Testes de hipótese em Modelos Multivariados de Covariância Linear Generalizada (McGLM)

Lineu Alberto Cavazani de Freitas
Orientador: Prof. Dr. Wagner Hugo Bonat

PPG Informática
Data Science & Big Data
Universidade Federal do Paraná

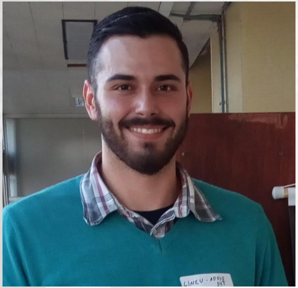
https://lineu96.github.io/st/
lineuacf@gmail.com

1 Quem sou eu

McGLM

Quem sou eu

- Estatístico formado pela Universidade Federal do Paraná (UFPR) em 2019.
- Atualmente mestrando no Programa de Pós Graduação em Informática da UFPR.
- Inserido na área de concentração Ciência da Computação, linha de pesquisa Tecnologia da Informação e grupo de pesquisa Data Science & Big Data.



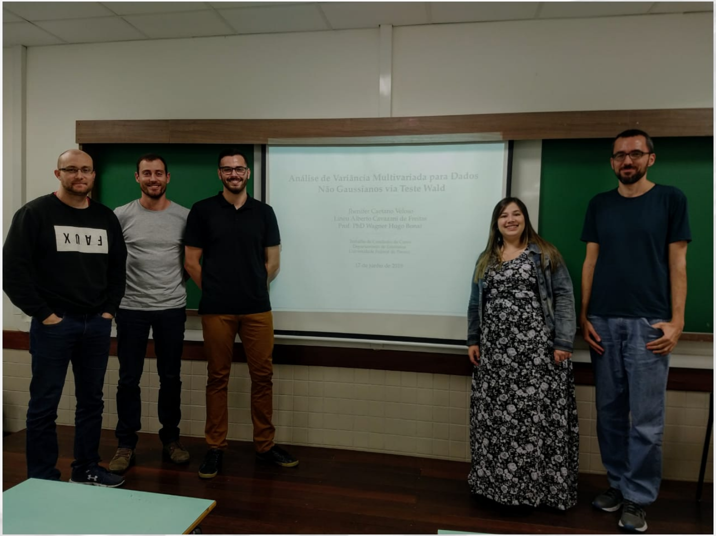
Sumário

2

Introdução

Onde tudo começou

- O projeto teve início em 2018 quando eu e uma colega de curso (Jhenifer Caetano Veloso), desenvolvemos nosso TCC sob orientação do professor Wagner.
- O título do trabalho foi "Análise de Variância Multivariada para Dados Não Gaussianos via Teste Wald".


Análise de Variância Multivariada para Dados
Não Gaussianos via Teste Wald
Jhenifer Caetano Veloso
Lineu Alberto Cavazani de Freitas
Prof. PhD Wagner Hugo Bonat

Plano para o mestrado

- Na graduação avaliamos o teste Wald para gerar quadros de análise de variância multivariadas do tipo III (MANOVA).
- A ideia é dar continuidade e ampliar o foco do trabalho que teve início na graduação.
- O título atual do trabalho é "Testes de hipótese em Modelos Multivariados de Covariância Linear Generalizada (McGLM)".
- Nosso objetivo é explorar o teste Wald para testar hipóteses gerais sobre parâmetros de regressão, dispersão ou potência de um McGLM.
- Bem como obter quadros de ANOVA e MANOVA para parâmetros de regressão e dispersão.

As etapas do trabalho são:

- Adaptar o teste Wald para realização de testes de hipótese gerais sobre parâmetros de Modelos Multivariados de Covariância Linear Generalizada (McGLM).
- Implementar funções para efetuar tais testes, bem como funções para efetuar Análises de Variância e Análises de Variância Multivariadas para os McGLM.
- Demonstrar as propriedades e comportamento dos testes propostos com base em estudos de simulação.
- Demonstrar o potencial de aplicação das metodologias discutidas com base na aplicação a conjuntos de dados reais.

3

McGLM

Para definição de um McGLM considere:

- $Y_{N \times R} = \{Y_1, \dots, Y_R\}$ uma matriz de variáveis resposta.
- $M_{N \times R} = \{\mu_1, \dots, \mu_R\}$ uma matriz de valores esperados.
- Σ_r, $r = 1,\dots, R$, a matriz de variância e covariância para cada resposta r, de dimensão NxN.
- Σ_b uma matriz de correlação, de ordem $R \times R$, que descreve a correlação entre as variáveis resposta.
- X_r denota uma matriz de delineamento $N \times k_r$.
- β_r denota um vetor $k_r \times 1$ de parâmetros de regressão.

Os McGLMs são definidos por:

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

Em que:

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

Matriz de variância e covariância

- Para variáveis resposta contínuas, binárias, binomiais, proporções ou índices a matriz de variância e covariância $\boldsymbol{\Sigma}_r$ é dada por:

$$\Sigma_r = V\left(\mu_r; p_r\right)^{1/2} \left(\boldsymbol{\Omega}\left(\boldsymbol{\tau}_r\right)\right) V\left(\mu_r; p_r\right)^{1/2}$$

- No caso de variáveis resposta que sejam contagens a matriz de variância e covariância para cada variável resposta fica dada por:

$$\Sigma_r = \mathrm{diag}(\mu_r) + V\left(\mu_r; p_r\right)^{1/2} \left(\boldsymbol{\Omega}\left(\boldsymbol{\tau}_r\right)\right) V\left(\mu_r; p_r\right)^{1/2}$$

- $V\left(\mu_r; p_r\right) = \mathrm{diag}(\vartheta(\mu_r; p_r))$ denota uma matriz diagonal na qual as entradas são dadas pela função de variância $\vartheta(\cdot; p_r)$ aplicada aos elementos do vetor μ_r.

Função de variância

- Função de variância potência:
  - Caracteriza a família Tweedie de distribuições.
  - É dada por $\vartheta(\cdot; p_r) = \mu_r^{p_r}$.
  - Casos particulares: Normal ($p = 0$), Poisson ($p = 1$), gama ($p = 2$) e Normal inversa ($p = 3$).
- Função de dispersão Poisson–Tweedie:
  - Visa contornar a inflexibilidade da utilização da função de variância potência para respostas que caracterizam contagens.
  - É dada por $\vartheta(\cdot; p) = \mu + \tau\mu^p$ em que τ é o parâmetro de dispersão.
  - Casos particulares: Hermite ($p = 0$), Neyman tipo A ($p = 1$), binomial negativa ($p = 2$) e Poisson–inversa gaussiana ($p = 3$).
- Função de variância binomial:
  - Indicada quando a variável resposta é binária, restrita a um intervalo ou quando tem-se o número de sucessos em um número de tentativas.
  - É dada por $\vartheta(\cdot; p_r) = \mu_r^{p_{r1}}(1 - \mu_r)^{p_{r2}}$

Parâmetro de potência

- O parâmetro de potência p aparece em todas as funções de variância discutidas.
- Este parâmetro tem especial importância pois trata-se de um índice que distingue diferentes distribuições de probabilidade.
- Pode ser utilizado como uma ferramenta para seleção automática da distribuição de probabilidade que mais se adequa ao problema.

Preditor linear matricial

- A matriz de dispersão $\boldsymbol{\Omega}(\boldsymbol{\tau})$ descreve a parte da covariância dentro de cada variável resposta que não depende da estrutura média.
- Isto é, a estrutura de correlação entre as observações da amostra.
- A matriz de dispersão é modelada através de um preditor linear matricial combinado com uma função de ligação de covariância.
- O preditor linear matricial é dado por:

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

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

Comentários sobre a classe

- Os McGLM configuram uma estrutura geral para análise via modelos de regressão.
- Comporta múltiplas respostas não gaussianas.
- Não se faz suposições quanto à independência das observações.

4
Estimação e inferência

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:

$$\psi_{\beta}(\beta, \lambda) = \mathbf{D}^\top \mathbf{C}^{-1}(\mathcal{Y} - \mathcal{M})$$
$$\psi_{\lambda_i}(\beta, \lambda) = \mathrm{tr}(W_{\lambda i}(\mathbf{r}^\top \mathbf{r} - \mathbf{C})), i = 1,.., Q$$

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.
- $\mathcal{Y}$ é um vetor $NR \times 1$ com os valores da matriz de variáveis respostas $Y_{N \times R}$ empilhados.
- $\mathcal{M}$ é um vetor $NR \times 1$ com os valores da matriz de valores esperados $M_{N \times R}$ empilhados.
- $\mathbf{D} = \nabla_{\beta} \mathcal{M}$ é uma matriz $NR \times K$, e ∇_{β} denota o operador gradiente.
- $W_{\lambda i} = -\frac{\partial \mathbf{C}^{-1}}{\partial \lambda_i}$
- $\mathbf{r} = (\mathcal{Y} - \mathcal{M})$

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:

$$\beta^{(i+1)} = \beta^{(i)} - S_\beta^{-1}\psi\beta(\beta^{(i)}, \lambda^{(i)}),$$
$$\lambda^{(i+1)} = \lambda^{(i)}\alpha S_\lambda^{-1}\psi\lambda(\beta^{(i+1)}, \lambda^{(i)}).$$

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

5

Teste Wald

Teste Wald

- E um teste de hipóteses largamente empregado para avaliar suposições sobre parâmetros de um modelo de regressão.
- Isto é, verifcar se existe evidência suficiente para afirmar que o parâmetro é ou não estatísticamente igual a um valor qualquer.

Teste Wald

- A grosso modo, é um teste que avalia a distância entre a estimativa do parâmetro e o valor postulado sob a hipótese nula.
- Esta diferença é ainda ponderada por uma medida de precisão da estimativa do parâmetro.
- Quanto mais distante de 0 for o valor da distância ponderada, menor é a chance da hipótese de igualdade ser verdadeira, ou seja, do valor postulado ser igual ao valor estimado.

Teste Wald

- Além destes elementos o teste pressupõe que os estimadores dos parâmetros do modelo sigam distribuição assintótica Normal.
- Para avaliação da estatística de teste e verificação de significância estatística utiliza-se distribuição assintótica Qui-quadrado (χ^2).

Hipóteses

As hipóteses a serem testadas podem ser escritas como:

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

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:

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

Em que:

- $\mathbf{L}$ é a mesma matriz da especificação das hipóteses a serem testadas, tem dimensão $s \times h$.
- $\hat{\boldsymbol{\theta}}_{\boldsymbol{\beta},\boldsymbol{\tau},\mathbf{p}}$ 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.
- E $J^{-1}_{\boldsymbol{\beta},\boldsymbol{\tau},\mathbf{p}}$ é a inversa da matriz de informação de Godambe desconsiderando os parâmetros de correlação, de dimensão $h \times h$.

A matriz L

- Cada coluna da matriz $\mathbf{L}$ corresponde a um dos h parâmetros do modelo e cada linha a uma hipótese.
- Sua construção consiste basicamente em preencher a matriz com 0, 1 e eventualmente -1 de tal modo que o produto $\mathbf{L}\boldsymbol{\theta}_{\boldsymbol{\beta},\boldsymbol{\tau},\mathbf{p}}$ represente corretamente a hipótese de interesse.

Comentários finais

- É possível testar qualquer parâmetro individualmente ou até mesmo formular hipóteses para diversos parâmetros simultaneamente, sejam eles de regressão, dispersão ou potência.
- Independente do número de parâmetros testados, a estatística de teste W é um único valor que segue assintóticamente distribuição χ^2.
- Os graus de liberdade são dados pelo número de parâmetros testados, isto é, o número de linhas da matriz $\mathbf{L}$, denotado por s.

6

Construção da matriz L

Exemplos de hipóteses que podem ser testadas.
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
Considere a hipótese:

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

Esta hipótese pode ser reescrita na seguinte notação:

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

Em que:

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

Exemplo 2
Considere a hipótese:

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

Ou, da mesma forma:

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

Exemplo 2

A hipótese pode ser reescrita na seguinte notação:

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

Em que:

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

Exemplo 3

Considere a hipótese:

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

Esta hipótese pode ser reescrita na seguinte notação:

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

Em que:

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

7

ANOVA via teste Wald

A análise de variância

- 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 destes sucessivos testes são sumarizados numa tabela que contêm em cada linha:
  - A variável.
  - O valor de uma estatística de teste.
  - Os graus de liberdade.
  - E um p-valor.

Precauções

- Cuidados devem ser tomados no que diz respeito à forma como o quadro foi elaborado.
- Cada linha do quadro refere-se a uma hipótese e estas hipóteses podem ser elaboradas de formas distintas.
- Formas conhecidas de se elaborar o quadro são as chamadas ANOVAs do tipo I, II e III.
- Esta nomenclatura vem do software estatístico SAS, contudo as implementações não necessariamente correspondem ao que está implementado no SAS.
- Recomenda-se ao usuário estar seguro de qual tipo de análise está sendo utilizada pois, caso contrário, interpretações equivocadas podem ser tomadas.

ANOVA via teste Wald

- As Análises de Variância são sucessivos testes de hipótese que verificam a nulidade de determinados parâmetros.
- Isto geralmente é feito através de uma sequência de testes de Razão de Verossimilhança.
- Para as análises do tipo II e III é simples visualizar como gerar os quadros de Análise de Variância utilizando o teste Wald.
- Pois sempre estarão sendo comparados o modelo completo e o modelo sem determinada ou determinadas variáveis.
- Ou seja, basta então, para cada linha do quadro de Análise de Variância, especificar corretamente uma matriz L que represente de forma adequada a hipótese a ser testada.

MANOVA via teste Wald

- Do mesmo modo que é feito para um modelo univariado, podemos chegar também a uma Análise de Variância Multivariada (MANOVA).
- Basta realizar sucessivos testes para avaliar o efeito de determinada variável nas R respostas simultaneamente.
- Portanto, a pergunta que a ser respondida seria: esta variável tem efeito diferente de 0 para todas as respostas?

8

Funções implementadas

Baseando-nos nas funções do pacote *car*, temos funções implementadas para

- Análises de Variância por variável resposta (ANOVA).
- Análises de Variância multivariadas (MANOVA). Note que no caso da MANOVA os preditores devem ser iguais para todas as respostas sob análise.
- Análise de variância focados no preditor linear matricial. O objetivo é verificar a significância dos parâmetros de dispersão.
- Hipóteses lineares gerais em que todos os elementos são especificadas pelo usuário, na qual é possível testar hipóteses sobre parâmetros de regressão, dispersão ou potência.

Funções implementadas

Função	Descrição
mc_anova_I()	ANOVA tipo I (imita uma sequencial)
mc_anova_II()	ANOVA tipo II (nao bate com o car)
mc_anova_III()	ANOVA tipo III
mc_anova_disp()	ANOVA tipo III para dispersão
mc_manova_I()	MANOVA tipo I (imita uma sequencial)
mc_manova_II()	MANOVA tipo II (nao bate com o car)
mc_manova_III()	MANOVA tipo III
mc_manova_disp()	MANOVA tipo III para dispersão
mc_linear_hypothesis()	Hipóteses lineares gerais especificadas pelo usuário

Tabela 1

9

Próximos passos

O que temos até o momento

- Algum texto.
- Protótipo das funções.
- Conjuntos de dados para aplicação.

Tarefas a cumprir

- Entender porquê nossos resultados não batem com o **car** na ANOVA tipo II.
- Avaliar relevância da ANOVA sequencial.
- Formular e executar o estudo de simulação.
- Documentar e reportar.

Considerando a área de pesquisa, o trabalho teria as seguintes contribuições:

1. Adaptar um teste existente para uma classe de modelos não usual mas com alto potencial de aplicação.
2. Realizar um estudo pesado de simulação para verificar o funcionamento da forma que estamos propondo.
3. Análisar de dados provenientes de estudos reais para demonstrar a aplicabilidade das funcionalidades.

Obrigado!

Lineu Alberto Cavazani de Freitas
lineuacf@gmail.com
https://lineu96.github.io/st/
PPG Informática

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ