1. Introdução

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 morfológicas: comprimento e largura da sépala (CS, LS) e comprimento e largura da pétala (CP,LP). O objetivo original é quantificar a variação morfológica em relação a essas espécies com bases nas quatro variáveis de interesse. Em termos da presente análise o objetivo será traduzido como a comparação das médias dessa três espécies, os quais serão os grupos de interesse, em relação às variáveis medidas, através da metodologia de Análise de Variância Multivariada (MANOVA), veja Johnson and Wichern (2007).Todas as análise foram realizadas via pacote R versão 3.2.1 (R core team (2015)).

2. Análise descritiva

Na Figura 1 temos os diagramas de dispersão entre as variáveis, com os grupos em destaque. Podemos notar que parece haver correlação positiva entre todas as variáveis, para os três grupos. Além disso, vemos que os grupos se diferenciam, em relação às quatro variáveis, sendo a diferença mais acentuada entre o grupo setosa e os outros dois, sendo esses últimos mais semelhantes entre si.

As Tabelas de 1 a 4 apresentam algumas medidas resumo para as quatro variáveis por grupo. Podemos notar que as médias amostrais mostram-se bem

diferentes entre si (em relação aos grupos para cada variável), indicando uma diferença entre os grupos nesse quesito. Também observamos uma diferença entre as variâncias, o que indica uma possível heterocedasticidade dos dados (pela Figura 1 as covariâncias parecem ser diferentes também).

A Figura 3 apresenta o gráfico de quantis-quantis com envelopes, para a distância de Mahalanobis (Azevedo (2015)), para cada grupo. Vemos que a suposição de normalidade multivariada dos dados parece não ser razoável.

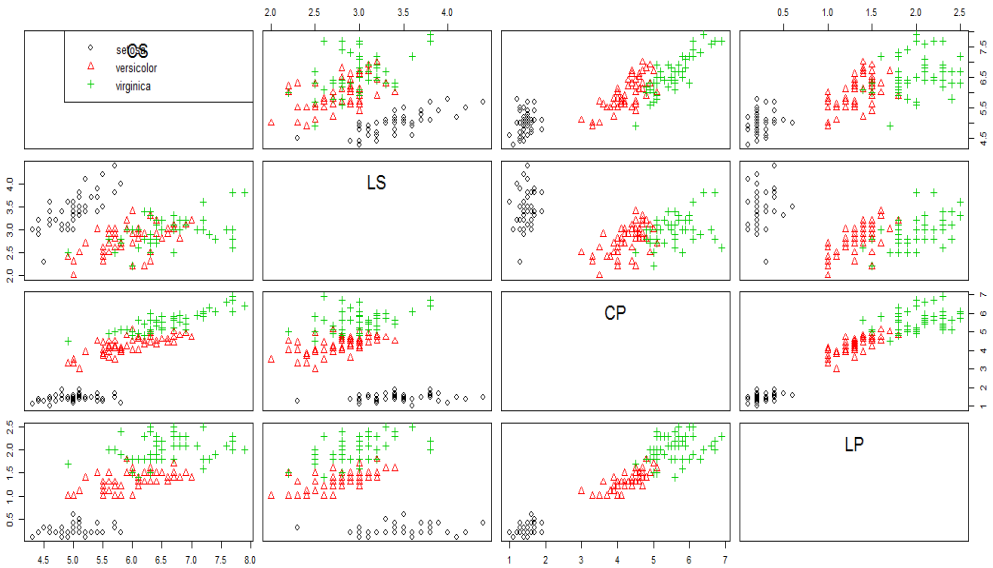

Figura 1: Matriz de gráficos de dispersão entre as variáveis

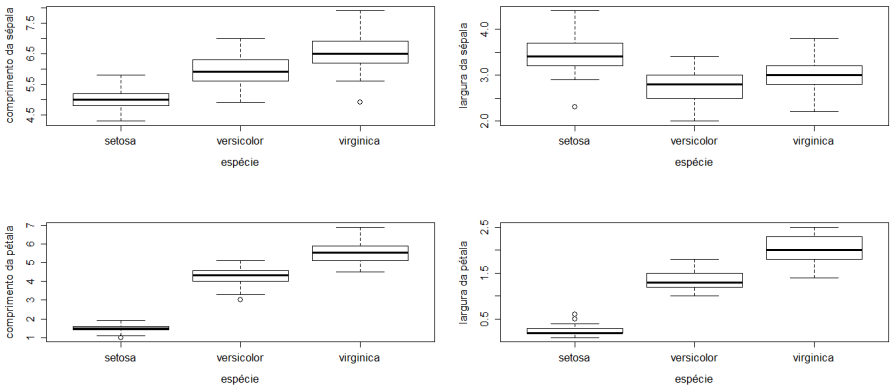

Figura 2: Box-plot das variáveis por grupo

Tabela 1: Medidas resumo por grupo para a variável CS

Espécie	Média	DP	Var.	CV(%)	Min.	Med.	Máx.
Setosa	5,01	0,35	0,12	7,04	4,30	5,00	5,80
Versicolor	5,94	0,52	0,27	8,70	4,90	5,90	7,00
Virgínica	6,59	0,64	0,40	9,65	4,90	6,50	7,90

Tabela 2: Medidas resumo por grupo para a variável LS

Espécie	Média	DP	Var.	CV(%)	Min.	Med.	Máx.
Setosa	3,43	0,38	0,14	11,06	2,30	3,40	4,40
Versicolor	2,77	0,31	0,10	11,33	2,00	2,80	3,40
Virgínica	2,97	0,32	0,10	10,84	2,20	3,00	3,80

Tabela 3: Medidas resumo por grupo para a variável CP

Espécie	Média	DP	Var.	CV(%)	Min.	Med.	Máx.
Setosa	1,46	0,17	0,03	11,88	1,00	1,50	1,90
Versicolor	4,26	0,47	0,22	11,03	3,00	4,35	5,10
Virgínica	5,55	0,55	0,30	9,94	4,50	5,55	6,90

Tabela 4: Medidas resumo por grupo para a variável LP

Espécie	Média	DP	Var.	CV(%)	Min.	Med.	Máx.
Setosa	0,25	0,11	0,01	42,48	0,10	0,20	0,60
Versicolor	1,33	0,20	0,04	14,91	1,00	1,30	1,80
Virgínica	2,03	0,27	0,08	13,56	1,40	2,00	2,50

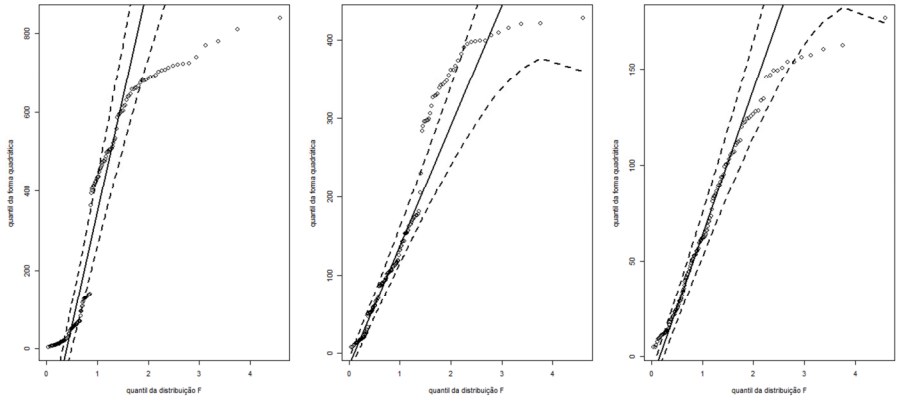

Figura 3: Gráfico de quantil-quantil com envelopes para a distância de Mahalanobis

3. Análise Inferencial

Com o objetivo de comparar os grupos, o seguinte modelo foi ajustado

$$Y_{ijk} = \mu_k + \alpha_{ik} + \xi_{ijk}, \alpha_{1k} = 0, \vee k, \xi_{ijk} \sim N_3(\underset{\sim}{0}, \underset{\sim}{\Sigma}), i = 1,2,3(grupo, 1 - setosa, 2 - verisolcor, 3 - virginica), j = 1,2,...,50(indivíduo), k = 1,2,3,4(\text{var}iável, 1 - CS, 2 - LS, 3 - CP, 4 - LP)$$

O modelo em questão foi ajustado via mínimos quadrados generalizados (veja Azevedo (2015)) e as quatro estatísticas multivariadas foram calculadas de acordo com a metodologia MANOVA (Johnson and Wichern (2007)). A Tabela 5 apresenta os resultados referentes às quatro estatísticas multivariadas enquanto que a Tabela 6 apresenta as estimativas dos parâmetros do modelo.

Tabela 5: Resultados da MANOVA

Estatística	Valor	Aproximação pela distribuição F	p-valor
Wilks	0,02	119,15	< 0,0001
Pillai	1,19	53,46	< 0,0001
Hotelling-Lawley	32,47	580,53	< 0,0001
Roy	32,19	1167,00	< 0,0001

Tabela 6: Estimativa dos parâmetros do modelo

Variável: CS				
Parâmetro	Estimativa	Ep	Estatística t	p-valor
μ_1	5,00	0,07	68,72	< 0,0001
α_{21}	0,93	0,10	9,03	< 0,0001
α_{31}	1,58	0,10	15,37	< 0,0001
Variável: LS				
Parâmetro	Estimativa	Ep	Estatística t	p-valor
μ_2	3,43	0,05	71,36	< 0,0001
α_{22}	-0,66	0,07	-9,69	< 0,0001
α_{32}	-0,45	0,07	-6,68	< 0,0001
Variável: CP				
Parâmetro	Estimativa	Ep	Estatística t	p-valor
μ_3	1,43	0,06	24,02	< 0,0001
α_{23}	2,80	0,09	32,51	< 0,0001
α_{33}	4,09	0,09	47,52	< 0,0001
Variável: LP				
Parâmetro	Estimativa	Ep	Estatística t	p-valor
μ_3	0,25	0,03	8,50	< 0,0001
α_{23}	1,08	0,04	26,39	< 0,0001
α_{33}	1,78	0,04	43,49	< 0,0001

Pela Tabela 5 vemos, claramente, que existe algum padrão de diferença entre as médias (em relação aos grupos e variáveis). Pelos resultados da Tabela 6 podemos concluir que as médias, para cada variável, do grupo de referência (setosa) são diferentes em relação às médias dos dois outros grupos. Mais especificamente, a média do grupo setosa é maior do que as dos outros dois para a variável LS e menor para as outras variáveis. Utilizaremos agora a metodologia para testar hipóteses do tipo $H_0: CBU = M$, como descrita em Azevedo (2015), para identificar possíveis diferenças entre as médias dos grupos versicolor e virginica, para cada uma das variáveis. Os resultados se encontram na Tabela 7. Podemos concluir que as médias desses dois grupos, para cada uma das variáveis, são diferentes.

As estimativas das médias preditas pelo modelo, para cada grupo e para cada variável, bem como os respectivos intervalos de confiança assintóticos, calculados utilizando-se resultados apresentados em Azevedo (2015), encontram-se na Figura 4. Podemos verificar, de fato, que as médias dos grupos são diferentes entre si, para todas as variáveis e que, com exceção da variável LS, as médias dos grupos em ordem crescente é: virginica, versicolor e setosa. Para a variável LS temos setosa, virginica e versicolor, nessa ordem. Assim, temos que os grupos de flores são bem diferentes entre si, em relação às características morfológicas estudadas, seguindo o supramencionado padrão. Como as médias foram preditas pelo modelo completo (considerando todos os grupos) elas coincidem com as médias amostrais (Tabelas de 1 a 4).

Tabela 7: Resultados da comparação entre as médias dos grupos versicolor e virginica

Variável	Estatística	p-valor
CS	40,1	< 0,0001
LS	9,02	0,0027
CP	225,35	< 0,0001
LP	292,49	< 0,0001

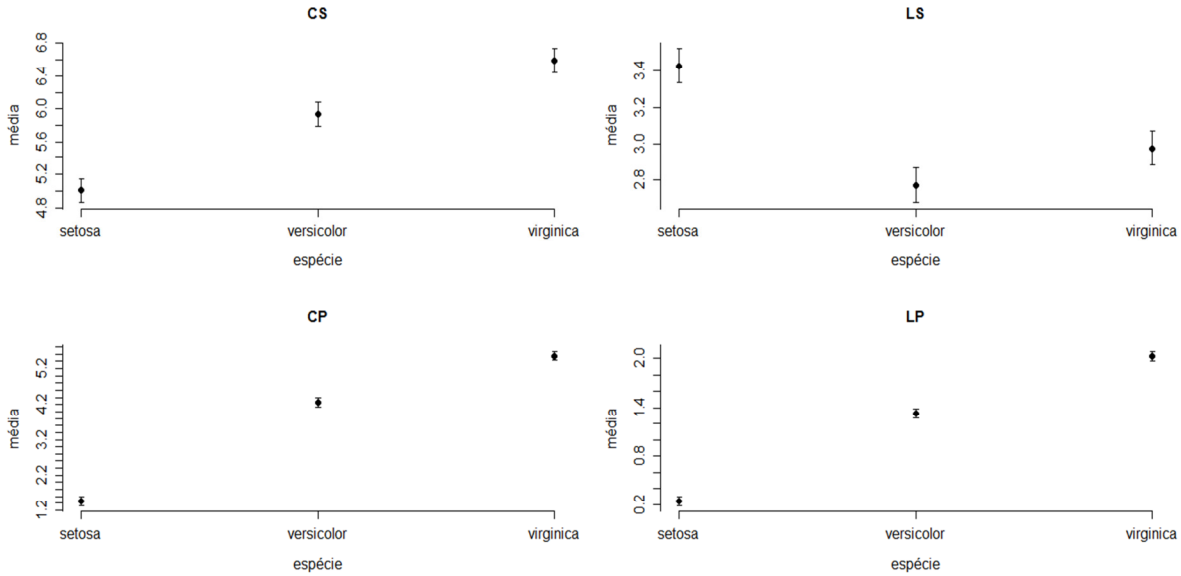

Figura 4: Médias preditas pelo modelo e respectivos intervalos de confiança para cada variável.

As Figuras 5, 6, 7 e 8 apresentam gráficos para o resíduo studentizado para cada variável, veja Azevedo (2015). Notamos uma presença de heterocedasticidade para as variáveis CS, CP e LP, devido à variabilidade oscilante, ao longo dos índices, vista nos gráficos de resíduo x índice (primeiro gráfico) e ao longo dos valores preditos, vista no gráfico de resíduo x valor predito (segundo gráfico sentido horário). Pelos histogramas (quarto gráfico no sentido horário) observamos uma aparente simetria para a variável somente para a variável CP, enquanto que para as outras observamos uma leve assimetria negativa. Esses resultados podem indicar uma possível não normalidade dos resíduos, para as variáveis CS, LS e LP. Os gráficos de envelopes (terceiro no sentido horário) indicam uma leve concavidade para cima para a variável CS e uma maior quantidade de pontos acima da linha de referência (nas caudas) para a variável LS, o que sugere um comportamento sistemático e, em ambos os casos, uma possível assimetria positiva. Para a variável CP vemos alguns pontos (nas caudas) fora dos respectivos envelopes, sendo que na cauda inferior os pontos estão abaixo da linha de referência, enquanto que na superior os pontos estão acima. O padrão de "escada" observado para a variável LP pode ser devido à pequena variabilidade dos dados o que, não necessariamente, indica uma não normalidade.

Em resumo, o modelo para não ter se ajustado bem aos dados devido à presença de heterocedasticidade (em relação à algumas variáveis) e uma possível não normalidade (em relação à algumas variáveis). Assim sugere-se ajustar um modelo em que ao menos a heterocedasticadade seja contemplada. Se ainda houver indícios, para esse novo modelo, de não normalidade dos resíduos, um terceiro

modelo, que considere uma distribuição mais flexível do que a normal, além da heterocedasticidade, deve ser considerado.

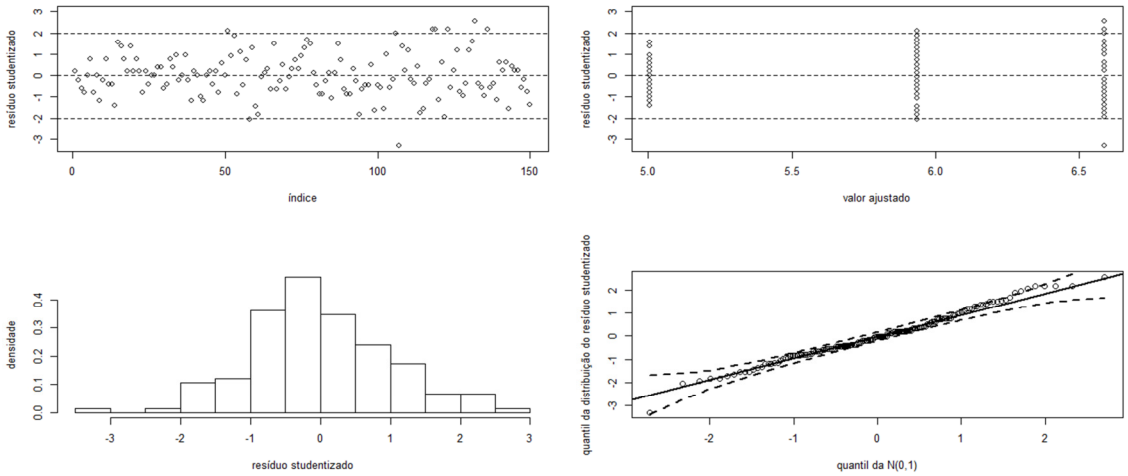

Figura 5: Gráficos para os resíduos para a variável CS

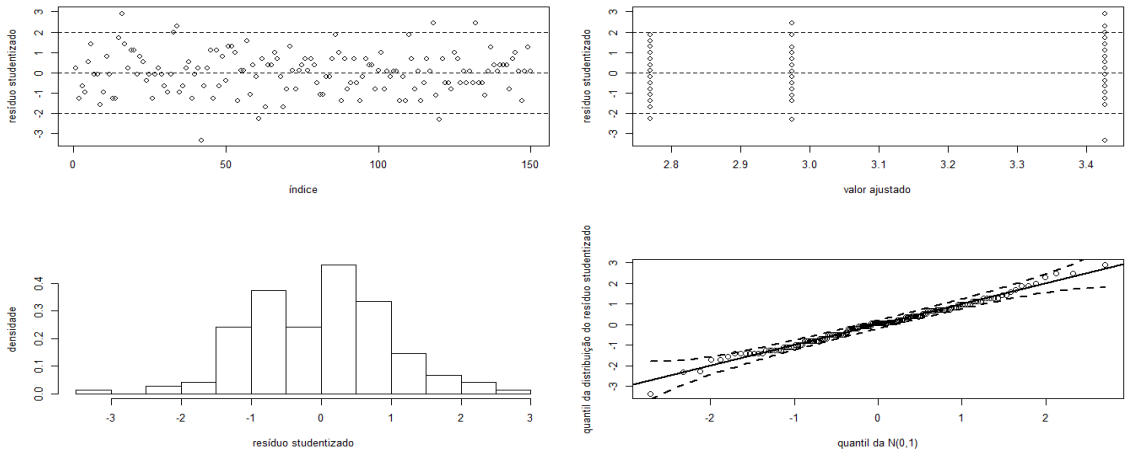

Figura 6: Gráficos para os resíduos para a variável LS

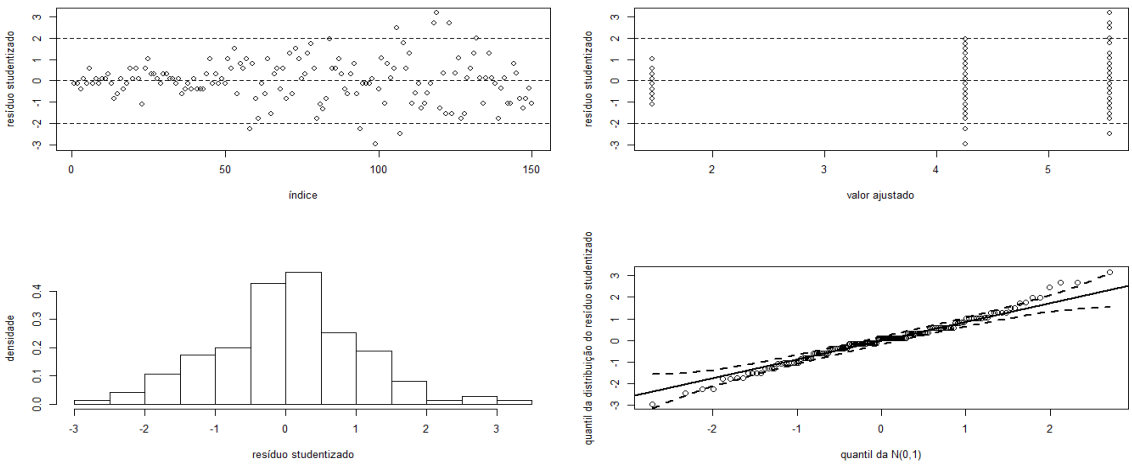

Figura 7: Gráficos para os resíduos para a variável CP.

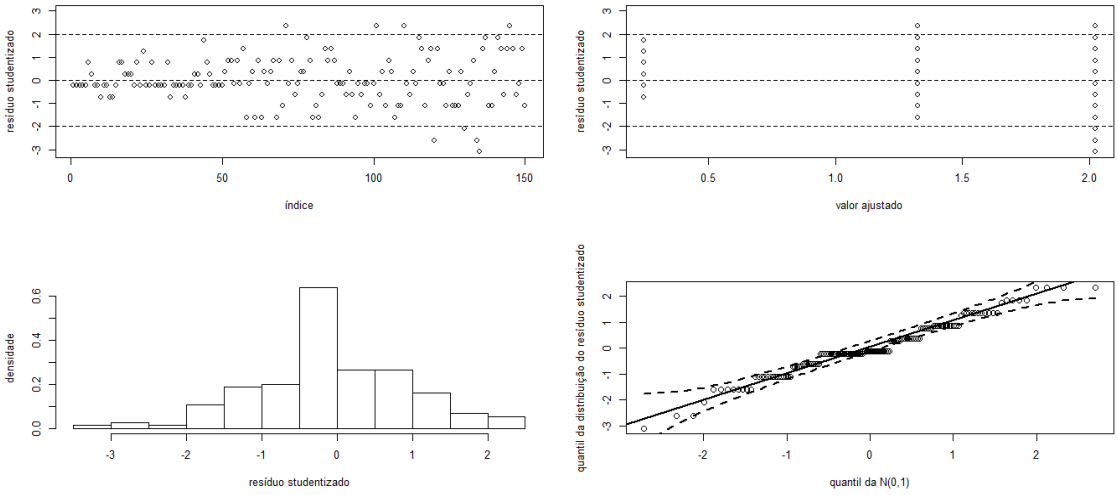

Figura 8: Gráficos para os resíduos para a variável LP

4. Conclusões

O modelo ajustado nos permitiu concluir que, de fato, sob a suposição de normalidade, os grupos se diferenciam mutuamente em relação às variáveis medidas. Além disso, em termos de predição pontual das médias, como era de se esperar, o modelo se comportou de forma adequada. No entanto, como o modelo não se ajustou bem a predição intervalar pode não ser adequada, haja vista que pelo menos algumas das suposições do modelo não foram satisfeitas, conforme indicou a análise residual. Assim, outras abordagens devem ser utilizadas para uma correta análise do conjunto de dados.

5. Bibliografia

- Azevedo, C. L. N (2015). Notas de aula sobre análise multivariada de dados http://www.ime.unicamp.br/~cnaber/Material_AM_2S_2015.htm
- Johson, R. A. and Wichern, D. W. (2007). Applied Multivariate Statistical Analysis, 7ª edição, Upper Saddle River, NJ : Prentice-Hall.
- R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.