Modelo normal linear multivariado: Parte 1

Prof. Caio Azevedo

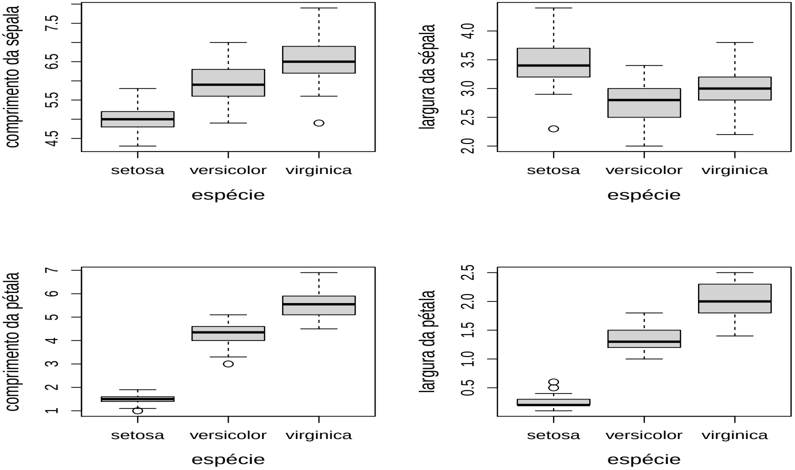
Recapitulando: dados das Iris de Fisher (veja)

- Os dados consistem de 50 unidades amostrais de três espécies (setosa, virginica, versicolor) de íris (uma espécie de planta), ou seja, temos um total de 150 unidades amostrais.
- De cada uma delas mediu-se quatro variáveis: comprimento e largura da sépala (CS, LS) e comprimento e largura da pétala (CP,LP).
- Objetivo original: quantificar a variação morfológica em relação à essas espécies com bases nas quatro variáveis de interesse.

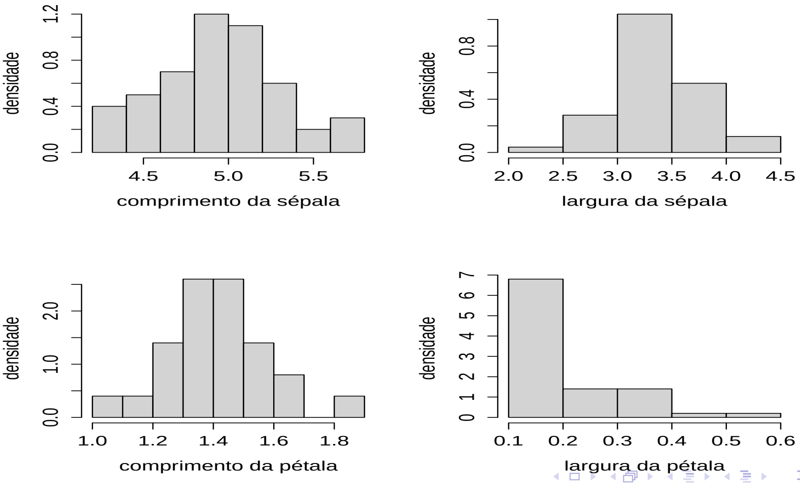
Modelagem dos dados

- Seja Y_{ijk} : o valor da k-ésima variável (k=1,2,3,4), para a j-ésima unidade experimental (planta) (j=1,...,50) do i-ésimo grupo ($i = 1, 2, 3$).
- Suposição $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3}, Y_{ij4}) \overset{ind.}{\sim} N_4(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$.
- Tem-se, portanto, dados balanceados (em relação aos grupos) e homocedasticidade (mesma matriz de variâncias e covariâncias).

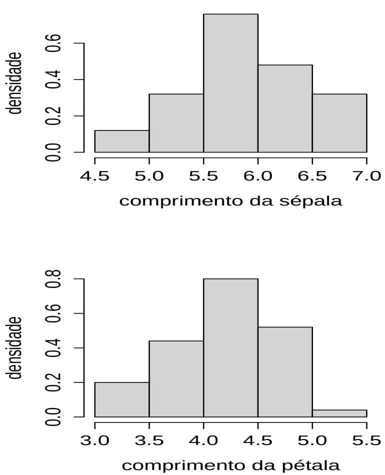
Boxplot das variáveis por grupo



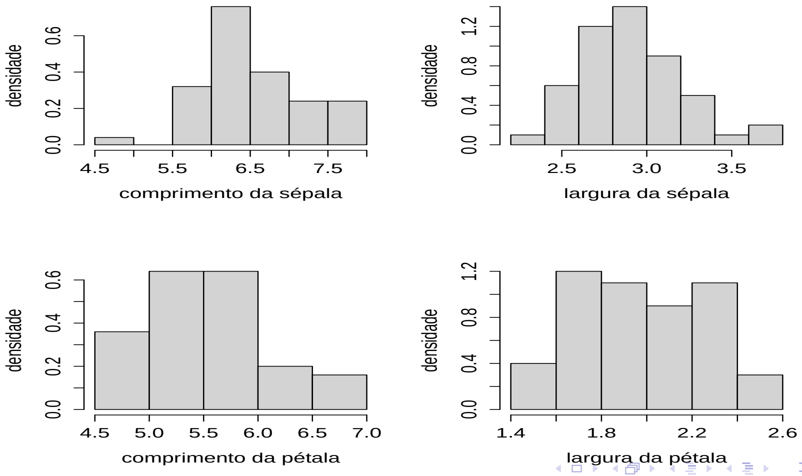
Histograma das variáveis: iris setosa



Histograma das variáveis: iris versicolor



Histograma das variáveis: iris virginica



Matriz de dispersão: ■ - S, ● - Ver, ▲ - Vir

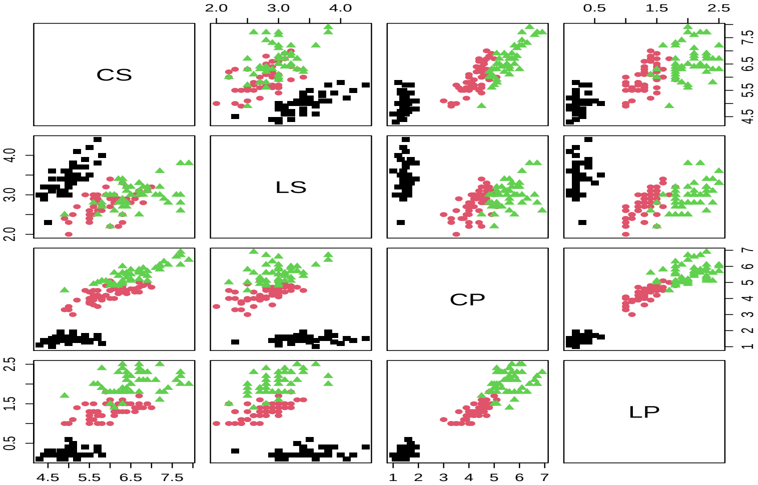


Diagrama de dispersão: setosa

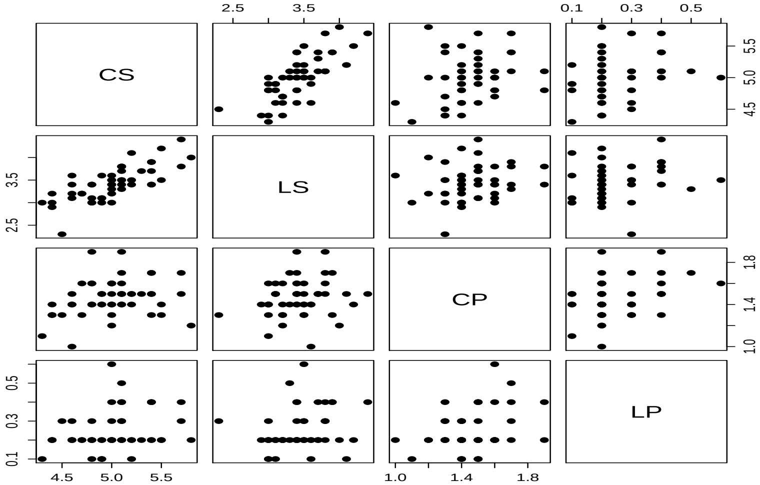


Diagrama de dispersão: versicolor

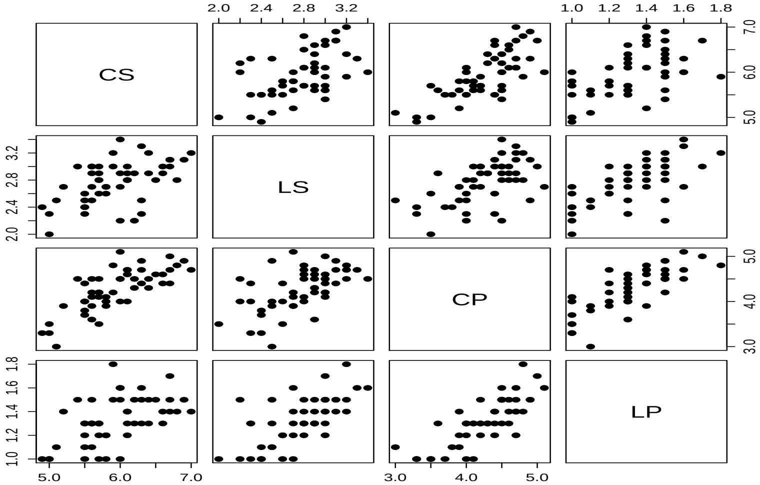
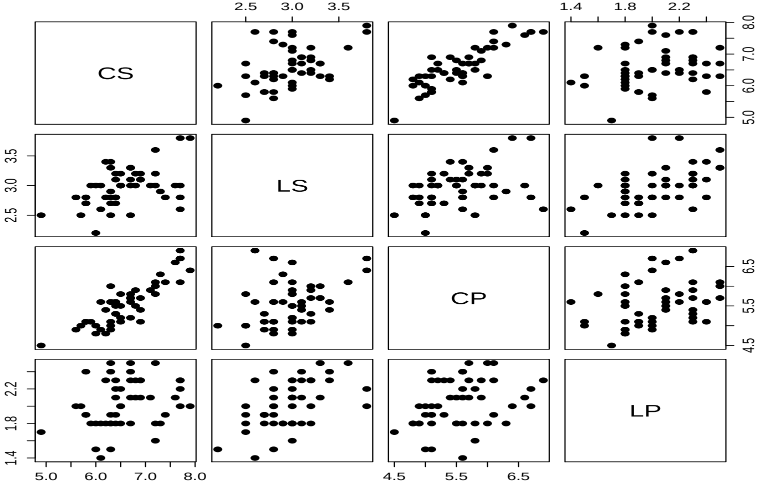


Diagrama de dispersão: virginica



Variâncias na diagonal principal (dp), covariâncias abaixo e correlações acima (da dp)

$$\widetilde{\Psi}_{\text{setosa}} = \begin{array}{c} CS \\ LS \\ CP \\ LP \end{array} \begin{bmatrix} 0,12 & 0,74 & 0,27 & 0,28 \\ 0,10 & 0,14 & 0,18 & 0,23 \\ 0,02 & 0,01 & 0,03 & 0,33 \\ 0,01 & 0,01 & 0,01 & 0,01 \end{bmatrix}$$

$$\widetilde{\Psi}_{\text{versicolor}} = \begin{array}{c} CS \\ LS \\ CP \\ LP \end{array} \begin{bmatrix} 0,27 & 0,53 & 0,75 & 0,55 \\ 0,08 & 0,10 & 0,56 & 0,66 \\ 0,18 & 0,08 & 0,22 & 0,79 \\ 0,06 & 0,04 & 0,07 & 0,04 \end{bmatrix}$$

$$\widetilde{\Psi}_{\text{virginica}} = \begin{array}{c} CS \\ LS \\ CP \\ LP \end{array} \begin{bmatrix} 0,40 & 0,46 & 0,86 & 0,28 \\ 0,09 & 0,10 & 0,40 & 0,54 \\ 0,30 & 0,07 & 0,30 & 0,32 \\ 0,05 & 0,05 & 0,05 & 0,07 \end{bmatrix}$$

Medidas resumo: comprimento da sépala

Espécie	Média	DP	Var	CV(%)	Mín.	Med.	Máx.	CA.	Curt.	n
setosa	5,01	0,35	0,12	7,04	4,30	5,00	5,80	0,11	-0,45	50
versicolor	5,94	0,52	0,27	8,70	4,90	5,90	7,00	0,10	-0,69	50
virginica	6,59	0,64	0,40	9,65	4,90	6,50	7,90	0,11	-0,20	50

Medidas resumo: largura da sépala

Espécie	Média	DP	Var	CV(%)	Mín.	Med.	Máx.	CA.	Curt.	n
setosa	3,43	0,38	0,14	11,06	2,30	3,40	4,40	0,04	0,60	50
versicolor	2,77	0,31	0,10	11,33	2,00	2,80	3,40	-0,34	-0,55	50
virginica	2,97	0,32	0,10	10,84	2,20	3,00	3,80	0,34	0,38	50

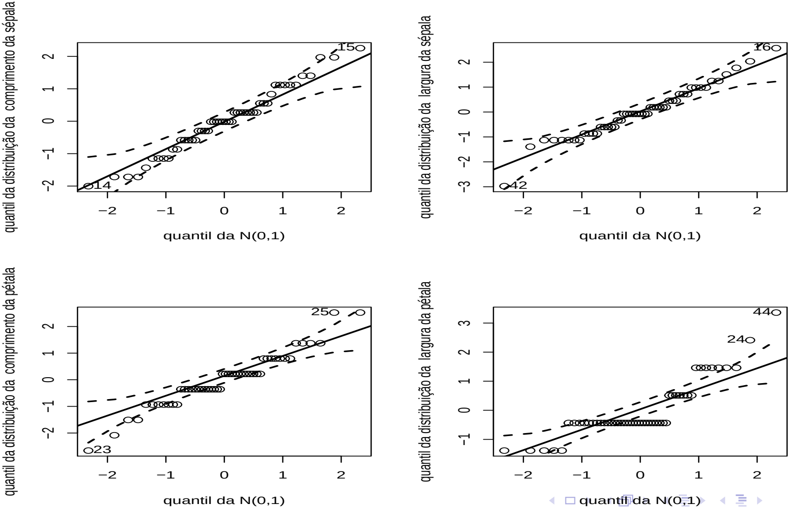
Medidas resumo: comprimento da pétala

Espécie	Média	DP	Var	CV(%)	Mín.	Med.	Máx.	CA.	Curt.	n
setosa	1,46	0,17	0,03	11,88	1,00	1,50	1,90	0,10	0,65	50
versicolor	4,26	0,47	0,22	11,03	3,00	4,35	5,10	-0,57	-0,19	50
virginica	5,55	0,55	0,30	9,94	4,50	5,55	6,90	0,52	-0,37	50

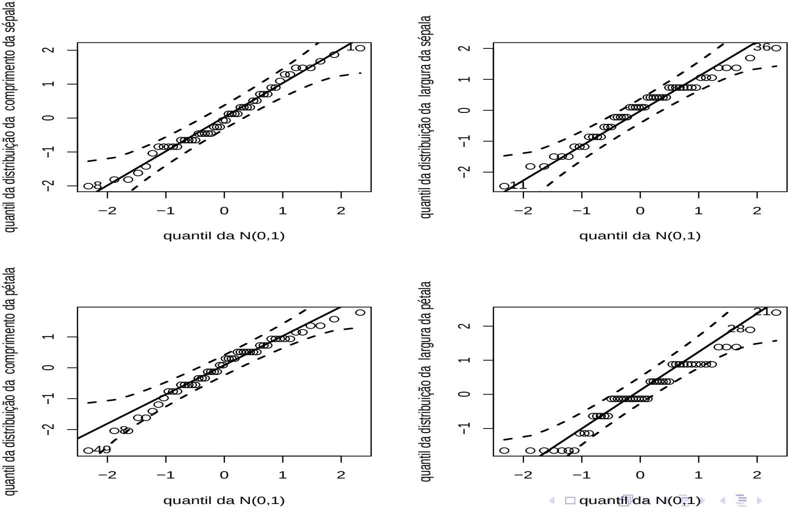
Medidas resumo: largura da pétala

Espécie	Média	DP	Var	CV(%)	Mín.	Med.	Máx.	CA.	Curt.	n
setosa	0,25	0,11	0,01	42,84	0,10	0,20	0,60	1,18	1,26	50
versicolor	1,33	0,20	0,04	14,91	1,00	1,30	1,80	-0,03	-0,59	50
virginica	2,03	0,27	0,08	13,56	1,40	2,00	2,50	-0,12	-0,75	50

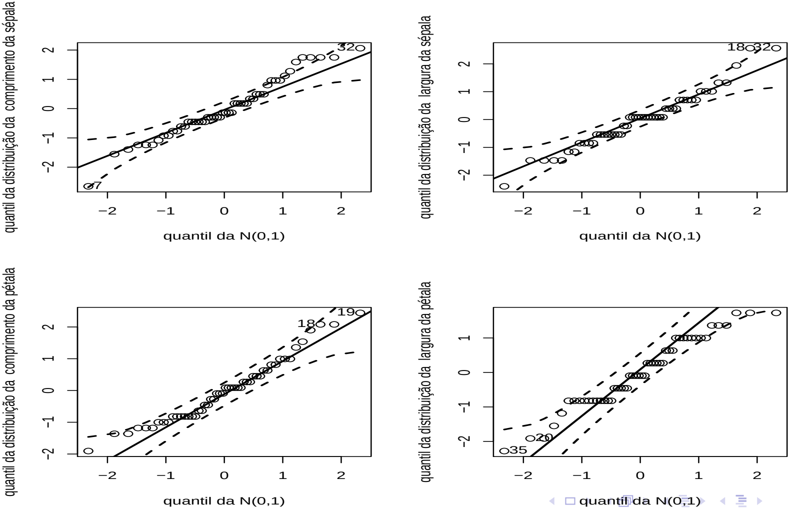
Gráficos de quantis-quantis com envelopes: setosa



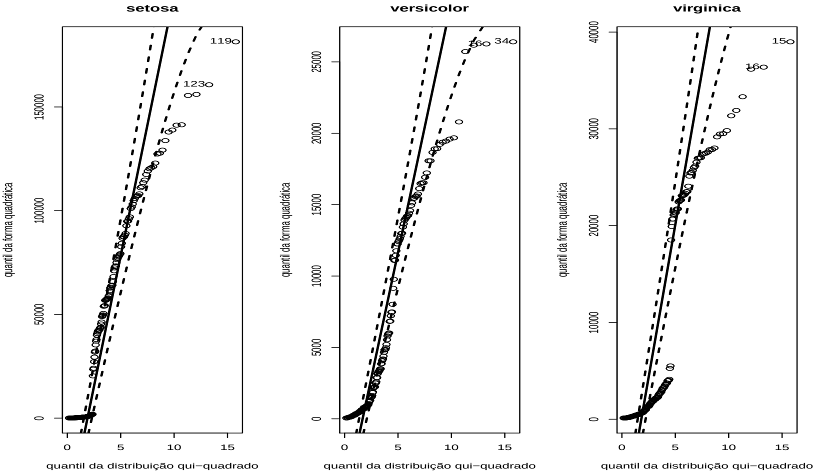
Gráficos de quantis-quantis com envelopes: versicolor



Gráficos de quantis-quantis com envelopes: virginica



Gráficos de Q-Q com envelopes: forma quadrática



Matriz de dados: iris de Fisher

$$\boldsymbol{Y}_{(n \times p)} = \begin{bmatrix} Y_{111} & Y_{112} & Y_{113} & Y_{114} \\ Y_{121} & Y_{122} & Y_{123} & Y_{124} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{1(50)1} & Y_{1(50)2} & Y_{1(50)3} & Y_{1(50)4} \\ ---&---&---&--- \\ Y_{211} & Y_{212} & Y_{213} & Y_{214} \\ Y_{221} & Y_{222} & Y_{223} & Y_{224} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{2(50)1} & Y_{2(50)2} & Y_{2(50)3} & Y_{2(50)4} \\ ---&---&---&--- \\ Y_{311} & Y_{312} & Y_{313} & Y_{314} \\ Y_{321} & Y_{322} & Y_{323} & Y_{324} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{3(50)1} & Y_{3(50)2} & Y_{3(50)3} & Y_{3(50)4} \end{bmatrix}$$

Suposições

- Suponha um conjunto de G populações independentes da qual retiramos G amostras de tamanho n_i, $i = 1, ..., G$,
- Por suposição, temos que $\boldsymbol{Y}_{ij} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, em que i = 1, 2,...,G (grupo) e j = 1,2,...,n_i (indivíduo). Notação: Y_{ijk} observação referente à variável k do indivíduo j do grupo i.
- Homocedasticidades: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = ... = \boldsymbol{\Sigma}_G = \boldsymbol{\Sigma}$.
- Assim, temos a seguinte matriz de dados ($n = \sum_{i=1}^{G} n_i$):

$$
\boldsymbol{Y}_{(n\times p)} = \begin{bmatrix}
Y_{111} & Y_{112} & \ldots & Y_{11p} \\
Y_{121} & Y_{122} & \ldots & Y_{12p} \\
\vdots & \vdots & \ddots & \vdots \\
Y_{1n_1 1} & Y_{1n_1 2} & \ldots & Y_{1n_1 p} \\
--- & --- & --- & --- \\
Y_{211} & Y_{212} & \ldots & Y_{21p} \\
Y_{221} & Y_{222} & \ldots & Y_{22p} \\
\vdots & \vdots & \ddots & \vdots \\
Y_{2n_2 1} & Y_{2n_2 2} & \ldots & Y_{2n_2 p} \\
--- & --- & --- & --- \\
\vdots & \vdots & \ddots & \vdots \\
--- & --- & --- & --- \\
Y_{G11} & Y_{G12} & \ldots & Y_{G1p} \\
Y_{G21} & Y_{G22} & \ldots & Y_{G2p} \\
\vdots & \vdots & \ddots & \vdots \\
Y_{Gn_G 1} & Y_{Gn_G 2} & \ldots & Y_{Gn_G p}
\end{bmatrix}
$$

Hipóteses de interesse

- Queremos testar $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_G$ vs H_1 : há pelo menos uma diferença.
- Uma abordagem: análise de variância multivariada (MANOVA).
- Comparar médias atráves do estudo da decomposição da matriz de variâncias-covariâncias total.
- Como resumir a informação das matrizes de covariâncias de interesse? Uma possibilidade: variâncias generalizadas.
- À rigor, a metodologia MANOVA está acomodada na classe de modelos de regressão normais lineares multivariados, a qual é uma extensão dos respectivos modelos univariados (link).

Modelo de regressão normal linear multivariado (MRNLM)

$$\boldsymbol{Y}_{(n \times p)} = \boldsymbol{X}_{(n \times q)} \boldsymbol{B}_{(q \times p)} + \boldsymbol{\xi}_{(n \times p)}$$

- $\boldsymbol{Y}_{(n \times p)}$: matriz de dados
- $\boldsymbol{X}_{(n \times q)}$: matriz de planejamento, conhecida e não-aleatória.
- $\boldsymbol{B}_{(q \times p)}$: parâmetros de regressão, desconhecido e não aleatório.
- $\boldsymbol{\xi}_{(n \times p)}$: matriz de erros, sendo que cada uma das suas linhas corresponde a uma vetor aleatório tal que $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, mutuamente independentes.
- Tal estrutura se aplica a diferentes situaçãoes, não somente em situação como a dos dados da iris de Fisher.
- Com efeito, qualquer modelo que possa ser escrito sob estrutura, é um MRNLM.

MRNLM (cont.)

- As suposições do MRNLN são:
  - Em relação aos erros associados a cada unidade amostral/experimental (linhas da matriz de dados):
    - Normalidade multivariada.
    - Matriz de variâncias e covariâncias constante ao longo das observações, conhecida como homocedasticidade.
    - Independência entre elas.
  - Linearidade entre a média $\mathcal{E}(\boldsymbol{Y})$ e as covariáveis $\boldsymbol{X}$, ou seja $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{XB}$.
- Em geral o MRNLM não é robusto a violação de pelo menos uma das suposições descritas acima (artigo 1, artigo 2)

Matriz de planejamento de matriz de parâm. de regressão

$$
\boldsymbol{X}_{(n \times q)} = \begin{bmatrix}
X_{111} & X_{112} & \ldots & X_{11G} \\
X_{121} & X_{122} & \ldots & X_{12G} \\
\vdots & \vdots & \vdots & \vdots \\
X_{1n_1 1} & X_{1n_1 2} & \ldots & X_{1n_1 G} \\
--- & --- & --- & --- \\
\vdots & \vdots & \vdots & \vdots \\
--- & --- & --- & --- \\
X_{G11} & X_{G12} & \ldots & X_{G1G} \\
X_{G21} & X_{G22} & \ldots & X_{G2G} \\
\vdots & \vdots & \vdots & \vdots \\
X_{Gn_G 1} & X_{Gn_G 2} & \ldots & X_{Gn_G G}
\end{bmatrix} ; \boldsymbol{B}_{(q \times p)} = \begin{bmatrix}
\mu_{11} & \mu_{12} & \ldots & \mu_{1p} \\
\mu_{21} & \mu_{22} & \ldots & \mu_{2p} \\
\vdots & \vdots & \vdots & \vdots \\
\mu_{G1} & \mu_{G2} & \ldots & \mu_{Gp}
\end{bmatrix}
$$

$$
\boldsymbol{\xi}_{(n\times p)} = \begin{bmatrix}
\xi_{111} & \xi_{112} & \cdots & \xi_{11p} \\
\xi_{121} & \xi_{122} & \cdots & \xi_{12p} \\
\vdots & \vdots & \ddots & \vdots \\
\xi_{1n_1 1} & \xi_{1n_1 2} & \cdots & \xi_{1n_1 p} \\
--- & --- & --- & --- \\
\xi_{211} & \xi_{212} & \cdots & \xi_{21p} \\
\xi_{221} & \xi_{222} & \cdots & \xi_{22p} \\
\vdots & \vdots & \ddots & \vdots \\
\xi_{2n_2 1} & \xi_{2n_2 2} & \cdots & \xi_{2n_2 p} \\
--- & --- & --- & --- \\
\vdots & \vdots & \ddots & \vdots \\
--- & --- & --- & --- \\
\xi_{G11} & \xi_{G12} & \cdots & \xi_{G1p} \\
\xi_{G21} & \xi_{G22} & \cdots & \xi_{G2p} \\
\vdots & \vdots & \ddots & \vdots \\
\xi_{Gn_G 1} & \xi_{Gn_G 2} & \cdots & \xi_{Gn_G p}
\end{bmatrix}
$$

Voltando aos dados da Iris de Fisher:

- G $= 3$, p $= 4$ e $n_i = 50, i = 1, 2, 3$ (3 grupos, 4 variáveis e 50 indivíduos por grupo).
- Objetivo: modelar o vetor de médias de cada grupo e compará-los entre si.
- $\boldsymbol{Y}_{(150 \times 4)}$.

Cont.: Parametrização de médias

- $\boldsymbol{X} = \boldsymbol{I}_{(3\times 3)} \otimes \mathbf{1}_{(50\times 1)}$ ($\otimes$ denota o produto de Kronecker à esquerda). $\boldsymbol{X} = \begin{bmatrix} \mathbf{1}_{(50\times 1)} & \mathbf{0}_{(50\times 1)} & \mathbf{0}_{(50\times 1)} \\ \mathbf{0}_{(50\times 1)} & \mathbf{1}_{(50\times 1)} & \mathbf{0}_{(50\times 1)} \\ \mathbf{0}_{(50\times 1)} & \mathbf{0}_{(50\times 1)} & \mathbf{1}_{(50\times 1)} \end{bmatrix}$
- $\boldsymbol{B} = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} \\ \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} \\ \mu_{31} & \mu_{32} & \mu_{33} & \mu_{34} \end{bmatrix}$
- μ_{ik} : média da variável k do grupo i.

Cont.: Parametrização casela de referência

- $$\boldsymbol{X} = \begin{bmatrix} \mathbf{1}_{(50\times1)} & \mathbf{0}_{(50\times1)} & \mathbf{0}_{(50\times1)} \\ \mathbf{1}_{(50\times1)} & \mathbf{1}_{(50\times1)} & \mathbf{0}_{(50\times1)} \\ \mathbf{1}_{(50\times1)} & \mathbf{0}_{(50\times1)} & \mathbf{1}_{(50\times1)} \end{bmatrix}$$
- $$\boldsymbol{B} = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \end{bmatrix}$$
- μ_{1k} : média da variável k do grupo 1 (grupo de referência - setosa).
- $\alpha_{ik} = \mu_{ik} - \mu_{1k}, i = 2, 3$: é a diferença entre a média do grupo i e a média do grupo 1, em relação a variável k.

Decomposição da matriz de covariâncias total

- Pode-se demonstrar, para o exemplo em questão, que:

$$
\begin{aligned}
\underbrace{\sum_{i=1}^{G}\sum_{j=1}^{n_i}\left(\boldsymbol{Y}_{ij}-\overline{\boldsymbol{Y}}\right)\left(\boldsymbol{Y}_{ij}-\overline{\boldsymbol{Y}}\right)'}_{\text{Matriz de SQ Total}} &= \underbrace{\sum_{i=1}^{G} n_i\left(\overline{\boldsymbol{Y}}_i-\overline{\boldsymbol{Y}}\right)\left(\overline{\boldsymbol{Y}}_i-\overline{\boldsymbol{Y}}\right)'}_{\text{Matriz de SQ do Modelo}} \\
&+ \underbrace{\sum_{i=1}^{G}\sum_{j=1}^{n_i}\left(\boldsymbol{Y}_{ij}-\overline{\boldsymbol{Y}}_i\right)\left(\boldsymbol{Y}_{ij}-\overline{\boldsymbol{Y}}_i\right)'}_{\text{Matriz de SQ do Erro}} \\
\boldsymbol{T} &= \boldsymbol{M}+\boldsymbol{E}
\end{aligned}
$$

- De uma forma geral: $\boldsymbol{T} = \boldsymbol{Y}'(\boldsymbol{I} - n^{-1}\boldsymbol{J})\boldsymbol{Y}$; $\boldsymbol{M} = \boldsymbol{Y}'\left(\boldsymbol{H} - n^{-1}\boldsymbol{J}\right)\boldsymbol{Y}$; $\boldsymbol{E} = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$, $\boldsymbol{J} = \boldsymbol{1}\boldsymbol{1}'$, $\boldsymbol{I}$ é uma matriz identidade de ordem n e $\boldsymbol{1}$ é um vetor coluna unitário de dimensão n.

Variância generalizada

- Seja $\mathbf{\Sigma}_{(p \times p)}$ uma matriz de covariâncias.
- Variância generalizada $|\mathbf{\Sigma}|$ (resume a informação contida em $\mathbf{\Sigma}$).
- Suponha p = 2.
- Assim $|\mathbf{\Sigma}| = \sigma_1^2\sigma_2^2 - \sigma_{11}^2 = \sigma_1^2\sigma_2^2 - \sigma_1^2\sigma_2^2\rho^2 = \sigma_1^2\sigma_2^2(1-\rho^2)$.
- Estamos supondo que $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = ... = \mathbf{\Sigma}_G = \mathbf{\Sigma}$ (teste de Box para igualdade de matrizes de covariâncias).

As quatro estatísticas "tradicionais"

- Lembre-se de que $\boldsymbol{T} = \boldsymbol{M} + \boldsymbol{E}$.
- Sejam $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_s$ os autovalores diferentes de zero da matriz $\boldsymbol{E}^{-1}\boldsymbol{M}$, em que $s = \min(p, G-1)$.
- Lambda de Wilks: $\displaystyle\prod_{i=1}^{s} \frac{1}{1+\lambda_i} = \frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{M}|}$.
- Traço de Pillai: $\displaystyle\sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i} = tr[\boldsymbol{M}(\boldsymbol{M}+\boldsymbol{E})^{-1}]$.
- Traço de Lawley-Hotelling: $\displaystyle\sum_{i=1}^{s} \lambda_i^{-1} = tr[\boldsymbol{M}\boldsymbol{E}^{-1}]$.
- Máxima raiz de Roy: $\dfrac{\lambda_1}{1+\lambda_1}$.

Cont.

- Quanto menor o valor das estatística de Wilks e maior os valores das estatísticas de Pillai, Lawley-Hotelling e de Roy, mais evidências tem-se contra H_0.
- Existem aproximações pela distribuição F, para cada uma destas estatísticas (link manual do SAS, livro: Multivariate Statistics: High-Dimensional and Large-Sample Approximations).

Aplicação

- Dados da iris: as quatro variáveis apresentadas e os três grupos.
- O teste de Box indica a rejeição da igualdade entre as matrizes de covariâncias (estatística $= 140{,}94$; pvalor $< 0{,}0001$).
- Além da heterocedasticidade, a suposição de normalidade não parece ser razoável. Assim, os resultados obtidos podem não ser confiáveis.
- Utilização do pacote *manova* (manual) implementado na linguagem R (próximos slides).

Aplicação

- Outras opções de ajuste de modelos linares multivariados no R:
  - Função apresentada no Apêndice do Livro: An R Companion to Applied Regression.
  - Outras metodologias (funções) apresentadas em livro: An R Companion to Applied Regression.
  - Funções do pacote car, como por exemplo aqui.
  - Veja também o livro: Modern Multivariate Statistical Techniques Regression, Classification, and Manifold Learning.

Estatísticas calculadas e p-valores

Estatística	Valor	Aproxim. pela dist. F.	p-valor
Wilks	0,02	191,15	$< 0{,}0001$
Pillai	1,19	53,46	$< 0{,}0001$
Hotelling-Lawley	32,47	580,53	$< 0{,}0001$
Roy	32,19	1167,00	$< 0{,}0001$

A igualdade simultânea dos vetores de médias é rejeitada, de acordo com cada uma das estatísticas. Como realizar outras comparações de interesse?

Anovas univariadas

CS

FV	g.l.	SQ	QM	Estat. F	p-valor
Espécie	2	63,21	31,61	119,26	$<$ 0,0001
Resíduos	147	38,96	0,27		

LS

FV	g.l.	SQ	QM	Estat. F	p-valor
Espécie	2	11,34	5,67	49,16	$<$ 0,0001
Resíduos	147	16,96	0,12		

Anovas univariadas

CP

FV	g.l.	SQ	QM	Estat. F	p-valor
Espécie	2	437,10	218,55	1180,16	$<$ 0,0001
Resíduos	147	27,22	0,19		

LP

FV	g.l.	SQ	QM	Estat. F	p-valor
Espécie	2	80,41	40,21	960,01	$<$ 0,0001
Resíduos	147	6,16	0,04		

Comentários

- Realizar anovas univariadas, desconsidera a estruturas de correlação existente.
- Além disso, não permitem testar hipóteses envolvendo mais de uma variável, de modo simultâneo.
- Uma vez que se decide realizar, desde o início, análises multivariadas, é melhor considerar o MRNLM.
- Em geral, as hipóteses de interesse em MRNLM podem ser descritas como combinações lineares dos parâmetros $\boldsymbol{B}$.

Forma vetorial (MLNM)

- Considere novamente o modelo: $\boldsymbol{Y}_{(n \times p)} = \boldsymbol{X}_{(n \times q)} \boldsymbol{B}_{(q \times p)} + \boldsymbol{\xi}_{(n \times p)}$.

  Portanto, temos que (propriedades de matrizes):

$$
\begin{aligned}
\boldsymbol{Y}'_{(p \times n)} &= \boldsymbol{B}'_{(p \times q)} \boldsymbol{X}'_{(q \times n)} + \boldsymbol{\xi}'_{(p \times n)} \\
vec(\boldsymbol{Y}') &= \left(\boldsymbol{X}_{(n \times q)} \otimes \boldsymbol{I}_p\right) vec(\boldsymbol{B}') + vec(\boldsymbol{\xi}') \\
\boldsymbol{Y}^*_{(np \times 1)} &= \boldsymbol{X}^*_{(np \times pq)} \boldsymbol{\beta}_{(pq \times 1)} + \boldsymbol{\xi}^*_{(np \times 1)}
\end{aligned}
$$

  pois $vec(\boldsymbol{ABC}) = (\boldsymbol{C}' \otimes \boldsymbol{A}) vec(\boldsymbol{B})$.

- Note, assim, que as observações dos indivíduos foram concatenadas (uma abaixo da outra), nos vetores $\boldsymbol{Y}^*$ e $\boldsymbol{\xi}^*$.

Cont.

- Portanto (propriedades da DNM), temos que $\boldsymbol{Y}^* \sim N_{pn}(\boldsymbol{X}^*\boldsymbol{\beta}, \boldsymbol{\Sigma}^*)$, em que $\boldsymbol{\Sigma}^* = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}_{(p \times p)}$.
- O estimador de mínimos quadrados generalizados de $\boldsymbol{\beta}$ é obtido minimizando-se

$$(\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta})'\boldsymbol{\Sigma}^{*-1}(\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}).$$

- O que implica que $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{*'}\boldsymbol{\Sigma}^{*-1}\boldsymbol{X}^*\right)^{-1}\boldsymbol{X}^{*'}\boldsymbol{\Sigma}^{*-1}\boldsymbol{Y}^*$.
- Exercício: obtenha o estimador de MQO de $\boldsymbol{B}$.

Cont.

- Note que

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left[(\boldsymbol{X} \otimes \boldsymbol{I})' (\boldsymbol{I} \otimes \boldsymbol{\Sigma})^{-1} (\boldsymbol{X} \otimes \boldsymbol{I})\right]^{-1} (\boldsymbol{X} \otimes \boldsymbol{I})' (\boldsymbol{I} \otimes \boldsymbol{\Sigma})^{-1} \boldsymbol{Y}^* \\
&= \left[(\boldsymbol{X}'\boldsymbol{X} \otimes \boldsymbol{\Sigma}^{-1})\right]^{-1} \left[\boldsymbol{X}' \otimes \boldsymbol{\Sigma}^{-1}\right] \boldsymbol{Y}^* \\
&= \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}' \otimes \boldsymbol{I}\right] \boldsymbol{Y}^* \\
&= \boldsymbol{A}\boldsymbol{Y}^* \qquad (1)
\end{aligned}
$$

- Por outro lado, temos que

$$\mathcal{E}(\widehat{\boldsymbol{\beta}}) = \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}' \otimes \boldsymbol{I}\right] [\boldsymbol{X} \otimes \boldsymbol{I}] \boldsymbol{\beta} = \boldsymbol{\beta}.$$

- Além disso,

$$Cov(\widehat{\boldsymbol{\beta}}) = \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}' \otimes \boldsymbol{I}\right] [\boldsymbol{I} \otimes \boldsymbol{\Sigma}] \left[\boldsymbol{X} (\boldsymbol{X}'\boldsymbol{X})^{-1} \otimes \boldsymbol{I}\right] = (\boldsymbol{X}'\boldsymbol{X})^{-1} \otimes \boldsymbol{\Sigma}$$

Cont.

- Logo, $\widehat{\boldsymbol{\beta}} \sim N_{pq}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, em que $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \otimes \boldsymbol{\Sigma}$.
- Portanto, se $\boldsymbol{\Sigma}$ for conhecido (e consequentemente, o for $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$), então

$$\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' (\boldsymbol{\Sigma}_{\boldsymbol{\beta}})^{-1} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \chi^2_{(qp)}$$

- Por outro lado, de modo semelhante ao caso univariado, a grande maioria das hipóteses de interesse, podem ser escritas na forma

$$H_0 : \boldsymbol{C}_{(r\times q)}\boldsymbol{B}_{(q\times p)}\boldsymbol{U}_{(p\times s)} = \boldsymbol{M}_{(r\times s)} \text{ vs } H_1 : \boldsymbol{CBU} \neq \boldsymbol{M}$$

em que $r \leq q$ e $s \leq p$.

Cont.

- Note agora que $H_0 : \boldsymbol{U}'\boldsymbol{B}'\boldsymbol{C}' = \boldsymbol{M}'$ e, assim, temos que:

  $vec(\boldsymbol{U}'\boldsymbol{B}'\boldsymbol{C}') - \underbrace{vec(\boldsymbol{M}')}_{\boldsymbol{M}^*} = \underbrace{(\boldsymbol{C} \otimes \boldsymbol{U}')}_{\boldsymbol{C}^*}\boldsymbol{\beta} - \boldsymbol{M}^* = \boldsymbol{C}^*\boldsymbol{\beta} - \boldsymbol{M}^*.$

- Logo, podemos escrever as hipótese anteriores como:

  $H_0 : \boldsymbol{C}^*\boldsymbol{\beta} - \boldsymbol{M}^* = \boldsymbol{0}$ vs $H_1 : \boldsymbol{C}^*\boldsymbol{\beta} - \boldsymbol{M}^* \neq \boldsymbol{0}.$

- Além disso, $\widehat{\boldsymbol{\theta}} = \boldsymbol{C}^*\widehat{\boldsymbol{\beta}} - \boldsymbol{M}^* \sim N_{rs}\left(\boldsymbol{C}^*\boldsymbol{\beta} - \boldsymbol{M}, \boldsymbol{C}^*\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{C}^{*'}\right).$

- Se $\boldsymbol{\Sigma}_\beta$ for conhecido, temos que, sob H_0

$$\widehat{\boldsymbol{\theta}}'\left(\boldsymbol{C}^*\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{C}^{*'}\right)^{-1}\widehat{\boldsymbol{\theta}} \sim \chi^2_{(rs)}$$

Cont.

- Seja, então, $\widehat{\boldsymbol{\Sigma}}$, um estimador consistente de $\boldsymbol{\Sigma}$ $\left(\widehat{\boldsymbol{\Sigma}} \xrightarrow[n\to\infty]{P} \boldsymbol{\Sigma}\right)$. Logo $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \otimes \widehat{\boldsymbol{\Sigma}}$ o será para $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$.
- $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-q}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}\right)$, em que $\widehat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.
- Assim, pelo Teorema de Slutsky,

$$Q = \widehat{\boldsymbol{\theta}}'\left(\boldsymbol{C}^*\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\boldsymbol{C}^{*'}\right)^{-1}\widehat{\boldsymbol{\theta}} \xrightarrow[n\to\infty]{D} \chi^2_{(rs,\delta)},$$

em que

$$\delta = (\boldsymbol{C}^*\beta - \boldsymbol{M}^*)'\left(\boldsymbol{C}^*\boldsymbol{\Sigma}_{\beta}\boldsymbol{C}^{*'}\right)^{-1}(\boldsymbol{C}^*\beta - \boldsymbol{M}^*).$$

- Sob H_0 tem-se que $\delta = 0$.

Cont.

- Nível descritivo: $P(Q > q_{calc}|H_0)$, em que $Q \approx \chi^2_{(rs)}$, para n suficientemente grande e q_{calc} é o valor calculado da estatística Q.
- Função poder: $P(Q > q_c|H_1, \alpha)$, em que $Q \approx \chi^2_{(rs)}$ para n suficientemente grande e q_c é o valor crítico para um dado α (nível de significância).
- Assim, o poder estimado do teste é dado por: $P(\widetilde{Q} > q_c|H_1, \alpha)$, em que $\widetilde{Q} \approx \chi^2_{(rs,\widetilde{\delta})}$ para n suficientemente grande, q_c é o valor crítico para um dado α (nível de significância), $\widetilde{\delta} = \left(\boldsymbol{C}^*\widetilde{\boldsymbol{\beta}} - \boldsymbol{M}^*\right)' \left(\boldsymbol{C}^*\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\boldsymbol{C}^{*'}\right)^{-1} \left(\boldsymbol{C}^*\widetilde{\boldsymbol{\beta}} - \boldsymbol{M}^*\right)$ e "$\widetilde{\ }$" representa a respectiva estimativa.

Cont.

- Voltando ao exemplo da íris, se quisermos testar se as médias entre os grupos, em relação à variável CS (que é a variável 1), são simultâneamente iguais, teríamos (parametrização de médias):

$$\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

$$\boldsymbol{U} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}; \boldsymbol{M} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}; \boldsymbol{CBU} = \begin{bmatrix} \mu_{11} - \mu_{21} \\ \mu_{11} - \mu_{31} \end{bmatrix}$$

Cont.

- Voltando ao exemplo da íris, se quisermos testar se as médias entre os grupos, em relação à variável CS (que é a variável 1), são simultâneamente iguais, teríamos (parametrização casela de referências): $\boldsymbol{C} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

$$\boldsymbol{U} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}; \boldsymbol{M} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}; \boldsymbol{CBU} = \begin{bmatrix} \alpha_{21} \\ \alpha_{31} \end{bmatrix}$$

Comparações múltiplas

- Comparações de interesse (para $k = 1, 2, 3, 4$)

$$H_0 : \mu_{1k} = \mu_{2k} \;\; vs \;\; H_1 : \mu_{1k} \neq \mu_{2k}$$

$$H_0 : \mu_{1k} = \mu_{3k} \;\; vs \;\; H_1 : \mu_{1k} \neq \mu_{3k}$$

$$H_0 : \mu_{2k} = \mu_{3k} \;\; vs \;\; H_1 : \mu_{2k} \neq \mu_{3k}$$

Resultados

- Estatística do teste e p-valor entre parênteses.
- Variável CS:
  - Setosa x Versicolor: 81,59 ($< 0{,}0001$).
  - Setosa x Virginica: 236,10 ($< 0{,}0001$).
  - Versicolor x Virgínica: 40,10 ($< 0{,}0001$).
- Variável LS:
  - Setosa x Versicolor: 93,81 ($< 0{,}0001$).
  - Setosa x Virginica: 44,60 ($< 0{,}0001$).
  - Versicolor x Virgínica: 9,02 (0,0027).

Resultados

- Estatística do teste e p-valor entre parênteses.
- Variável CP:
  - Setosa x Versicolor: 1056,87 ($<$ 0,0001).
  - Setosa x Virginica: 2258,26 ($<$ 0,0001).
  - Versicolor x Virgínica: 225,35 ($<$ 0,0001).
- Variável LP:
  - Setosa x Versicolor: 696,25 ($<$ 0,0001).
  - Setosa x Virginica: 1891,28 ($<$ 0,0001).
  - Versicolor x Virgínica: 292,49 ($<$0,0001).

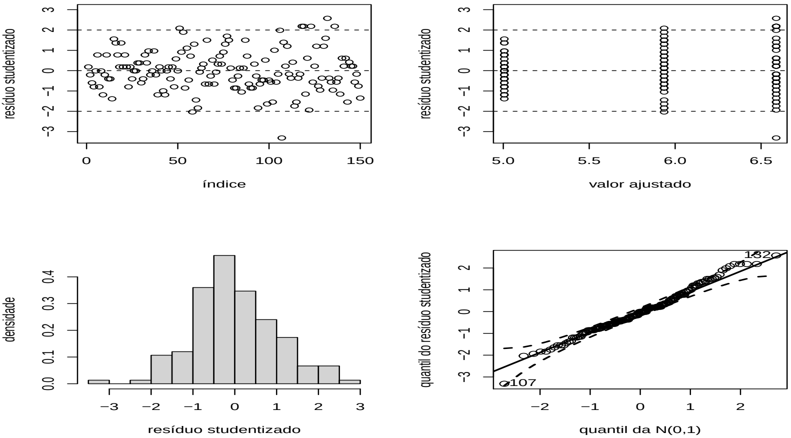
Análise de resíduos

- Veja, para o caso univariado: link.
- Resíduo ordinário: $\boldsymbol{R}^* = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}$. (distribuição de referência: N(0,a), exata, em que "a" depende de $\boldsymbol{\Sigma}$ e $\boldsymbol{X}$).
- Resíduo padronizado: $\boldsymbol{R} = \boldsymbol{R}^*\boldsymbol{D}^{-1}$, em que $\boldsymbol{D} = \sqrt{(\text{diag}(\widehat{\boldsymbol{\Sigma}}))}$. (distribuição de referência: N(0,1), assintótica)
- Resíduo studentizado (RS): $\boldsymbol{R}_s = \boldsymbol{R}./\sqrt{(\boldsymbol{1}_{(n\times p)} - \boldsymbol{h}_{(n\times p)})}$, em que ./ denota a divisão elemento por elemento, $\boldsymbol{h}$ é uma matriz em que cada coluna corresponde à diagonal principal da matriz $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. (distribuição de referência: N(0,1), assintótica)

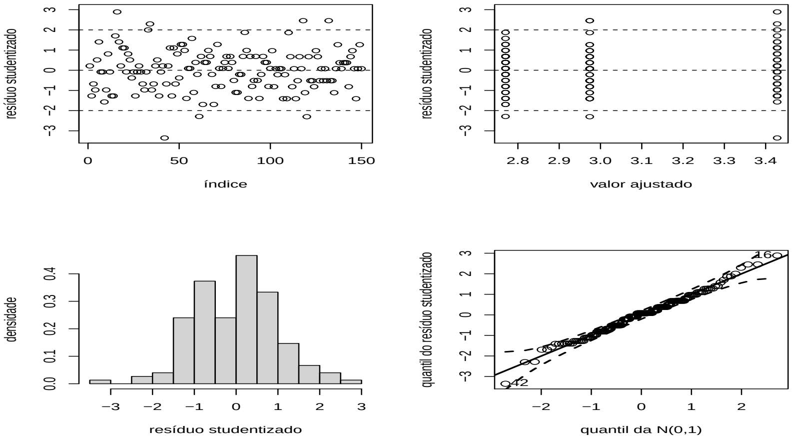
Análise de resíduos

- Resíduo multivariado (RM) : $\boldsymbol{R}_s^{(m)} = \boldsymbol{R}^*(chol(\widehat{\Sigma}))^{-1}$. (distribuição de referência: N(0,1), assintótica)
- Resíduo baseado na distância de Mahalanobis (via RM): $(\boldsymbol{R}_i^*)'\widehat{\Sigma}^{-1}\boldsymbol{R}_i^*$, em que $\boldsymbol{R}_i^* = \boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{B}}$, $i = 1, 2, ..., n$ é a i-ésima linha da matriz $\boldsymbol{R}^*$.
- Podemos também considerar uma combinação do RS e do RM, ou seja $\boldsymbol{R}_s^{*(m)} = \boldsymbol{R}_s^{(m)}./\sqrt{(\boldsymbol{1}_{(n\times p)} - \boldsymbol{h}_{(n\times p)})}$ (resíduo studentizado multivariado).
- Exercício: escrever os valores individuais, para cada um dos resíduos acima .

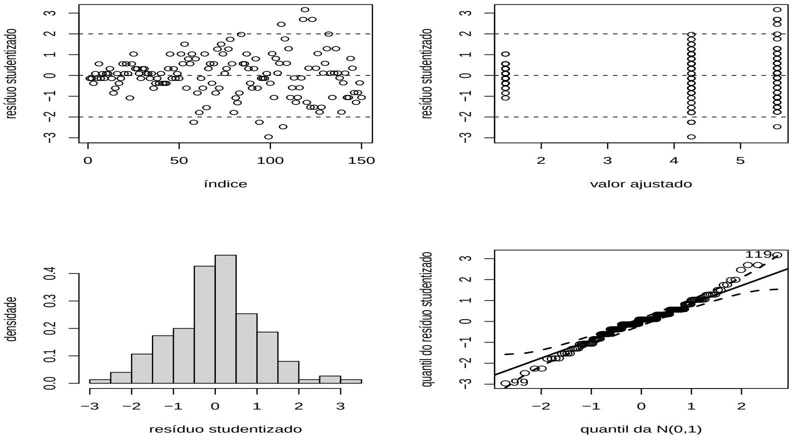
Análise de resíduos (RS) - comprimento da sépala



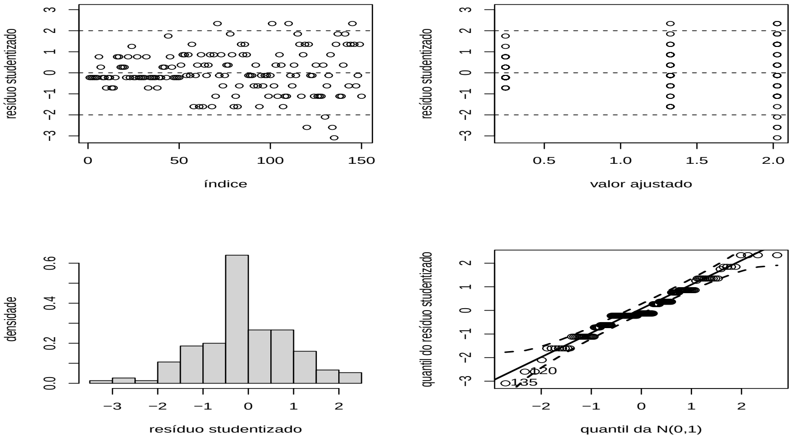
Análise de resíduos (RS) - largura da sépala



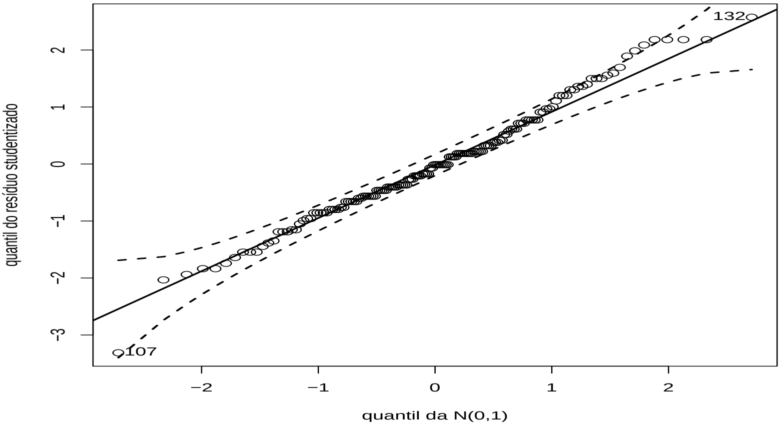
Análise de resíduos (RS) - comprimento da pétala



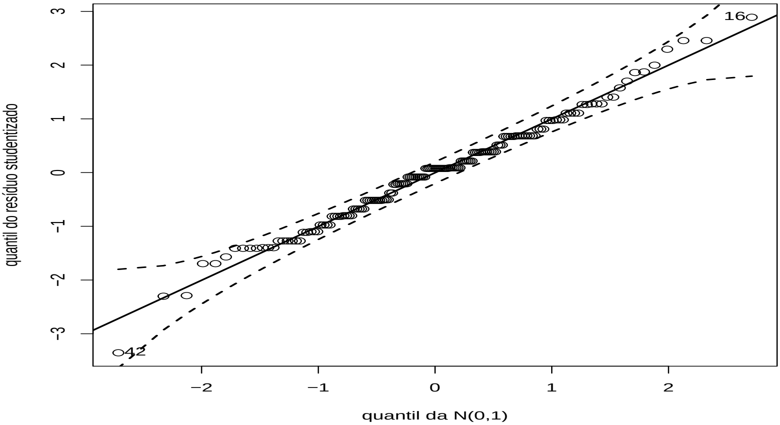
Análise de resíduos (RS) - largura da pétala



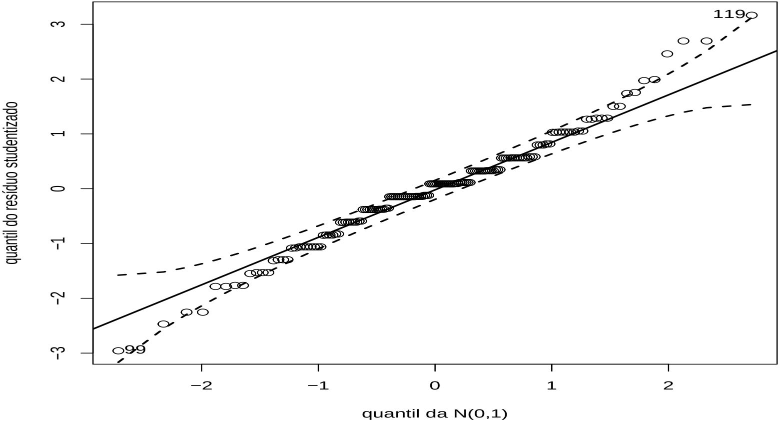
Envelopes (RS) - comprimento da sépala



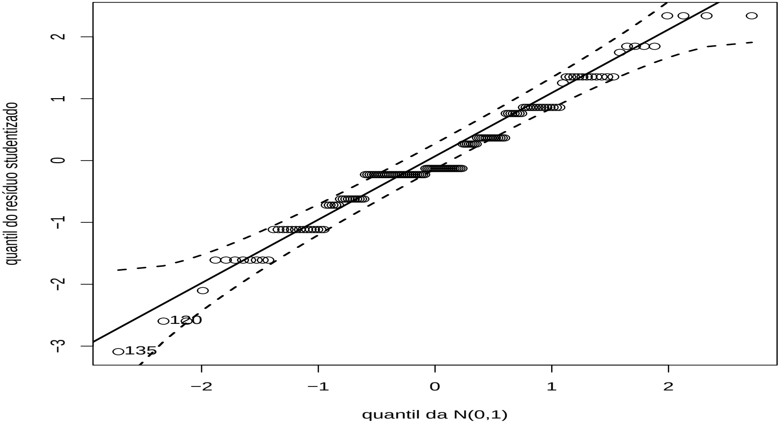
Envelopes (RS) - largura da sépala



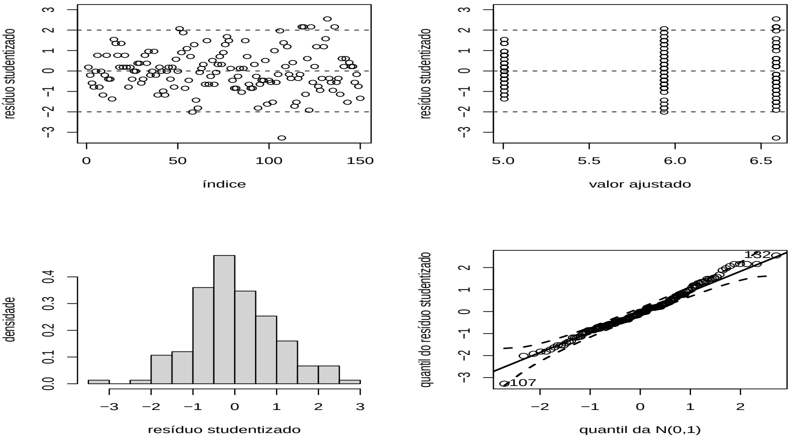
Envelopes (RS) - comprimento da pétala



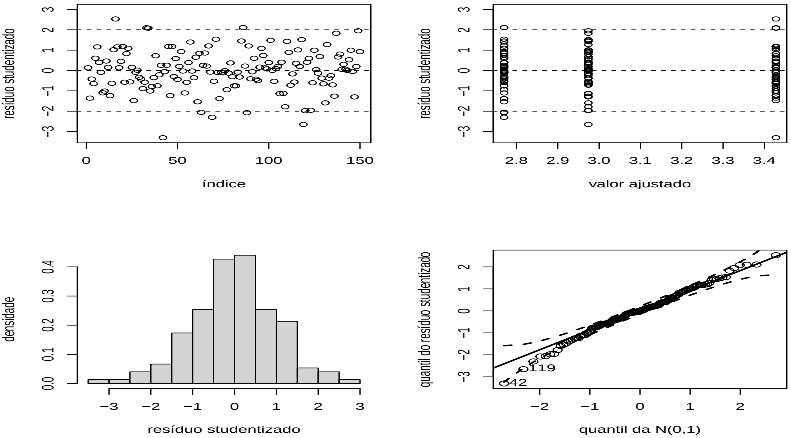
Envelopes (RS) - largura da pétala



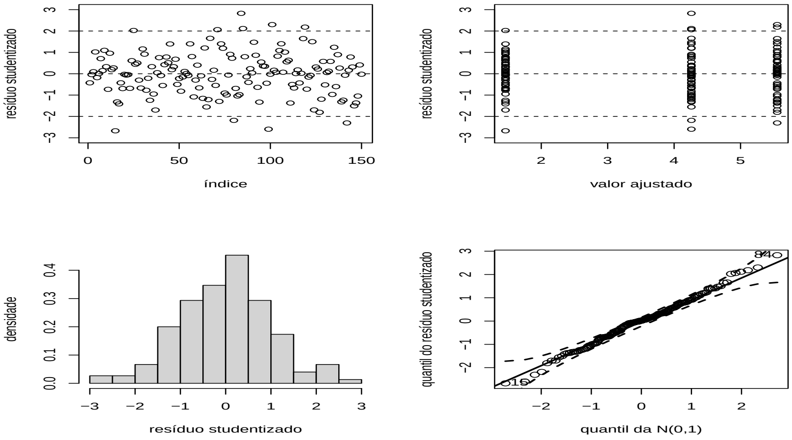
Análise de resíduos (RM) - comprimento da sépala



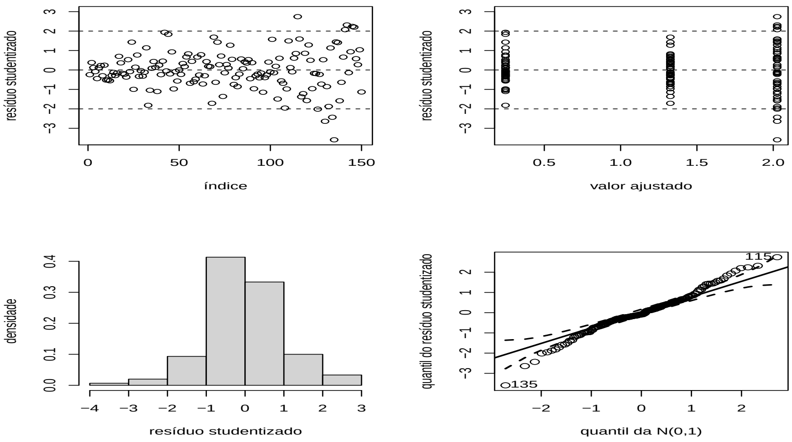
Análise de resíduos (RM) - largura da sépala



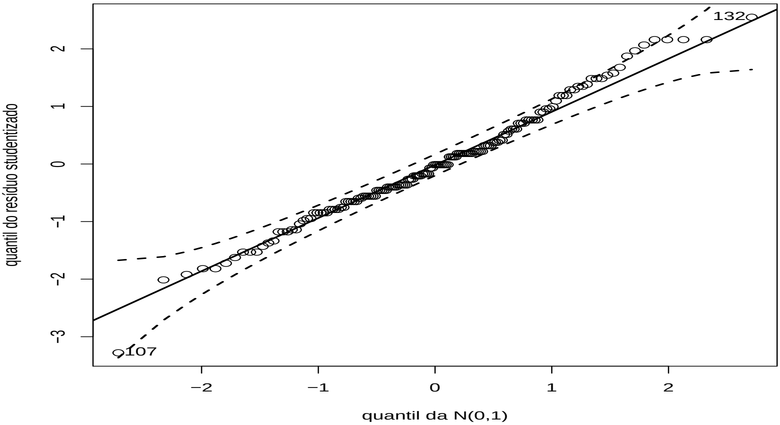
Análise de resíduos (RM) - comprimento da pétala



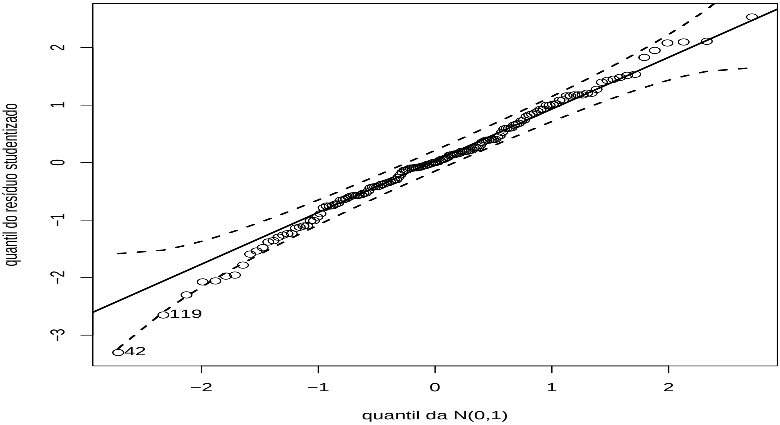
Análise de resíduos (RM) - largura da pétala



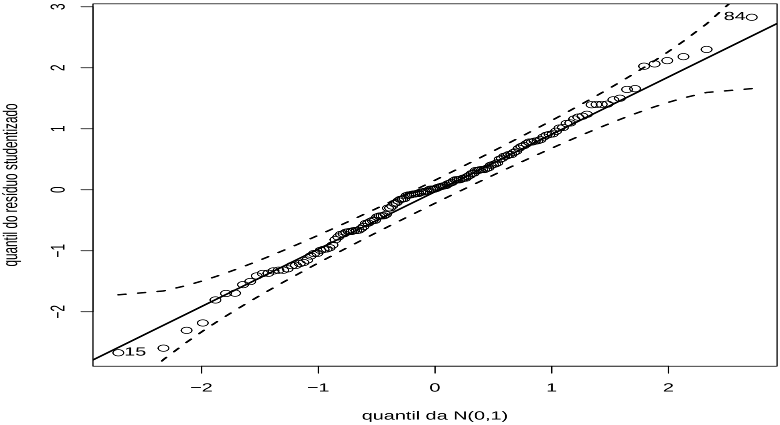
Envelopes (RM) - comprimento da sépala



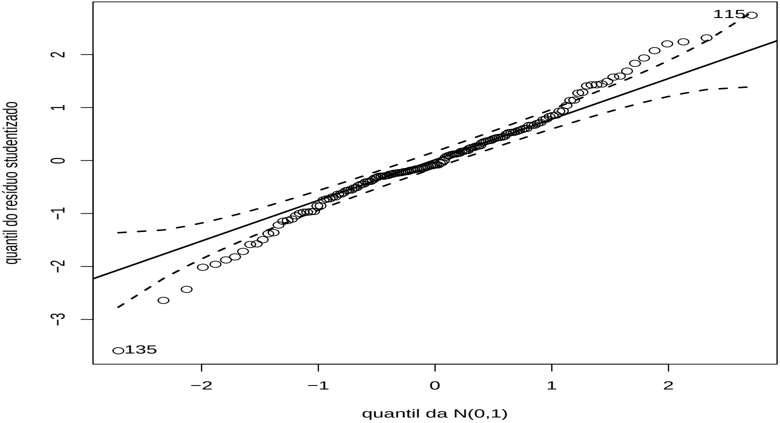
Envelopes (RM) - largura da sépala



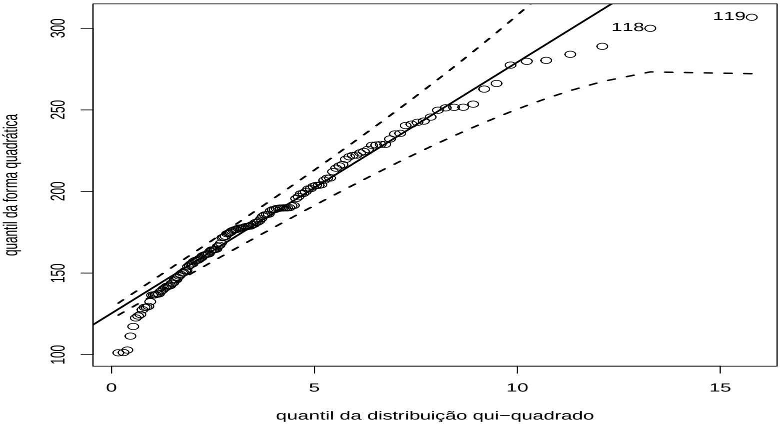
Envelopes (RM) - comprimento da pétala



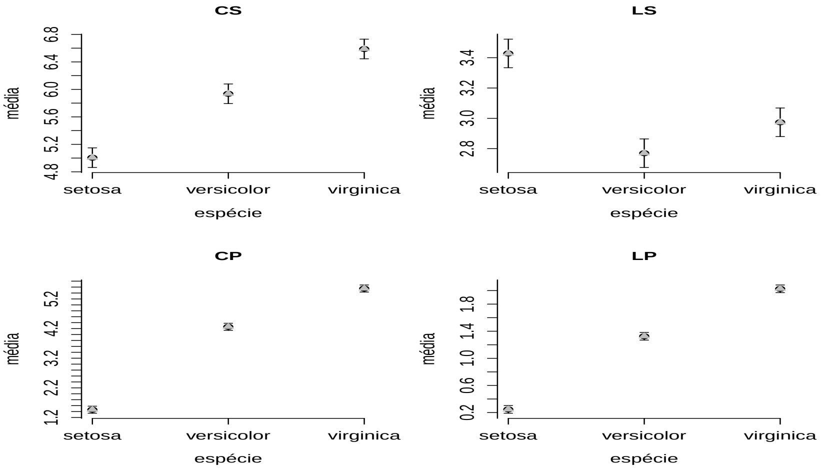
Envelopes (RM) - largura da pétala



Envelopes (RM) - distância de Mahalanobis



Médias observadas e preditas e respectivos IC(95%)



Comentários

- Os graficos de resíduos acusam:
  - Heterocedasticidade dos erros (para cada uma das variáveis).
  - Não normalidade dos erros (para cada uma das variáveis).
  - Independência entre as observações de diferentes flores (iris).
- Assim, podemos concluir que o modelo não está bem ajustado
- Lembrando que o MRNLM é, em geral, sensível (não robusto) à ausência das suposições sobre os erros, conclui-se que os resulados não são confiáveis.