa a função escore e $\mathbf{Var}(\hat{\beta})$ a matriz de variâncias avaliadas sob o modelo restrito (H_0).

- ▶ $WT \sim \chi^2_q$, em que q representa o número de parâmetros fixados sob H_0 .

LRT, WT & LMT



3

Proposta

UNIVERSIDADE DO PARANÁ

Adaptação do teste Wald para os McGLMs

Hipóteses

As hipóteses a serem testadas podem ser escritas como:

$$H_0 : \mathbf{L}\theta_{\beta,\tau,p} = \mathbf{c} \text{ vs } H_1 : \mathbf{L}\theta_{\beta,\tau,p} \neq \mathbf{c}.$$

Em que:

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

Estatística de teste

A generalização da estatística de teste para verificar a validade de uma hipótese sobre parâmetros de um McGLM é dada por:

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

Em que:

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

Exemplos de hipóteses nos McGLMs

Exemplos de hipóteses

Considere um modelo bivariado genérico, com preditor dado por:

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

- ▶ O índice r denota a variável resposta, $r = 1, 2$.
- ▶ β_{r0} representa o intercepto.
- ▶ β_{r1} um parâmetro de regressão associado a uma variável x_1 .
- ▶ Considere que cada resposta possui apenas um parâmetro de dispersão: τ_{r1} .
- ▶ Considere que os parâmetros de potência foram fixados.

Exemplo 1: único parâmetro

Considere a hipótese:

$$H_0 : \beta_{11} = 0 \text{ vs } H_1 : \beta_{11} \neq 0.$$

Esta hipótese pode ser reescrita na seguinte notação:

$$H_0 : \mathbf{L}\boldsymbol{\theta}_{\beta,\tau,p} = \mathbf{c} \text{ vs } H_1 : \mathbf{L}\boldsymbol{\theta}_{\beta,\tau,p} \neq \mathbf{c}.$$

Em que:

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

Exemplo 2: múltiplos parâmetros

Considere a hipótese:

$$H_0 : \beta_{r1} = 0 \text{ vs } H_1 : \beta_{r1} \neq 0.$$

Ou, da mesma forma:

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

Exemplo 2: múltiplos parâmetros

A hipótese pode ser reescrita na seguinte notação:

$$H_0 : \mathbf{L}\boldsymbol{\theta}_{\beta,\tau,p} = \mathbf{c} \text{ vs } H_1 : \mathbf{L}\boldsymbol{\theta}_{\beta,\tau,p} \neq \mathbf{c}.$$

Em que:

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

Exemplo 3: igualdade de efeitos

Considere a hipótese:

$$H_0 : \beta_{11} - \beta_{21} = 0 \text{ vs } H_1 : \beta_{11} - \beta_{21} \neq 0.$$

Esta hipótese pode ser reescrita na seguinte notação:

$$H_0 : \mathbf{L}\boldsymbol{\theta}_{\beta,\tau,p} = \mathbf{c} \text{ vs } H_1 : \mathbf{L}\boldsymbol{\theta}_{\beta,\tau,p} \neq \mathbf{c}.$$

Em que:

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

Obrigado!

Lineu Alberto Cavazani de Freitas


lineuacf@gmail.com


<https://lineu96.github.io/st/>


PPG Informática





Referências bibliográficas


 AITCHISON, J.; SILVEY, S. Maximum-likelihood estimation of parameters subject to restraints. **The annals of mathematical Statistics**, JSTOR, p. 813–828, 1958.


 ANDERSON, T. et al. Asymptotically efficient estimation of covariance matrices with linear structure. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 1, n. 1, p. 135–141, 1973.

 BONAT, W. H. Multiple response variables regression models in R: The mcglm package. **Journal of Statistical Software**, v. 84, n. 4, p. 1–30, 2018.

 BONAT, W. H.; JØRGENSEN, B. Multivariate covariance generalized linear models. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 65, n. 5, p. 649–675, 2016.

 BOX, G. E.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 211–252, 1964.

 EVANS, G.; SAVIN, N. E. Conflict among the criteria revisited; the w, lr and lm tests. **Econometrica: Journal of the Econometric Society**, JSTOR, p. 737–748, 1982.

 FISHER, R. A.; MACKENZIE, W. A. Studies in crop variation. ii. the manurial response of different potato varieties. **The Journal of Agricultural Science**, Cambridge University Press, v. 13, n. 3, p. 311–320, 1923.

 GALTON, F. Regression towards mediocrity in hereditary stature. **The Journal of the Anthropological Institute of Great Britain and Ireland**, JSTOR, v. 15, p. 246–263, 1886.