

# 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



# Conteúdo

- 1 Introdução
- 2 Revisão de literatura
  - McGLM
  - Testes de hipóteses
  - ANOVA & MANOVA
- 3 Proposta
  - Adaptação do teste Wald para os McGLM
  - Exemplos de hipóteses
  - ANOVA & MANOVA via teste Wald
- 4 Resultados preliminares
- 5 Considerações finais
- 6 Pendências

Introdução

Revisão de literatura

Proposta

Resultados preliminares

Considerações finais

Pendências

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.

Introdução

Revisão de literatura

Proposta

Resultados preliminares

Considerações finais

Pendências

# 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.

Introdução

**Revisão de literatura**

McGLM

Testes de hipóteses

ANOVA & MANOVA

Proposta

Resultados preliminares

Considerações finais

Pendências

## 2

# Revisão de literatura

# Revisão de literatura

A revisão de literatura compreende 3 temas:

1. Modelos multivariados de covariância linear generalizada.
2. Testes de hipóteses.
3. ANOVA & MANOVA.

# Modelos Multivariados de Covariância Linear Generalizada

Introdução

Revisão de literatura

McGLM

Testes de hipóteses

ANOVA & MANOVA

Proposta

Resultados preliminares

Considerações finais

Pendências

# Modelos multivariados de covariância linear generalizada

- ▶ Os GLMs 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:
  1. A incapacidade de lidar com observações dependentes.
  2. A incapacidade de lidar com múltiplas respostas simultaneamente.
  3. Leque reduzido de distribuições disponíveis.
- ▶ Os McGLMs contornam estas restrições.
- ▶ Vamos discutir os McGLM como uma extensão dos GLM, seguindo as ideias de (BONAT; JØRGENSEN, 2016) .

# GLM

TH MCGLM

Lineu Alberto

Introdução

Revisão de literatura

McGLM

Testes de hipóteses

ANOVA & MANOVA

Proposta

Resultados preliminares

Considerações finais

Pendências

Considere:

- ▶  $Y$  um vetor  $N \times 1$  de valores observados da variável resposta.
- ▶  $X$  uma matriz de delineamento  $N \times k$ .
- ▶  $\beta$  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  $\mu$ .
- ▶  $p$  é o parâmetro de potência.
- ▶  $\tau_0$  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).
  - ▶ Família Tweedie de distribuições.
  - ▶  $\vartheta(\mu; p) = \mu^p$ .
  - ▶ 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).
  - ▶ Família Poisson-Tweedie de distribuições.
  - ▶  $\vartheta(\mu; p) = \mu + \mu^p$ .
  - ▶ Casos particulares: Hermite ( $p = 0$ ), Neyman tipo A ( $p = 1$ ), binomial negativa ( $p = 2$ ) e Poisson–inversa gaussiana ( $p = 3$ ).
3. Função de variância binomial.
  - ▶  $\vartheta(\mu) = \mu(1 - \mu)$ .
  - ▶ Acomoda respostas binárias ou restritas a um intervalo.



# 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:

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

# McGLM

Um McGLM é descrito como:

$$\begin{aligned} E(Y) &= M = \{g_1^{-1}(X_1\beta_1), \dots, g_R^{-1}(X_R\beta_R)\} \\ \text{Var}(Y) &= C = \Sigma_R \overset{G}{\otimes} \Sigma_b, \end{aligned} \tag{4}$$

em que

- ▶  $\Sigma_R \overset{G}{\otimes} \Sigma_b = B \text{diag}(\tilde{\Sigma}_1, \dots, \tilde{\Sigma}_R)(\Sigma_b \otimes I) B \text{diag}(\tilde{\Sigma}_1^\top, \dots, \tilde{\Sigma}_R^\top)$  é o produto generalizado de Kronecker.
- ▶ A matriz  $\tilde{\Sigma}_r$  denota a matriz triangular inferior da decomposição de Cholesky da matriz  $\Sigma_r$ .
- ▶ O operador Bdiag denota a matriz bloco-diagonal.
- ▶ I uma matriz identidade  $N \times N$ .
- ▶ A metodologia do McGLM está implementada no pacote *mcglm* (BONAT, 2018) do software R.

# Estimação e inferência

Introdução

Revisão de literatura

McGLM

Testes de hipóteses

ANOVA & MANOVA

Proposta

Resultados preliminares

Considerações finais

Pendências

# 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.
- ▶  $\mathbf{W}_{\lambda_i} = -\frac{\partial \mathbf{C}^{-1}}{\partial \lambda_i}$
- ▶  $\mathbf{r} = (\mathbf{y} - \mathcal{M})$

[Introdução](#)[Revisão de literatura](#)[McGLM](#)[Testes de hipóteses](#)[ANOVA & MANOVA](#)[Proposta](#)[Resultados preliminares](#)[Considerações finais](#)[Pendências](#)

# 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  $\theta$ .
- ▶ A distribuição assintótica de  $\hat{\theta}$  é:

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

$J_{\theta}^{-1}$  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

Introdução

Revisão de literatura

McGLM

**Testes de hipóteses**

ANOVA & MANOVA

Proposta

Resultados preliminares

Considerações finais

Pendências



# 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.
- ▶ Seja  $\theta$  um parâmetro, um teste de hipóteses sobre  $\theta$  é dado por:

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

# Testes de hipóteses

Desfechos possíveis:

|                                    | <b>Rejeita <math>H_0</math></b> | <b>Não Rejeita <math>H_0</math></b> |
|------------------------------------|---------------------------------|-------------------------------------|
| <b><math>H_0</math> verdadeira</b> | Erro tipo I                     | Decisão correta                     |
| <b><math>H_0</math> falsa</b>      | 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.
- ▶ A probabilidade de a estatística de teste tomar um valor igual ou mais extremo do que aquele que foi observado recebe o nome de p-valor.

# 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  $\beta$  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.

# Teste Wald

- ▶ Considere um modelo de regressão ajustado com  $p$  parâmetros.
- ▶ Considere  $\beta$  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 melhora no modelo caso as variáveis omitidas fossem incluídas.

- ▶ Considere um modelo de regressão ajustado com  $p$  parâmetros.
- ▶ Considere  $\beta$  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

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

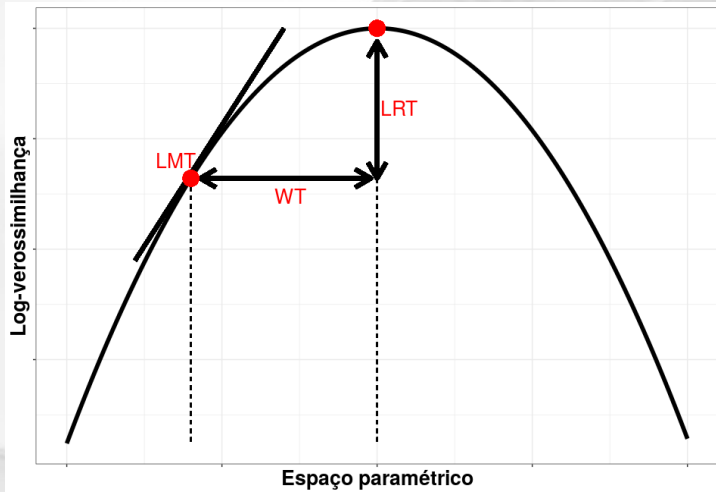
- ▶ A estatística de teste é dada por:

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

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

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

# LRT, WT & LMT



# ANOVA & MANOVA

Introdução

Revisão de literatura

McGLM

Testes de hipóteses

**ANOVA & MANOVA**

Proposta

Resultados preliminares

Considerações finais

Pendências

# ANOVA & MANOVA

- ▶ Formas de avaliar a significância de cada uma das variáveis de uma forma procedural.
- ▶ Consiste em efetuar testes sucessivos impondo restrições ao modelo original.
- ▶ O objetivo é testar se a ausência de determinada variável gera perda ao modelo.
- ▶ Os resultados são sumarizados numa tabela, o chamado quadro de análise de variância, que contém em cada linha:
  1. A variável.
  2. O valor de uma estatística de teste referente à hipótese de nulidade de todos os parâmetros associados à esta variável.
  3. Os graus de liberdade.
  4. O p-valor associado à hipótese testada naquela linha do quadro.

# ANOVA & MANOVA

- ▶ Na ANOVA (FISHER; MACKENZIE, 1923), avalia-se a relevância das variáveis sobre uma única resposta.
- ▶ Na MANOVA (SMITH; GNANADESIKAN; HUGHES, 1962), avalia-se a relevância das variáveis sobre mais de uma resposta.
- ▶ Formas conhecidas de se elaborar o quadro são as chamadas ANOVAs do tipo I, II e III.
- ▶ Esta nomenclatura vem do software estatístico SAS (INSTITUTE, 1985).
- ▶ 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; WEISBERG, 2019).

Introdução

Revisão de literatura

**Proposta**

Adaptação do teste Wald para os McGLM

Exemplos de hipóteses

ANOVA & MANOVA via teste Wald

Resultados preliminares

Considerações finais

Pendências

3

Proposta

UNIVERSIDADE DO PARANÁ

Introdução

Revisão de literatura

Proposta

**Adaptação do teste Wald para os McGLM**

Exemplos de hipóteses

ANOVA & MANOVA via teste Wald

Resultados preliminares

Considerações finais

Pendências

# 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  $h \times h$ .

[Introdução](#)[Revisão de literatura](#)[Proposta](#)[Adaptação do teste Wald para os McGLM](#)[Exemplos de hipóteses](#)[ANOVA & MANOVA via teste Wald](#)[Resultados preliminares](#)[Considerações finais](#)[Pendências](#)

Introdução

Revisão de literatura

Proposta

Adaptação do teste Wald para os McGLM

**Exemplos de hipóteses**

ANOVA & MANOVA via teste Wald

Resultados preliminares

Considerações finais

Pendências

## 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,  $r = 1, 2$ .
- ▶  $\beta_{r0}$  representa o intercepto.
- ▶  $\beta_{r1}$  um parâmetro de regressão associado a uma variável  $x_1$ .
- ▶ Considere que cada resposta possui apenas um parâmetro de dispersão:  $\tau_{r1}$ .
- ▶ 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.



# ANOVA & MANOVA via teste Wald

Introdução

Revisão de literatura

Proposta

Adaptação do teste Wald para os McGLM

Exemplos de hipóteses

**ANOVA & MANOVA via teste Wald**

Resultados preliminares

Considerações finais

Pendências

# ANOVA & MANOVA via teste Wald

- ▶ Com base na adaptação do teste Wald propostas, buscamos propor ANOVAs e MANOVAs via teste Wald.
- ▶ Propomos 3 tipos diferentes de análises de variância, nomeadas tipo I, II e III.
- ▶ Basicamente, cada linha do quadro corresponde uma hipótese. Portanto, basta especificar uma matriz  $L$ .
- ▶ Os procedimentos para análise de variância retornam um quadro para cada resposta.
- ▶ Os procedimentos para análise variância multivariadas retornam um único quadro.

[Introdução](#)[Revisão de literatura](#)[Proposta](#)[Adaptação do teste Wald para os McGLM](#)[Exemplos de hipóteses](#)[ANOVA & MANOVA via teste Wald](#)[Resultados preliminares](#)[Considerações finais](#)[Pendências](#)

# ANOVA & MANOVA via teste Wald

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

$$g_r(\mu_r) = \beta_{r0} + \beta_{r1}x_1 + \beta_{r2}x_2 + \beta_{r3}x_1x_2. \quad (5)$$

Em que:

- ▶ O índice  $r$  denota a variável resposta,  $r = 1, 2$ .
- ▶  $\beta_{r0}$  denota o intercepto de cada resposta.
- ▶ Temos três parâmetros de regressão para cada resposta:
  1.  $\beta_{r1}$  é o efeito de  $x_1$ .
  2.  $\beta_{r2}$  é o efeito de  $x_2$ .
  3.  $\beta_{r3}$  representa o efeito da interação entre as variáveis  $x_1$  e  $x_2$ .

# ANOVA tipo I

Nossa proposta de análise de variância do tipo I realiza testes sobre os parâmetros de regressão de forma sequencial. Neste cenário, os seguintes testes seriam efetuados:

1. Testa se todos os parâmetros são iguais a 0.
2. Testa se todos os parâmetros, exceto intercepto, são iguais a 0.
3. Testa se todos os parâmetros, exceto intercepto e os parâmetros referentes a  $x_1$ , são iguais a 0.
4. Testa se todos os parâmetros, exceto intercepto e os parâmetros referentes a  $x_1$  e  $x_2$ , são iguais a 0.

# ANOVA tipo II

Em nossa análise de variância do tipo II são feitos testes comparando o modelo completo contra o modelo sem todos os parâmetros que envolvem determinada variável.

1. Testa se o intercepto é igual a 0.
2. Testa se os parâmetros referentes a  $x_1$  são iguais a 0. Ou seja, é avaliado o impacto da retirada de  $x_1$  do modelo. Neste caso retira-se a interação pois nela há  $x_1$ .
3. Testa se os parâmetros referentes a  $x_2$  são iguais a 0. Ou seja, é avaliado o impacto da retirada de  $x_2$  do modelo. Neste caso retira-se a interação pois nela há  $x_2$ .
4. Testa se o efeito de interação é 0.

[Introdução](#)[Revisão de literatura](#)[Proposta](#)[Adaptação do teste Wald para os McGLM](#)[Exemplos de hipóteses](#)[ANOVA & MANOVA via teste Wald](#)[Resultados preliminares](#)[Considerações finais](#)[Pendências](#)

# ANOVA tipo III

Nossa análise de variância do tipo III considera o modelo completo contra o modelo sem determinada variável.

1. Testa se o intercepto é igual a 0.
2. Testa se os parâmetros de efeito principal referentes a  $x_1$  são iguais a 0. Ou seja, é avaliado o impacto da retirada de  $x_1$  nos efeitos principais do modelo.
3. Testa se os parâmetros de efeito principal referentes a  $x_2$  são iguais a 0. Ou seja, é avaliado o impacto da retirada de  $x_2$  nos efeitos principais do modelo.
4. Testa se o efeito de interação é 0.

# Comentários

- ▶ As análises de variância do tipo II e III geram os mesmos resultados quando aplicadas a modelos sem efeitos de interação.
- ▶ A diferença entre os tipos II e III é como são feitos os testes na presença de parâmetros de interação.
- ▶ As análises de variâncias multivariadas seguem o mesmo padrão de teste. Contudo a matriz  $L$  é expandida para comportar hipóteses sobre todas as respostas.
- ▶ De modo análogo ao que é feito para o preditor linear, propomos também uma análise de variância do tipo III para os parâmetros de dispersão do modelo.

[Introdução](#)[Revisão de literatura](#)[Proposta](#)[Adaptação do teste Wald para os McGLM](#)[Exemplos de hipóteses](#)[ANOVA & MANOVA via teste Wald](#)[Resultados preliminares](#)[Considerações finais](#)[Pendências](#)

4

# Resultados preliminares



## Funções implementadas

# Funções implementadas

Baseando-nos nas funcionalidades do pacote *car* (FOX; WEISBERG, 2019) e usando nossa adaptação do teste Wald implementamos uma série de funções:

| Função                              | Descrição  |
|-------------------------------------|--|
| <code>mc_linear_hypothesis()</code> | Hipóteses lineares gerais especificadas pelo usuário |
| <code>mc_anova_I()</code>           | ANOVA tipo I   |
| <code>mc_anova_II()</code>          | ANOVA tipo II  |
| <code>mc_anova_III()</code>         | ANOVA tipo III                                       |
| <code>mc_manova_I()</code>          | MANOVA tipo I  |
| <code>mc_manova_II()</code>         | MANOVA tipo II                                       |
| <code>mc_manova_III()</code>        | MANOVA tipo III                                      |
| <code>mc_anova_disp()</code>        | ANOVA tipo III para dispersão                        |
| <code>mc_manova_disp()</code>       | MANOVA tipo III para dispersão                       |

Tabela 2. Funções implementadas

5

# Considerações finais

# Considerações finais

- ▶ O McGLM contorna importantes restrições encontradas nas classes clássicas de modelos, como:
  1. A impossibilidade de modelar múltiplas respostas.
  2. A impossibilidade de modelar a dependência entre unidades.
  3. Disponibilidade de distribuições para modelagem.
- ▶ Nossa contribuição vai no sentido de fornecer ferramentas para uma melhor interpretação dos parâmetros estimados na classe.
- ▶ Nossa adaptação e implementações podem ser usadas para avaliar parâmetros de regressão, dispersão e potência.

# Considerações finais

- ▶ Todas as funções de prefixo *mc\_anova* e *mc\_manova* foram implementadas no sentido de facilitar o procedimento de análise da importâncias das variáveis.
- ▶ Implementamos funções para avaliar parâmetros de dispersão.
- ▶ A função *mc\_linear\_hypothesis()* da liberdade ao usuário de efetuar qualquer teste utilizando a estatística de Wald no contexto dos McGLM.
- ▶ A partir desta função é possível replicar os resultados de qualquer uma das funções de análise de variância e testar hipóteses mais gerais como:
  1. Igualdade de efeitos.
  2. Formular hipóteses com testes usando valores diferentes de zero .
  3. Formular hipóteses que combinem parâmetros de regressão, dispersão e potência.

6

# Pendências

# Pendências

- ▶ Propor e implementar procedimentos para realização de testes de comparações múltiplas.
- ▶ Adequar os testes para que sejam válidos para diferentes contrastes.
- ▶ Avaliar as propriedades e comportamento dos testes propostos com base em estudos de simulação.
- ▶ Motivar o potencial de aplicação das metodologias discutidas com base na aplicação a conjuntos de dados reais.


# Pendências


| Tarefa                                    | Data de início | Data de finalização |
|---|----------------|---------------------|
| Testes de comparações múltiplas           | 17/08          | 17/09               |
| Adaptação para diferentes contrastes      | 17/09          | 17/10               |
| Desenho e execução do estudo de simulação | 17/10          | 17/11               |
| Análise de dados                          | 17/11          | 17/12               |
| Sumarização dos resultados                | 17/12          | 17/01               |
| Entrega e defesa da dissertação           | 17/01          | 17/02               |


Tabela 3. Cronograma para cumprimento das pendências para titulação.





# Referências bibliográficas I


 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.

# Referências bibliográficas II



FOX, J.; WEISBERG, S. **An R Companion to Applied Regression**. Third. Thousand Oaks CA: Sage, 2019. Disponível em: <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.



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.



INSTITUTE, S. **SAS user's guide: Statistics**. [S.l.]: Sas Inst, 1985. v. 2.



JØRGENSEN, B. Exponential dispersion models. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 49, n. 2, p. 127–145, 1987.



JØRGENSEN, B. **The theory of dispersion models**. [S.l.]: CRC Press, 1997.




JØRGENSEN, B.; KOKONENDJI, C. C. Discrete dispersion models and their tweedie asymptotics. **AStA Advances in Statistical Analysis**, Springer, v. 100, n. 1, p. 43–78, 2015.





LEY, C.; BORDAS, S. P. What makes data science different? a discussion involving statistics 2.0 and computational sciences. **International Journal of Data Science and Analytics**, Springer, v. 6, n. 3, p. 167–175, 2018.


# Referências bibliográficas III


 NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society. Series A (General)**, v. 135, p. 370–384, 1972.


 PAULA, G. A. **Modelos de regressão: com apoio computacional**. [S.l.]: IME-USP São Paulo, 2004.


 POURAHMADI, M. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. **Biometrika**, Oxford University Press, v. 87, n. 2, p. 425–435, 2000.

 R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>.

 RAO, C. R. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In: CAMBRIDGE UNIVERSITY PRESS. **Mathematical Proceedings of the Cambridge Philosophical Society**. [S.l.], 1948. v. 44, n. 1, p. 50–57.

 SILVEY, S. D. The lagrangian multiplier test. **The Annals of Mathematical Statistics**, JSTOR, v. 30, n. 2, p. 389–407, 1959.

 SMITH, H. et al. Multivariate analysis of variance (manova). **Biometrics**, JSTOR, v. 18, n. 1, p. 22–41, 1962.

 WALD, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. **Transactions of the American Mathematical society**, JSTOR, v. 54, n. 3, p. 426–482, 1943.

# Referências bibliográficas IV



WEIHS, C.; ICKSTADT, K. Data science: the impact of statistics. **International Journal of Data Science and Analytics**, Springer, v. 6, n. 3, p. 189–194, 2018.



WILKS, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. **The annals of mathematical statistics**, JSTOR, v. 9, n. 1, p. 60–62, 1938.

# Obrigado!

Lineu Alberto Cavazani de Freitas

lineuacf@gmail.com

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

PPG Informática

