Teste Wald para avaliação de parâmetros de regressão e dispersão em modelos multivariados de covariância linear generalizada

Defesa de mestrado

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

PPG Informática UFPR

Sumário

Introdução

Introdução

Introdução

1. Motivação
   - Ciência de dados.
   - Modelos de regressão.
   - Testes de hipóteses.
   - Procedimentos baseados em testes de hipóteses.
2. Desafio e hipótese
3. Objetivo
4. Contribuição

Motivação

Motivação

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 and Ickstadt, 2018].

- Neste sentido, os **modelos de regressão** tem papel importante.

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.

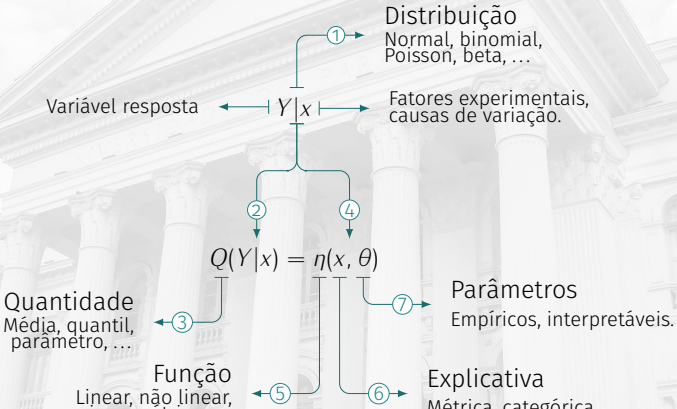
Modelos de regressão

- 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** por meio 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



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.**
   - Estudo observacional x estudo experimental.
   - Representação tabular.
3. **Análise dos dados via regressão.**
   - 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 (**GLMs**) [Nelder and 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 and Jørgensen, 2016] contornam estas restrições.

Modelos multivariados de covariância linear generalizada

- 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 observações** do conjunto de dados.

Modelos multivariados de covariância linear generalizada

- 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.
  - **Parâmetros correlação**: força de associação entre respostas.
- As estimativas dos parâmetros podem ser estudadas por meio de **testes de hipóteses**.

Testes de hipóteses em modelos de regressão

- Usados para avaliar o **efeito** das variáveis explicativas.
- Os três testes mais usados são:
  - O teste da razão de verossimilhanças [Wilks, 1938].
  - O teste do multiplicador de lagrange ou teste escore [Aitchison and Silvey, 1958, Silvey, 1959, Rao, 1948].
  - **O teste Wald [Wald, 1943].**
- São baseados na função de verossimilhança dos modelos.
- São **assintóticamente equivalentes** [Engle, 1984].

ANOVA, MANOVA e testes de comparações múltiplas

Existe uma série de **procedimentos baseados em testes de hipóteses**, tais como:

- Análise de variância (**ANOVA**).
- Análise de variância multivariada (**MANOVA**).
- Testes de **comparações múltiplas**.

Temas abordados até aqui

- Ciência de dados.
- Modelos de regressão.
- Classes de modelos de regressão (ênfase nos **McGLMs**).
- Testes de hipóteses em modelos de regressão (ênfase no **teste Wald**).
- Procedimentos baseados em testes de hipóteses (ênfase na **ANOVA, MANOVA e testes de comparações múltiplas**).

Desafio e hipótese

- Não há discussão a respeito da **construção de testes de hipóteses para os McGLMs.**
- Contudo, os McGLMs apresentam os elementos necessários para utilizar o **teste Wald**:
  1. Um vetor de estimativas dos parâmetros
  2. Uma matriz de variância e covariância destas estimativas.
- Das três opções clássicas de testes de hipóteses, o teste Wald se torna o mais atrativo para os McGLMs.

Objetivo

1. Propor a utilização do **teste Wald** para realização de testes de hipóteses gerais sobre parâmetros de **regressão** e **dispersão** de **McGLMs**.

2. Avaliar as propriedades e comportamento dos testes propostos com base em **estudos de simulação**.

3. **Implementar** funções em R para testes de hipóteses, ANOVA, MANOVA e testes de comparações múltiplas para os McGLMs.

4. Motivar o potencial de aplicação das metodologias discutidas com base na **aplicação a conjuntos de dados reais**.

Contribuição

- Formas de **avaliar os parâmetros** estimados pelos McGLMs.
- Fornecer ferramentas para uma melhor **interpretação dos parâmetros** estimados, permitindo responder questões comuns no contexto de modelagem.
- **Extrair mais informações e conclusões** a respeito dos problemas modelados por meio dos McGLMs.

Referencial teórico

Referencial teórico

Referencial teórico

1. McGLMs
   - Elementos.
   - Preditores lineares e matriciais.
   - Funções de variância.
   - Parâmetros.
   - Estimação.

2. Testes de hipóteses
   - Elementos de um teste de hipóteses.
   - Testes de hipóteses em modelos de regressão.
   - Teste Wald
   - ANOVA e MANOVA.
   - Testes de comparações múltiplas.

Modelos multivariados de covariância linear generalizada

Modelos multivariados de covariância linear generalizada

Modelos multivariados de covariância linear generalizada

Para definição de um McGLM considere:

- $Y_{N\times R} = \{Y_1, \ldots, Y_R\}$ uma matriz de **variáveis resposta**.
- $M_{N\times R} = \{\mu_1, \ldots, \mu_R\}$ uma matriz de **valores esperados**.
- X_r denota uma **matriz de delineamento** $N \times k_r$.
- β_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, $r = 1,\ldots, R$, a **matriz de variância e covariância** para cada resposta r, de dimensão NxN:

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

Em que:

- $\mathbb{V}_r\left(\boldsymbol{\mu}; p\right)$ é uma matriz diagonal em que as entradas principais são dadas pela função de variância aplicada ao vetor $\boldsymbol{\mu}$.
- p_r é o parâmetro de potência.
- $\boldsymbol{\Omega}\left(\boldsymbol{\tau}_r\right)$ a matriz de dispersão que descreve a parte da covariância dentro de cada variável resposta.

Preditor linear matricial

- A matriz $\boldsymbol{\Omega}(\boldsymbol{\tau}_r)$ 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:

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

  - $h()$ é a função de ligação de covariância.
  - Z_{rd} com $d = 0, \ldots, D$ 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.

Funções de variância

1. **Função de variância potência** [Jørgensen, 1987, 1997].
   - Família **Tweedie** de distribuições.
   - $\vartheta(\boldsymbol{\mu}; p) = \mu^p$.
   - Casos particulares: normal (p = 0), Poisson (p = 1), gama (p = 2) e normal inversa (p = 3).
2. **Função de dispersão Poisson–Tweedie** [Jørgensen and Kokonendji, 2015].
   - Família **Poisson-Tweedie** de distribuições.
   - $\vartheta(\boldsymbol{\mu}; p) = \mu + \mu^p$.
   - 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.**
   - $\vartheta(\boldsymbol{\mu}) = \mu(1 - \mu)$.
   - Acomoda **respostas binárias** ou **restritas a um intervalo**.

Modelos multivariados de covariância linear generalizada}

Os McGLMs são definidos por:

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

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

Em que:

- $\boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b = \mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \ldots, \tilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b \otimes \boldsymbol{I})\mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^\top, \ldots, \tilde{\boldsymbol{\Sigma}}_R^\top)$ é o produto generalizado de Kronecker.
- $\tilde{\boldsymbol{\Sigma}}_r$ denota a matriz triangular inferior da decomposição de Cholesky da matriz $\boldsymbol{\Sigma}_r$.
- Bdiag() denota a matriz bloco-diagonal.
- $\boldsymbol{I}$ 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

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \boldsymbol{D}^\top \boldsymbol{C}^{-1}(\mathcal{Y} - \mathcal{M})$$

$$\psi_{\boldsymbol{\lambda}_i}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \text{tr}(W_{\boldsymbol{\lambda} i}(\boldsymbol{r}^\top \boldsymbol{r} - \boldsymbol{C})), i = 1,.., Q$$

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.
- $\mathcal{Y}$ 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.
- $\boldsymbol{D} = \nabla_{\boldsymbol{\beta}} \mathcal{M}$ é uma matriz $NR \times K$, e $\nabla_{\boldsymbol{\beta}}$ denota o operador gradiente.
- $W_{\boldsymbol{\lambda} i} = -\frac{\partial \boldsymbol{C}^{-1}}{\partial \boldsymbol{\lambda}_i}$
- $\boldsymbol{r} = (\mathcal{Y} - \mathcal{M})$

Distribuição assintótica e algoritmo de estimação

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

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

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

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

$J_{\boldsymbol{\theta}}^{-1}$ é a inversa da matriz de informação de Godambe, dada por

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

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

Testes de hipóteses

Testes de hipóteses

Testes de hipóteses

- Inferência: **tirar conclusões** 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 π um parâmetro, um teste de hipóteses sobre π é dado por:

$$\begin{cases} H_0 : \pi = \pi_0 \\ H_1 : \pi \neq \pi_0 \end{cases}$$

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

- 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 **valor-p**.

Testes de hipóteses em modelos de regressão

- **Testes de hipóteses são ferramentas gerais** que podem ser aplicadas em problemas de regressão.
- Modelos de regressão: modelar uma ou mais variáveis em função de um conjunto de variáveis explicativas.
- Modelos contêm parâmetros que são quantidades desconhecidas que estabelecem a relação entre as variáveis sob o modelo.
- Para avaliar os parâmetros de um modelo testes de hipóteses são usados.

Teste Wald em modelos de regressão

- Verificar se existe evidência para afirmar que um ou mais **parâmetros** são iguais a **valores especificados**.
- Avalia a **distância** entre as **estimativas** dos parâmetros e um conjunto de **valores postulados**.
- Esta **distância** é **padronizada** por medidas de precisão das estimativas dos parâmetros.
- Quanto maior for esta distância padronizada, menores são as evidências a favor da hipótese de que os valores estimados são iguais aos valores postulados.

Teste Wald em modelos de regressão

Considere um problema de regressão em que:

- β um vetor $k \times 1$ **parâmetros de regressão**.
- $\hat{\beta}$ as **estimativas** dos parâmetros.
- L uma **matriz de especificação das hipóteses**, de dimensão $s \times k$.
- c um vetor de **valores postulados** de dimensão s.

Teste Wald em modelos de regressão

- As hipóteses podem ser descritas como:

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

- A estatística de teste é dada por:

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

- WT $\sim \chi^2_s$.

ANOVA e MANOVA

- **Testes sucessivos impondo restrições** ao modelo original.
- Avaliar o efeito das variáveis explicativas.
- Caso univariado: **ANOVA**. Caso multivariado: **MANOVA**.
- O quadro de ANOVA ou MANOVA contém em cada linha:
  1. A variável.
  2. O valor da estatística de teste.
  3. Os graus de liberdade.
  4. Um valor-p.
- É possível gerar quadros de análise de variância por meio do teste Wald.

Testes de comparações múltiplas

- Usado quando a ANOVA aponta para **efeito significativo de uma variável categórica**.
- **Comparações aos pares** a fim de detectar para quais níveis da variável categórica os valores da resposta se alteram.
- Pode ser avaliada utilizando o teste Wald.
- Por meio da correta especificação da matriz L, é possível **avaliar hipóteses sobre qualquer possível contraste** entre os níveis de uma determinada variável categórica.

Teste Wald para McGLMs

Teste Wald para McGLMs

Teste Wald para McGLMs

1. Definição das hipóteses.
2. Estatística de teste.
3. Distribuição.
4. Hipóteses comuns.
5. Construção da matriz L.
6. Exemplos.

Hipóteses

$$H_0 : \boldsymbol{L}\boldsymbol{\theta}^* = \boldsymbol{c} \; vs \; H_1 : \boldsymbol{L}\boldsymbol{\theta}^* \neq \boldsymbol{c}.$$

Em que:

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

Estatística de teste

$$W = (L\hat{\theta}^* - \boldsymbol{c})^T \, (L \, J^{*-1} \, L^T)^{-1} \, (L\hat{\theta}^* - \boldsymbol{c}).$$

Em que:

- $\boldsymbol{L}$ é a matriz da especificação das hipóteses, tem dimensão $s \times h$.
- $\hat{\theta}^*$ é o vetor de dimensão $h \times 1$ com todas as estimativas dos parâmetros de regressão, dispersão e potência.
- $\boldsymbol{c}$ é um vetor de dimensão $s \times 1$ com os valores sob hipótese nula.
- E J^{*-1} é a inversa da matriz de informação de Godambe desconsiderando os parâmetros de correlação, de dimensão $h \times h$.
- $W \sim \chi^2_s$

Hipóteses comuns

- Costuma ser de interesse formular hipóteses:
  1. Para **parâmetros individuais**.
  2. Para **múltiplos parâmetros**.
  3. Para avaliar **igualdade de parâmetros**.
  4. Sobre parâmetros de regressão ou dispersão para **respostas sob mesmo preditor**.
  5. Sobre **contrastes**.
- O elemento chave é a correta especificação da matriz L.

Construção da matriz L

- Cada coluna da matriz L corresponde a um dos h parâmetros de θ^*.
- Cada linha corresponde a uma restrição.
- A matriz é composta por valores iguais a 0, 1 e eventualmente -1.
- O produto $L\theta^*$ deve resultar nas hipóteses de interesse.

Exemplo

Considere o problema:

- Investigar se duas variáveis numéricas (X_1 e X_2) possuem efeito sobre Y_1 e Y_2.
- Amostra com N observações, para cada observação obteve-se X_1, X_2, Y_1 e Y_2.
- Um McGLM bivariado, pode ter preditor dado por

$$g_r(\mu_r) = \beta_{r0} + \beta_{r1}X_1 + \beta_{r2}X_2 + \beta_{r3}X_1X_2.$$

- 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.
- β_{r2} um parâmetro de regressão associado a uma variável X_2.
- β_{r3} um parâmetro de regressão associado a interação entre X_1 e X_2.
- Apenas um parâmetro de dispersão τ_{r0}.
- Unidades independentes, logo $Z_0 = I$.
- Parâmetros de potência foram fixados.

Exemplo

$$g_r(\mu_r) = \beta_{r0} + \beta_{r1}X_1 + \beta_{r2}X_2 + \beta_{r3}X_1X_2.$$

Trata-se de um problema com:

- Duas variáveis resposta.
- Duas variáveis explicativas.
- Observações independentes.

Podem ser perguntas de interesse:

- Existe efeito da variável X_1 apenas sobre a primeira resposta?
- É possível que a variável X_1 possua efeito sobre as duas respostas ao mesmo tempo?
- É possível que o efeito da variável seja o mesmo para ambas as respostas?

Exemplo: hipótese para múltiplos parâmetros

- Efeito da variável explicativa X_1 sobre ambas as respostas.
- Hipóteses:

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

- De forma equivalente

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

Exemplo: hipótese para múltiplos parâmetros

Notação do teste Wald:

$$H_0 : \boldsymbol{L}\boldsymbol{\theta}^* = \boldsymbol{c} \; vs \; H_1 : \boldsymbol{L}\boldsymbol{\theta}^* \neq \boldsymbol{c}.$$

- $\boldsymbol{\theta}^{*T} = \left[\beta_{10} \; \beta_{11} \; \beta_{20} \; \beta_{21} \; \tau_{11} \; \tau_{21}\right]$.
- $\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$.
- $\boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

ANOVA, MANOVA e testes de comparações múltiplas

- Propomos 3 tipos diferentes de análises de variância usando o teste Wald, nomeadas tipo I, II e III.
- Cada linha do quadro corresponde uma hipótese. Portanto, basta especificar uma matriz L.
- As ANOVAs retornam **um quadro para cada resposta**.
- As MANOVAs retornam **um único quadro**.
- Também chegamos a procedimentos para avaliação aos pares de níveis de variáveis categóricas.

ANOVA e MANOVA tipo II

$$g_r(\mu_r) = \beta_{r0} + \beta_{r1}X_1 + \beta_{r2}X_2 + \beta_{r3}X_1X_2.$$

Linha 1 Testa se o intercepto é igual a 0.

Linha 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.

Linha 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.

Linha 4 Testa se o efeito de interação é 0.

Estudo de simulação

Estudo de simulação

Estudo de simulação

1. Visão geral.
2. Parâmetros de regressão.
3. Parâmetros de dispersão.

Visão geral

- Avaliar o **poder do teste Wald** em testes de hipóteses sobre parâmetros de McGLMs.
- Comportamento da proposta para três distribuições de probabilidade:
  - **Normal.**
  - **Poisson.**
  - **Bernoulli.**
- Cenários **univariados** e **trivariados**.
- Diferentes tamanhos amostrais.

Visão geral

- O procedimento consistiu em:
  - **Simular conjuntos de dados** com valores dos parâmetros definidos.
  - **Ajustar modelos** com os dados simulados.
  - **Obter as estimativas** (espera-se que sejam próximas dos valores definidos).
  - **Testas hipóteses** usando as estimativas.

- A ideia é verificar o que ocorre quando a **hipótese se afasta dos valores dos parâmetros**.

- Espera-se que no primeiro ponto haja um percentual de rejeição baixo.

- Para os demais pontos espera-se que o percentual de rejeição aumente gradativamente.

Visão geral

- Correlação entre respostas no caso trivariado:

$$\Sigma_b = \begin{bmatrix} 1 & 0{,}75 & 0{,}5 \\ 0{,}75 & 1 & 0{,}25 \\ 0{,}5 & 0{,}25 & 1 \end{bmatrix}$$

- Funções de ligação e variância utilizadas nos modelos para cada distribuição simulada.

Distribuição	Função de ligação	Função de variância
Normal	Identidade	Constante
Poisson	Logarítmica	Tweedie
Bernoulli	Logito	Binomial

Visão geral

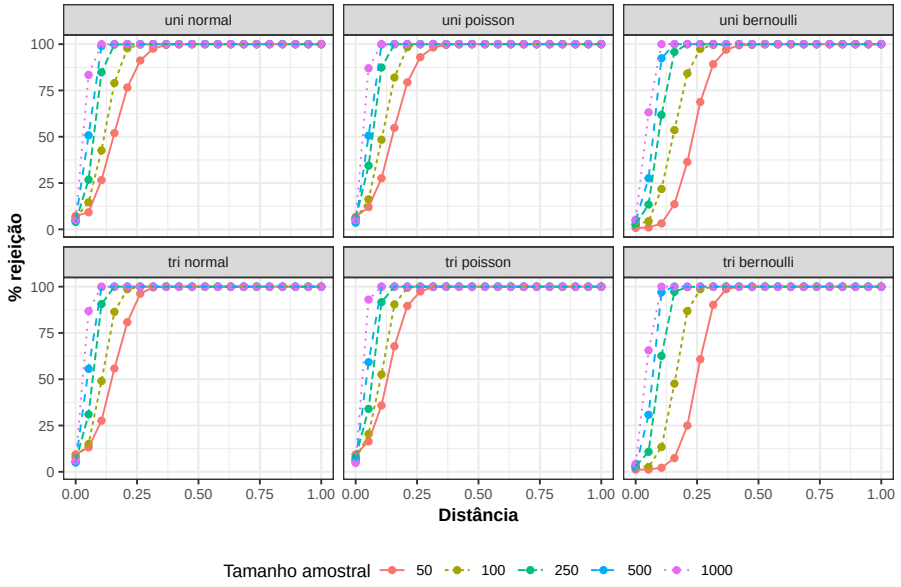
- Estudo realizado na linguagem **R**.
- **Conjuntos de dados univariados**: bibliotecas padrões do R.
- **Conjuntos de dados multivariados**:
  - Biblioteca *mvtnorm* [Genz et al., 2021], [Genz and Bretz, 2009] para distribuição Normal.
  - Método NORTA [Cario and Nelson, 1997], biblioteca *NORTARA* [Su, 2014] para distribuição Poisson e Bernoulli.

Parâmetros de regressão

- **5 tamanhos amostrais**: 50, 100, 250, 500 e 1000.
- **500** conjuntos de dados.
- Uma variável **explicativa categórica de 4 níveis**.
- Parâmetros por distribuição:
  - Normal: $\beta_0 = 5$, $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$. CV de 20%.
  - Poisson: $\beta_0 = 2{,}3$, $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$. Contagems próximas de 10.
  - Bernoulli: $\beta_0 = 0{,}5$, $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$. p aproximadamente 0,6.
- Cenários **univariados** e **trivariados**.
- Para cada amostra gerada foi ajustado um McGLM.

Parâmetros de regressão

- 20 diferentes hipóteses:
  - Decréscimo em β_0.
  - Distribuição deste decréscimo nos demais βs da hipótese.
- Decréscimo por distribuição:
  - Normal: 0,15.
  - Poisson: 0,05.
  - Bernoulli: 0,25.
- Percentual de rejeição da hipótese nula por ponto.

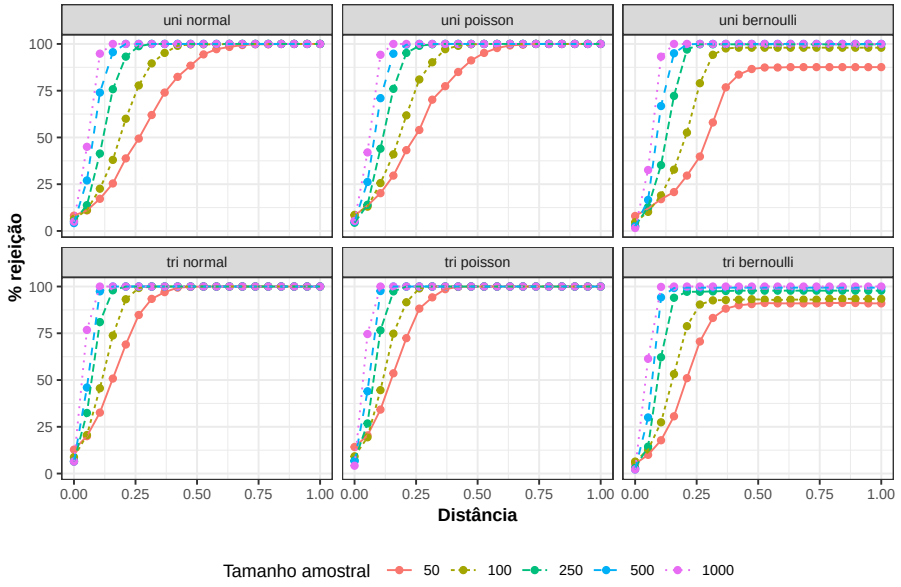

uni normal
uni poisson
uni bernoulli
tri normal
tri poisson
tri bernoulli
% rejeição
0
25
50
75
100
0.00
0.25
0.50
0.75
1.00
Distância
Tamanho amostral
50
100
250
500
1000

Parâmetros de dispersão

- **5 tamanhos amostrais**: 50, 100, 250, 500 e 1000.
- **500** conjuntos de dados.
- Cada unidade fornece 5 medidas ao conjunto de dados.
- Sem variáveis explicativas.
- Parâmetros de dispersão: $\tau_0 = 1$ e $\tau_1 = 0$.
- Parâmetros por distribuição:
  - Normal: Média 5, desvio padrão 1.
  - Poisson: Taxa igual a 10.
  - Bernoulli: Probabilidade igual 0,6.

Parâmetros de dispersão

- Cenários **univariados** e **trivariados**.
- Para cada amostra gerada foi ajustado um McGLM.
- 20 diferentes hipóteses.
  - Decréscimo de 0,02 em τ_0.
  - Acréscimo de 0,02 em τ_1.
- Percentual de rejeição da hipótese nula por ponto.

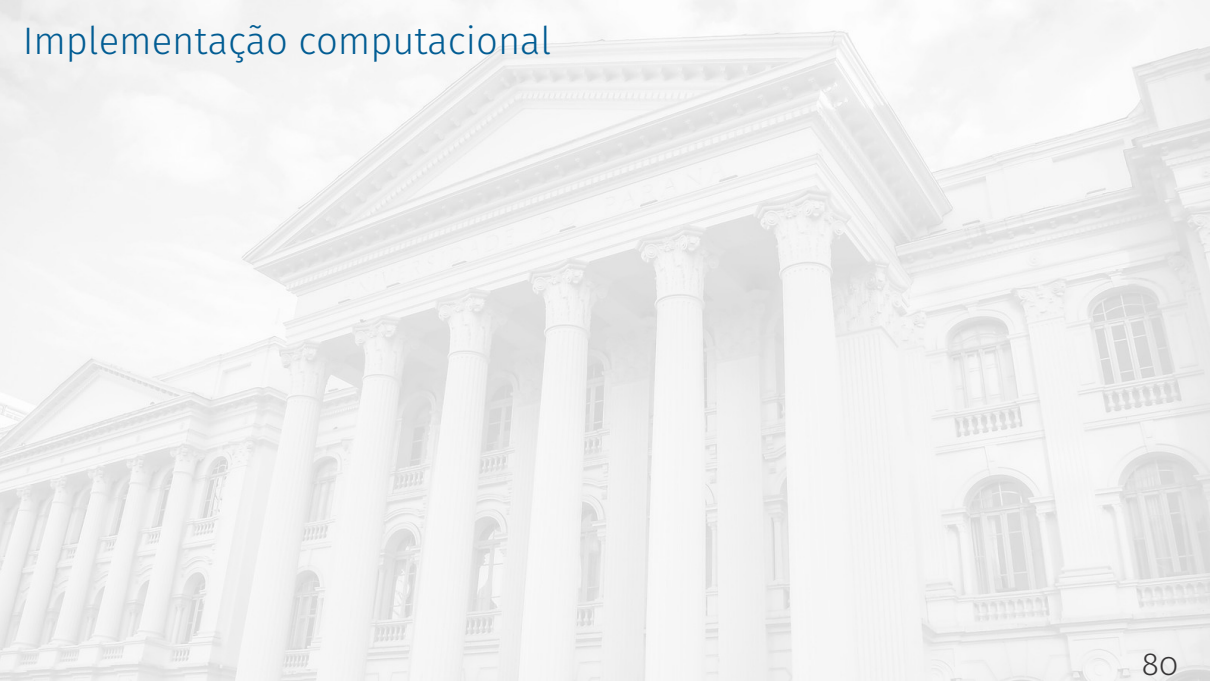

uni normal
uni poisson
uni bernoulli
tri normal
tri poisson
tri bernoulli
% rejeição
0
25
50
75
100
0.00
0.25
0.50
0.75
1.00
Distância
Tamanho amostral
50
100
250
500
1000

Resultados

- Quanto mais distante a hipótese é dos valores inicialmente simulados, maior é o percentual de rejeição.
- Os menores percentuais foram observados na hipótese igual aos valores simulados.
- Conforme aumenta-se o tamanho amostral, o percentual de rejeição aumenta para hipóteses pouco diferentes dos valores simulados dos parâmetros.

Implementação computacional

Implementação computacional



Implementação computacional

- **Implementar e disponibilizar** publicamente os testes apresentados.
- Complementar as já possíveis análises permitidas pelo pacote *mcglm* [Bonat, 2018].
- O pacote *htmcglm* já está disponível no **Comprehensive R Archive Network (CRAN)**.
- As funções implementadas geram resultados mostrando graus de liberdade e valores-p baseados no teste Wald aplicado a um McGLM.

Implementação computacional

Função	Descrição
mc_anova_I()	ANOVA tipo I
mc_anova_II()	ANOVA tipo II
mc_anova_III()	ANOVA tipo III
mc_manova_I()	MANOVA tipo I
mc_manova_II()	MANOVA tipo II
mc_manova_III()	MANOVA tipo III
mc_anova_dispersion()	ANOVA tipo III para dispersão
mc_manova_dispersion()	MANOVA tipo III para dispersão
mc_multcomp()	Testes de comparações múltiplas por resposta
mc_mult_multcomp()	Testes de comparações múltiplas multivariado
mc_linear_hypothesis()	Hipóteses lineares gerais especificadas pelo usuário

Análise de dados

Análise de dados

Análise de dados

1. Contexto.
2. Análise exploratória.
3. Especificação do modelo.
4. Resultados do ajuste.
5. Testes de hipóteses.

Contexto

- Avaliar se o **uso de probióticos** é capaz de controlar **vício** e **transtorno da compulsão alimentar** em pacientes submetidos à **cirurgia bariátrica**.
- Conjunto de indivíduos divididos em 2 grupos: **placebo** ou **tratamento**.
- Indivíduos avaliados ao longo do tempo:
  - **T0**: primeira avaliação, **antes** da cirurgia.
  - **T1**: segunda avaliação, **3 meses** após a cirurgia.
  - **T2**: terceira avaliação, **1 ano** após a cirurgia.

Contexto

- Avaliou-se os níveis de **vício** e **transtorno da compulsão alimentar** nos pacientes.
- Compulsão alimentar foi feita com base na **escala de compulsão alimentar (BES)**: escore, varia de 0 a 46.
- Vício alimentar foi utilizada a **escala de vício alimentar (YFAS)**: número de sintomas de vício, varia de 0 a 7.
- Para fins de análise, **YFAS** e **BES** foram transformados para a **escala unitária**, considerando que tratam-se de **variáveis restritas**.

Contexto

- **71 indivíduos** (33 grupo placebo, 38 grupo tratamento). **184 observações**.
- **Duas variáveis resposta**: BES e YFAS.
- As observações **não** são **independentes**.
- Técnicas de modelagem tradicionais seriam de difícil aplicação.
- Cenário ideal para resolução via McGLM.
- Testes de hipóteses podem ser empregados para avaliar o efeito da interação entre momento e uso do probiótico sobre vício e compulsão alimentar.

Contexto

Variáveis:

- **id**: identificadora de indivíduo.
- **momento**: identificadora de momento (T0, T1, T2).
- **grupo**: identificadora de grupo (placebo, probiótico)
- **YFAS**: vício na escala unitária.
- **BES**: compulsão na escala unitária.

Análise exploratória

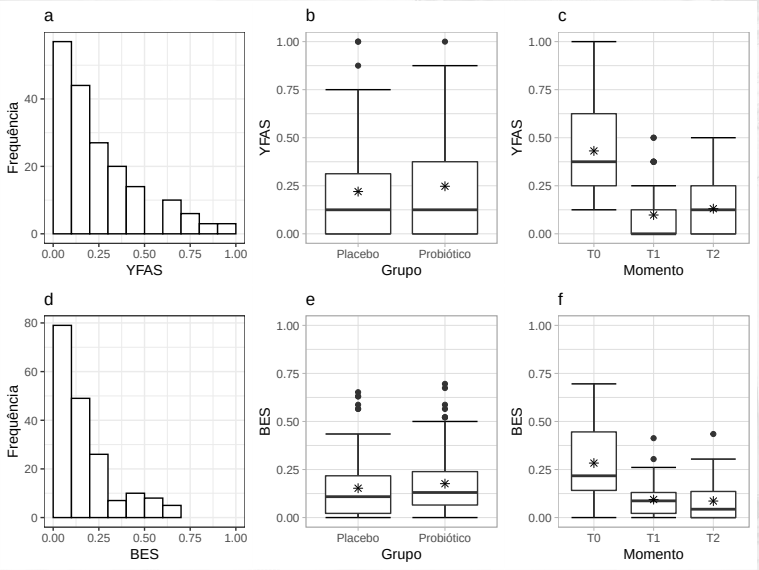


Figura 1. Análise exploratória gráfica: (a) histograma YFAS, (b) boxplots YFAS em função de grupo, (c) boxplots YFAS em função de momento, (d) histograma BES, (b) boxplots BES em função de grupo, (c) boxplots BES em função de momento. O asterísco nos boxplots indica a média.

Especificação do modelo

- Respostas foram transformadas para a **escala unitária**.
- Função de ligação **logito**.
- Função de variância **binomial**.
- Parâmetro de potência foram estimados.
- Foram consideradas categorias de referência o grupo placebo e o momento To.

Preditores lineares

$$g_r(\mu_r) = \beta_{0r} + \beta_{1r} T1 + \beta_{2r} T2 + \beta_{3r} Probiotico + \beta_{4r} T1 * Probiotico + \beta_{5r} T2 * Probiotico$$

- r: variáveis respostas (1 para YFAS, 2 para BES).
- β_{0r}: intercepto.
- β_{1r}: efeito do momento T1.
- β_{2r}: efeito do momento T2.
- β_{3r}: o efeito de probiótico.
- β_{4r}: efeito da interação entre T1 e probiótico.
- β_{5r}: efeito da interação entre T2 e probiótico.

Preditores matriciais

- Iguais para ambas as respostas.
- A função $h(.)$ utilizada foi a identidade.

$$h\left\{\boldsymbol{\Omega}(\boldsymbol{\tau})\right\} = \tau_0 Z_0 + \tau_1 Z_1$$

  - τ_0 e τ_1: parâmetros de dispersão.
  - Z_0: matriz identidade 184×184.
  - Z_1: matriz 184×184 especificada de forma a explicitar que as **medidas provenientes do mesmo indivíduo são correlacionadas.**

Análise de resíduos

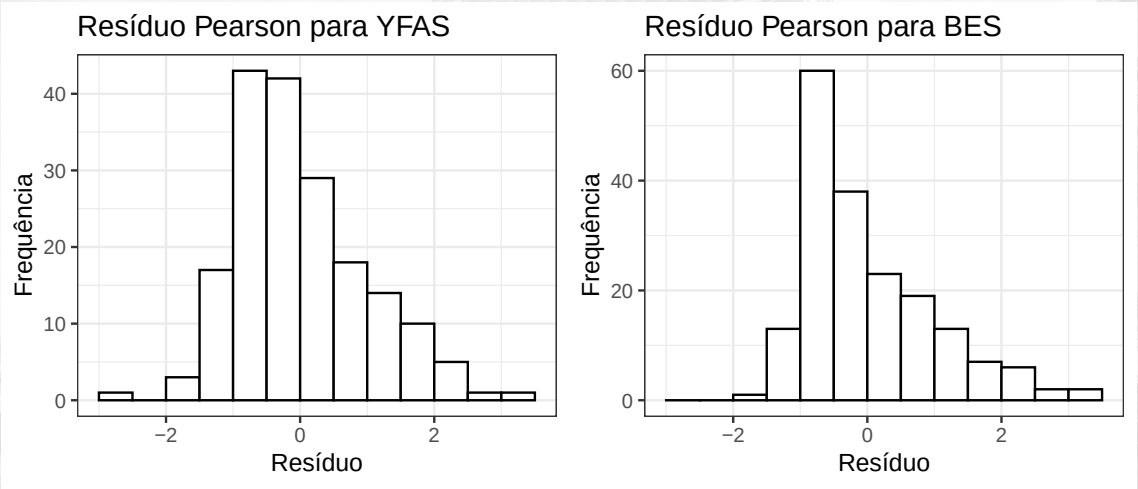


Figura 2. Histograma dos resíduos de Pearson por resposta.

Análise de resíduos

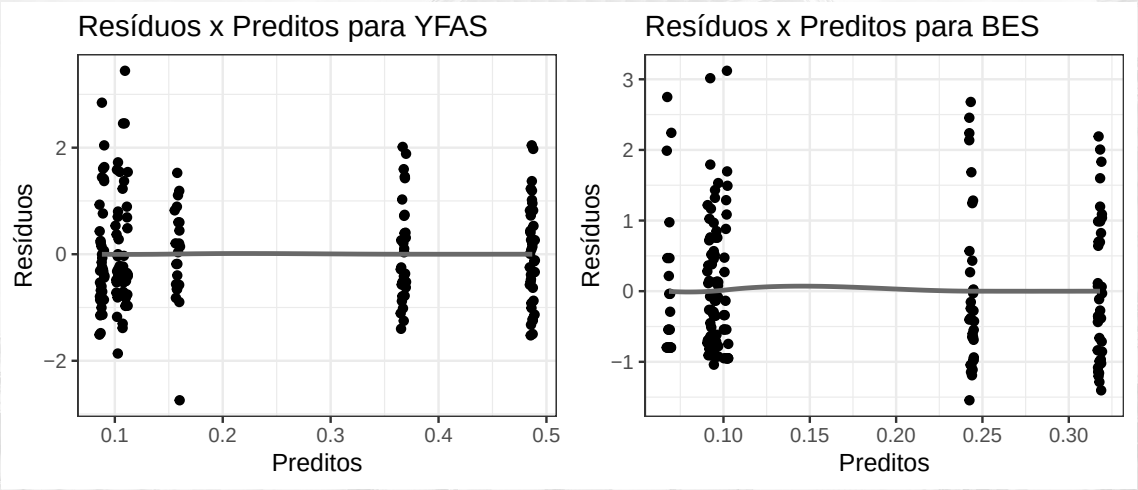


Figura 3. Gráfico de resíduos Pearson versus preditos com linha de tendência suave para cada resposta.

Análise de resíduos

- Resíduos de Pearson para YFAS e BES apresentam **média 0** e **desvio padrão próximo de 1**.
- Histogramas dos resíduos de Pearson: distribuição aproximadamente simétrica com a maior parte dos dados entre -2 e 2.
- Gráficos de resíduos versos preditos: mostram que não parece haver qualquer tipo de relação entre resíduos e preditos.
- De forma geral, o modelo parece estar razoavelmente bem ajustado aos dados.

Estimativas

Parâmetro	YFAS			BES		
	Estimativa	Intervalo de confiança	Valor-p	Estimativa	Intervalo de confiança	Valor-p
β_0	-0,54	(-0,87;-0,22)	<0,01	-1,13	(-1,44;-0,83)	<0,01
β_1	-1,55	(-2,17;-0,94)	<0,01	-1,16	(-1,62;-0,69)	<0,01
β_2	-1,13	(-1,75;-0,51)	<0,01	-1,05	(-1,58;-0,52)	<0,01
β_3	0,49	(0,05;0,93)	0,0284	0,37	(-0,03;0,77)	0,0733
β_4	-0,73	(-1,60;0,14)	0,0	-0,33	(-0,96;0,30)	0,3081
β_5	-0,98	(-1,93;-0,03)	0,0429	-0,80	(-1,58;-0,02)	0,0449
τ_0	0,18	(0,01;0,35)	0,0411	0,17	(0,00;0,34)	0,0458
τ_1	0,01	(-0,02;0,04)	0,5718	0,04	(-0,01;0,10)	0,1357
p	0,91	(0,47;1,34)	<0,05	1,23	(0,77;1,68)	<0,05

Tabela 1. Estimativas dos parâmetros, intervalos com 95% de confiança e valores-p do modelo.

Preditos

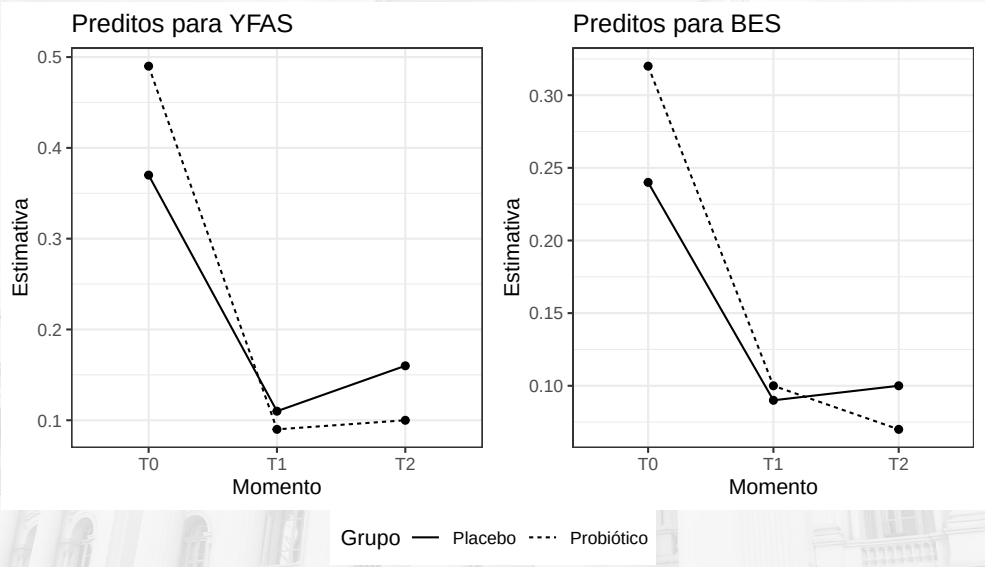


Figura 4. Gráfico de preditos pelo modelo para cada combinação entre momento e grupo.

Testes de hipóteses

- Até este ponto foi apresentada uma análise padrão.
- Podemos fazer uso dos testes de hipóteses para explorar o modelo.
  - Análise de variância multivariada do tipo II.
  - Análise de variância univariada do tipo II.
  - Comparações duas a duas entre momentos.
  - Comparações entre grupos para cada momento.
  - Análise da correlação de medidas tomadas em um mesmo indivíduo.

MANOVA tipo II

Variável	Graus de liberdade	Qui-quadrado	Valor-p
Intercepto	2	53,1581	<0,01
Momento	8	139,0161	<0,01
Grupo	6	8,4928	0,2042
Momento*Grupo	4	6,9923	0,1363

Tabela 2. Análise de variância multivariada do tipo II.

Existência do efeito de momento e ausência de efeito de grupo.

ANOVA tipo II

		YFAS		BES	
Variável	GL	Qui-quadrado	Valor-p	Qui-quadrado	Valor-p
Intercepto	1	10,6128	<0,01	53,1473	<0,01
Momento	4	102,9875	<0,01	99,5681	<0,01
Grupo	3	6,6837	0,0827	5,3083	0,1506
Momento*Grupo	2	5,5984	0,0609	4,2477	0,1196

Tabela 3. Análise de variância univariada do tipo II.

Existência do efeito de momento e ausência de efeito de grupo.

Comparações múltiplas

Contraste	Graus de liberdade	Qui-quadrado	Valor-p
T0-T1	2	97,9874	<0,01
T0-T2	2	67,2462	<0,01
T1-T2	2	2,4730	0,8712

Tabela 4. Comparações duas a duas entre momentos para ambas as respostas.

- $T0 \neq T1$.
- $T0 \neq T2$.
- $T1 = T2$.

Comparações múltiplas

Contraste	Graus de liberdade	Qui-quadrado	Valor-p
T0:Placebo-T0:Probiótico	2	5,5819	0,9204
T1:Placebo-T1:Probiótico	2	0,6096	1
T2:Placebo-T2:Probiótico	2	1,7645	1

Tabela 5. Comparações duas a duas entre grupos para cada momento para ambas as respostas.

Ausência de diferença entre grupos em cada momento.

MANOVA para dispersão

Variável	Graus de liberdade	Qui-quadrado	Valor-p
τ_0	2	7,1936	0,0274
τ_1	2	2,3201	0,3135

Tabela 6. Análise de variância multivariada do tipo III para parâmetros de dispersão.

Não há evidência para crer que as medidas tomadas em um mesmo indivíduo apresentam correlação.

Considerações finais

Considerações finais

Considerações finais

- Objetivo: teste **Wald** em **McGLMs**.
- Chegamos a procedimentos para:
  - **Testes de hipóteses lineares gerais**.
  - **ANOVA**.
  - **MANOVA**.
  - **Testes de comparações múltiplas**.
- **Proposta implementada** na biblioteca R *htmcglm*, complementa a biblioteca *mcglm*.

Conclusões gerais

- As propriedades dos testes foram avaliadas com base em **estudos de simulação**.
- Os estudos de simulação mostraram que:
  - Quanto mais distante a hipótese é dos valores inicialmente simulados, maior é o percentual de rejeição.
  - Os menores percentuais foram observados na hipótese igual aos valores simulados.
  - Conforme aumenta-se o tamanho amostral, o percentual de rejeição aumenta para hipóteses pouco diferentes dos valores simulados dos parâmetros.

Conclusões gerais

- Aplicação: efeito do uso de **probióticos** no controle de **vícios e compulsões alimentares**.

- Os resultados, baseados nos testes propostos indicam que:
  - Existe efeito de momento.
  - Existem diferenças entre o primeiro versus segundo e primeiro versus terceiro momento, mas os dois últimos momentos não diferem entre si.
  - Ausência de diferença entre grupos em cada momento.
  - Não há evidência para crer que as medidas tomadas em um mesmo indivíduo apresentam correlação.

Limitações

Casos não explorados nos estudos de simulação, tais como:

- Desempenho dos testes ao definir hipóteses lineares que combinem **parâmetros de diferentes tipos**.
- Impacto de um **número diferente de observações** por indivíduos em problemas longitudinais ou de medidas repetidas.
- Impacto no poder do teste conforme o **número de parâmetros testados** aumenta.
- Comportamento do teste em problemas multivariados com **distribuições de probabilidade diferentes** das exploradas.

Trabalhos futuros

Avaliação de parâmetros de McGLMs para um melhor entendimento do impacto dos elementos em problemas de modelagem.

Algumas possibilidades são:

- Correções de valores-p de acordo com o tamanho das hipóteses testadas.
- Procedimentos além do teste Wald (como o teste Escore e o teste da razão de verossimilhanças).
- Outros procedimentos para comparações múltiplas.
- Lidar com contrastes alternativos aos usuais.
- Explorar procedimentos para seleção covariáveis.

Obrigado!

Lineu Alberto Cavazani de Freitas

lineuacf@gmail.com

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

PPG Informática

John Aitchison and SD Silvey. Maximum-likelihood estimation of parameters subject to restraints. *The annals of mathematical Statistics*, pages 813–828, 1958.

Wagner Hugo Bonat. Multiple response variables regression models in R: The mcglm package. *Journal of Statistical Software*, 84(4):1–30, 2018. doi: 10.18637/jss.v084.i04.

Wagner Hugo Bonat and Bent Jørgensen. Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(5): 649–675, 2016.

Marne C Cario and Barry L Nelson. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Citeseer, 1997.

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

Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

Alan Genz and Frank Bretz. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg, 2009. ISBN 978-3-642-01688-2.

Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2021. URL https://CRAN.R-project.org/package=mvtnorm. R package version 1.1-3.

Bent Jørgensen. Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(2):127–145, 1987.

Bent Jørgensen. *The theory of dispersion models*. CRC Press, 1997.

Bent Jørgensen and Célestin C Kokonendji. Discrete dispersion models and their tweedie asymptotics. *AStA Advances in Statistical Analysis*, 100(1):43–78, 2015.

John Ashworth Nelder and Robert William Maclagan Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135:370–384, 1972.

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

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

Po Su. *NORTARA: Generation of Multivariate Data with Arbitrary Marginals*, 2014. URL https://CRAN.R-project.org/package=NORTARA. R package version 1.0.0.

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

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

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