

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 Ciência de dados
- 2 Modelos de regressão
 - Modelos multivariados de covariância linear generalizada
- 3 Testes de hipóteses
- 4 Proposta
- 5 Resultados preliminares
 - Adaptação do teste Wald para os McGLM
 - Exemplos de hipóteses
 - ANOVA & MANOVA via teste Wald
 - Funções implementadas
- 6 Próximas etapas
- 7 Considerações finais

Ciência de dados

Modelos de regressão

Testes de hipóteses

Proposta

Resultados preliminares

Próximas etapas

Considerações finais

1

Ciência de dados

Ciência de dados

- ▶ **Ciência de dados** é campo de estudo interdisciplinar que incorpora conhecimento de áreas como:
 1. Estatística.
 2. Ciência da computação.
 3. Matemática.
- ▶ 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.

2

Modelos de regressão

Modelos de regressão

Três conceitos são importantes para entender minimamente o funcionamento de um modelo de regressão:

- ▶ **Fenômeno aleatório.**
- ▶ **Variável aleatória.**
- ▶ **Distribuição de probabilidade.**

Modelos de regressão

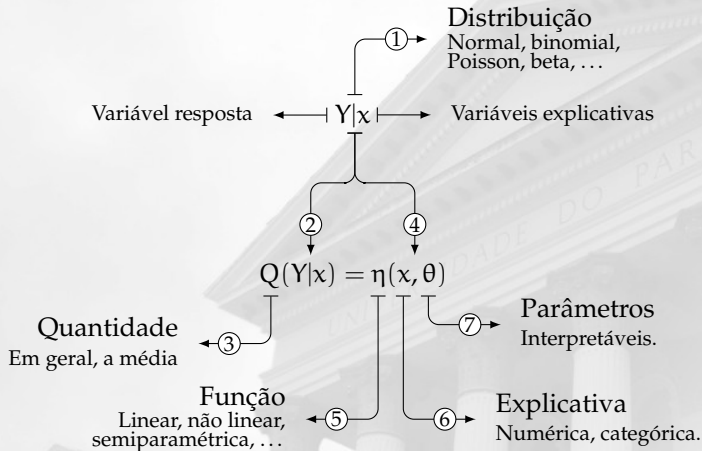
- ▶ **Fenômeno aleatório:** situação na qual diferentes observações podem fornecer diferentes desfechos.
- ▶ **Variáveis aleatórias:** mecanismos que associam um valor numérico a cada desfecho possível do fenômeno.
 - ▶ Podem ser discretas ou contínuas.
 - ▶ Existem probabilidades associadas aos valores de uma variável aleatória.
 - ▶ Estas probabilidades podem ser descritas por funções.
- ▶ **Distribuições de probabilidade:** modelos probabilísticos que buscam descrever as probabilidades de variáveis aleatórias.

Modelos de regressão

- ▶ Na prática, 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

- ▶ Em regressão **modelamos parâmetros** das distribuições como uma função de **variáveis explicativas**.
- ▶ O parâmetro de interesse é decomposto em uma combinação linear de novos parâmetros que associam as **variáveis explicativas** à **variável resposta**.
- ▶ Obtém-se uma **equação que explique a relação** entre as variáveis.



Modelos de regressão

1. Definição do problema.

- ▶ Qual o fenômeno aleatório de interesse?
- ▶ Que fatores externos podem afetar este fenômeno?

2. Planejamento do estudo e coleta de dados.

3. Análise dos dados via regressão.

- ▶ Escolha da distribuição de probabilidade.
- ▶ Especificação do modelo.
- ▶ Obtenção dos parâmetros (ajuste).
- ▶ Diagnóstico.

4. Interpretação dos resultados.

- ▶ Quais os fatores externos apresentam ou não impacto sobre o fenômeno.
- ▶ Qual a dimensão desse impacto.

Modelos de regressão

- ▶ Existem modelos univariados e multivariados.
 - ▶ **Univariados:** apenas uma variável resposta.
 - ▶ **Multivariados:** mais de uma variável resposta.
- ▶ Em ambos os casos o interesse é avaliar o **efeito de variáveis explicativas**.
- ▶ Existem inúmeras classes de modelos de regressão, dentre elas:
 - ▶ Modelo linear normal.
 - ▶ Modelos lineares generalizados.
 - ▶ Modelos multivariados de covariância linear generalizada.

Modelo linear normal

- ▶ O modelo linear normal (GALTON, 1886) ficou famoso por suas facilidades computacionais.
- ▶ Possui pressupostos difíceis de serem atendidos na prática.
 - ▶ Independência.
 - ▶ Normalidade.
 - ▶ Variância constante.
- ▶ Diversas técnicas foram propostas para solucionar casos em que os pressupostos fossem violados.

Modelos lineares generalizados

- ▶ O avanço computacional permitiu o surgimento de modelos mais gerais que necessitavam de processos iterativos para estimação dos parâmetros.
- ▶ Surgem os modelos lineares generalizados (GLM) (NELDER; WEDDERBURN, 1972).
- ▶ Os GLMs permitem utilizar qualquer membro da família exponencial de distribuições.
- ▶ Casos especiais: Bernoulli, binomial, Poisson, normal, gama, normal inversa, entre outras.

Modelos multivariados de covariância linear generalizada

- ▶ Apesar do grande potencial, os GLMs apresentam 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 modelos multivariados de covariância linear generalizada (McGLMs) (BONAT; JØRGENSEN, 2016) contornam estas restrições.

Modelos multivariados de covariância linear generalizada

Para definição de um McGLM 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.
- ▶ \mathbf{X}_r denota uma matriz de delineamento $N \times k_r$.
- ▶ $\boldsymbol{\beta}_r$ denota um vetor $k_r \times 1$ de parâmetros de regressão.

Modelos multivariados de covariância linear generalizada

Considere ainda:

- ▶ Σ_b uma matriz de correlação entre variáveis resposta, de ordem $R \times R$.
- ▶ Σ_r , $r = 1, \dots, R$, a matriz de variância e covariância para cada resposta r , de dimensão $N \times N$:

$$\Sigma_r = V_r (\mu_r; p_r)^{1/2} (\Omega(\tau_r)) V_r (\mu_r; p_r)^{1/2}.$$

Em que:

- ▶ $V_r (\mu; p)$ é uma matriz diagonal em que as entradas principais são dadas pela função de variância aplicada ao vetor μ .
- ▶ p_r é o parâmetro de potência.
- ▶ $\Omega(\tau_r)$ a matriz de dispersão que descreve a parte da covariância dentro de cada variável resposta.

Preditor linear matricial

- ▶ A matriz $\mathbf{\Omega}(\boldsymbol{\tau}_r)$ descreve a estrutura de correlação entre as observações da amostra.
- ▶ É modelada através de um preditor linear matricial combinado com uma função de ligação de covariância:

$$h\{\mathbf{\Omega}(\boldsymbol{\tau}_r)\} = \tau_{r0}Z_0 + \dots + \tau_{rD}Z_D$$

- ▶ $h()$ é a função de ligação de covariância.
- ▶ Z_{rd} com $d = 0, \dots, D$ são matrizes que representam a estrutura de covariância presente em cada variável resposta r .
- ▶ $\boldsymbol{\tau}_r = (\tau_{r0}, \dots, \tau_{rD})$ é um vetor $(D + 1) \times 1$ de parâmetros de dispersão.

Funções de variância

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.

Modelos multivariados de covariância linear generalizada

Os McGLMs são definidos por:

$$E(\mathbf{Y}) = \mathbf{M} = \{g_1^{-1}(\mathbf{X}_1\boldsymbol{\beta}_1), \dots, g_R^{-1}(\mathbf{X}_R\boldsymbol{\beta}_R)\}$$

$$\text{Var}(\mathbf{Y}) = \mathbf{C} = \boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b$$

Em que:

- ▶ $\boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b = \text{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \dots, \tilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b \otimes \mathbf{I})\text{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^T, \dots, \tilde{\boldsymbol{\Sigma}}_R^T)$ é o produto generalizado de Kronecker.
- ▶ $\tilde{\boldsymbol{\Sigma}}_r$ denota a matriz triangular inferior da decomposição de Cholesky da matriz $\boldsymbol{\Sigma}_r$.
- ▶ $\text{Bdiag}()$ denota a matriz bloco-diagonal.
- ▶ \mathbf{I} uma matriz identidade $N \times N$.
- ▶ $g_r()$ são as tradicionais funções de ligação.

Modelos multivariados de covariância linear generalizada

- ▶ Parâmetros estimados nos McGLMs:
 1. Regressão.
 2. Dispersão.
 3. Potência.
 4. Correlação.
- ▶ Todas estas quantidades são interpretáveis e são estimadas com base nos dados.
- ▶ A estimação é feita por meio de **funções de estimação**.
 1. **Função quasi-score** para parâmetros de regressão.
 2. **Função de estimação de Pearson** para os demais parâmetros.

Funções de estimação

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

Em que:

- ▶ β_r denota um vetor $k_r \times 1$ de parâmetros de regressão.
- ▶ λ 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.
- ▶ $\mathbf{D} = \nabla_{\beta} \mathcal{M}$ é uma matriz $NR \times K$, e ∇_{β} 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 $S_{\theta}^{-\top} = (S_{\theta}^{-1})^{\top}$.

Comentários sobre os McGLMs

- ▶ Configuram uma estrutura geral para análise via modelos de regressão.
- ▶ Comporta múltiplas respostas de diferentes naturezas.
- ▶ Pode-se ajustar modelos com diferentes preditores e distribuições para cada resposta.
- ▶ Os modelos levam em conta a correlação entre indivíduos do conjunto de dados.

Comentários sobre os McGLMs

- ▶ Os parâmetros são interpretáveis:
 - ▶ **Parâmetros de regressão:** efeito das variáveis explicativas sobre as respostas.
 - ▶ **Parâmetros de dispersão:** impacto da correlação entre unidades.
 - ▶ **Parâmetros de potência:** indicativo de qual distribuição se adequa ao problema.
- ▶ A metodologia do McGLM está implementada no pacote *mcglm* (BONAT, 2018) do software R.

3

Testes de hipóteses

Testes de hipóteses

- ▶ Inferência: **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?

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 θ um parâmetro, um teste de hipóteses sobre θ é dado por:

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

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 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

- ▶ Usados para verificar se a retirada de determinada variável explicativa do modelo geraria uma perda no ajuste.
- ▶ Os três testes mais usados 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).
- ▶ São baseados na função de verossimilhança dos modelos.
- ▶ São assintoticamente equivalentes.

Teste Wald

- ▶ Requer apenas um modelo ajustado.
- ▶ 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.

Teste Wald

Considere um modelo de regressão em que:

- ▶ β um vetor com p parâmetros de regressão.
- ▶ $\hat{\beta}$ as estimativas dos parâmetros.
- ▶ c um vetor de valores postulados de dimensão q .
- ▶ L uma matriz de especificação das hipóteses, de dimensão $q \times p$.

Teste Wald

- ▶ As hipóteses podem ser descritas como:

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

- ▶ A estatística de teste é dada por:

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

- ▶ $WT \sim \chi^2_q$.

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.

ANOVA & MANOVA

- ▶ Os resultados são sumarizados numa tabela, o chamado **quadro de análise de variância**.
- ▶ 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.

Tópicos abordados até aqui...

1. Importância dos modelos de regressão em ciência de dados.
2. Classes relevantes.
3. McGLMs e importância dos parâmetros de regressão, dispersão e potência.
4. Elementos de testes de hipóteses.
5. Testes de hipóteses em modelos de regressão, ênfase no teste Wald.
6. Procedimentos baseados em testes de hipóteses: ANOVA e MANOVA.

TH MCGLM

Lineu Alberto

Ciência de dados

Modelos de regressão

Testes de hipóteses

Proposta

Resultados preliminares

Próximas etapas

Considerações finais

4

Proposta

Proposta

- ▶ Considerando os McGLMs, não há discussão a respeito da construção de testes de hipóteses.
- ▶ Considerando o alto potencial de aplicação dos McGLMs em ciência de dados, nossos objetivos gerais são:
 - ▶ Desenvolvimento de testes de hipóteses para avaliação dos parâmetros de McGLMs.
 - ▶ Adaptação do teste de Wald clássico utilizado em modelos lineares para os McGLMs.

Etapas

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

Ciência de dados

Modelos de regressão

Testes de hipóteses

Proposta

Resultados preliminares

Adaptação do teste Wald para os McGLM

Exemplos de hipóteses

ANOVA & MANOVA via teste Wald

Funções implementadas

Próximas etapas

Considerações finais

5

Resultados preliminares

Adaptação do teste Wald para os McGLMs

Hipóteses

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

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

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

Em que:

- ▶ \mathbf{L} é a matriz da especificação das hipóteses, 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.
- ▶ \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

Ciência de dados

Modelos de regressão

Testes de hipóteses

Proposta

Resultados preliminares

Adaptação do teste Wald para os McGLM

Exemplos de hipóteses

ANOVA & MANOVA via teste Wald

Funções implementadas

Próximas etapas

Considerações finais

Exemplos de hipóteses

- ▶ Considere o exemplo disponível na seção 4.3.6 do livro **Modelos de Regressão com Apoio Computacional** (PAULA, 2004):
- ▶ Os dados referem-se a um estudo sobre demanda de TV's a cabo em 40 regiões dos Estados Unidos.
- ▶ Algumas das variáveis coletadas foram:
 - ▶ Número de assinantes de TV a cabo (em milhares).
 - ▶ Percentual de domicílios com TV a cabo.
 - ▶ Renda per capita por domicílio com TV a cabo (em USD).
 - ▶ Custo médio mensal de manutenção de TV a cabo (em USD).

Exemplos de hipóteses

► Variáveis resposta:

1. Número de assinantes de TV a cabo (em milhares).
2. Percentual de domicílios com TV a cabo.

► Variáveis explicativas:

1. Renda per capita por domicílio com TV a cabo (em USD).
2. Custo médio mensal de manutenção de TV a cabo (em USD).

- No problema existem duas variáveis resposta, de diferentes naturezas: uma contagem e uma proporção.
- A estrutura do problema é ideal para utilizar um modelo multivariado.

Para análise do problema, considere um modelo bivariado:

$$g_r(\mu_r) = \beta_{r0} + \beta_{r1}\text{renda} + \beta_{r2}\text{custo} + \beta_{r3}\text{renda} : \text{custo}.$$

Em que:

- ▶ O índice r denota a variável resposta, $r = 1, 2$.
 - ▶ 1: Número de assinantes de TV a cabo (em milhares).
 - ▶ 2: Percentual de domicílios com TV a cabo.
- ▶ β_{r0} denota o intercepto de cada resposta.
- ▶ Temos três parâmetros de regressão para cada resposta:
 1. β_{r1} é o efeito de renda.
 2. β_{r2} é o efeito de custo.
 3. β_{r3} representa o efeito da interação entre as variáveis renda e custo.
- ▶ Considere ainda que:
 - ▶ Cada resposta possui apenas um parâmetro de dispersão: τ_{r0} .
 - ▶ As unidades em estudo são independentes, logo $Z_0 = \mathbf{I}$.
 - ▶ Os parâmetros de potência foram fixados.

Exemplos de hipóteses

Algumas perguntas de interesse podem ser:

- ▶ Existe efeito de renda per capita por domicílio com TV a cabo (em USD) sobre o número de assinantes de TV a cabo (em milhares)?
- ▶ Existe efeito de custo médio mensal de manutenção de TV a cabo (em USD) sobre o número de assinantes de TV a cabo (em milhares) E percentual de domicílios com TV a cabo?

Exemplo 1: efeito de renda sobre número de assinantes

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:

- ▶ $\boldsymbol{\theta}_{\beta,\tau,p}^T = [\beta_{10} \ \beta_{11} \ \beta_{12} \ \beta_{13} \ \beta_{20} \ \beta_{21} \ \beta_{22} \ \beta_{23} \ \tau_{10} \ \tau_{20}]$.
- ▶ $\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$.
- ▶ $\mathbf{c} = [0]$, é o valor da hipótese nula.

Exemplo 2: efeito de custo com manutenção sobre número de assinantes e percentual de domicílios com TV a cabo

Considere a hipótese:

$$H_0 : \beta_{r2} = 0 \text{ vs } H_1 : \beta_{r2} \neq 0.$$

Ou, da mesma forma:

$$H_0 : \begin{pmatrix} \beta_{12} \\ \beta_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ vs } H_1 : \begin{pmatrix} \beta_{12} \\ \beta_{22} \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Exemplo 2: efeito de custo com manutenção sobre número de assinantes e percentual de domicílios com TV a cabo

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:

- ▶ $\boldsymbol{\theta}_{\beta,\tau,p}^T = [\beta_{10} \ \beta_{11} \ \beta_{12} \ \beta_{13} \ \beta_{20} \ \beta_{21} \ \beta_{22} \ \beta_{23} \ \tau_{10} \ \tau_{20}]$.
- ▶ $\mathbf{L} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$.
- ▶ $\mathbf{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, é o valor da hipótese nula.

Ciência de dados

Modelos de regressão

Testes de hipóteses

Proposta

Resultados preliminares

Adaptação do teste Wald para os McGLM

Exemplos de hipóteses

ANOVA & MANOVA via teste Wald

Funções implementadas

Próximas etapas

Considerações finais

ANOVA & MANOVA via teste Wald

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.
- ▶ 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.

[Ciência de dados](#)[Modelos de regressão](#)[Testes de hipóteses](#)[Proposta](#)[Resultados preliminares](#)[Adaptação do teste Wald para os McGLM](#)[Exemplos de hipóteses](#)[ANOVA & MANOVA via teste Wald](#)[Funções implementadas](#)[Próximas etapas](#)[Considerações finais](#)

ANOVA & MANOVA tipo II

- ▶ São feitos testes comparando o modelo completo contra o modelo sem determinada variável.
- ▶ Portanto, considerando o exemplo:
 1. Testa se os parâmetros referentes a **renda** são iguais a 0. Ou seja, é avaliado o impacto da retirada de renda do modelo. Neste caso retira-se a interação pois nela há renda.
 2. Testa se os parâmetros referentes a **custo** são iguais a 0. Ou seja, é avaliado o impacto da retirada de custo do modelo. Neste caso retira-se a interação pois nela há custo.
 3. Testa se o efeito de **interação** é 0.

Funções implementadas

Ciência de dados

Modelos de regressão

Testes de hipóteses

Proposta

Resultados preliminares

Adaptação do teste Wald para os McGLM

Exemplos de hipóteses

ANOVA & MANOVA via teste Wald

Funções implementadas

Próximas etapas

Considerações finais

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

6

Próximas etapas

Próximas etapas

- ▶ 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.

Cronograma

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.

Ciência de dados

Modelos de regressão

Testes de hipóteses

Proposta

Resultados preliminares

Próximas etapas

Considerações finais


7


Considerações finais


Considerações finais


- ▶ O McGLM contorna importantes restrições encontradas nas classes clássicas de modelos.
- ▶ Nossa contribuição vai no sentido de fornecer ferramentas para uma melhor interpretação dos parâmetros estimados na classe.
- ▶ Nossa contribuição visa formas de responder sobre:
 1. Importância das variáveis explicativas no problema.
 2. Impacto das medidas correlacionadas no conjunto de dados.
 3. Qual distribuição se adequa ao problema.


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.

 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.









 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.

 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.


 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.

Referências bibliográficas II

-  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.
-  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.
-  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 III

 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



Ciência de dados

Modelos de regressão

Testes de hipóteses

Proposta

Resultados preliminares

Próximas etapas

Considerações finais