Universidade Estadual de Campinas
Métodos em Análise Multivariada

Trabalho ME731

Parte 1

Anita Maria de Oliveira Lobo 154692
Bruna Mendes Heyn 154836
Gabriel Vivaldini Silva 155483
Natalia Rodrigues da Silva 156831
Ricardo Grella Vieira Simões Ferreira 139742

Professor: Caio Lucidius Naberezny Azevedo

Dezembro/2017

Sumário

1 Introdução

Os dados foram obtidos do arquivo moscas.txt, extraído do site http://www.ime.unicamp.br/~cnaber/Moscas.txt. Este se refere a sete variáveis medidas em duas espécies das moscas chamadas bitting fly (Leptoconops carteri e Leptoconops torrens), sendo elas espécie (0 - Leptoconops torrens e 1- Leptoconops carteri), comprimento da asa, largura da asa, comprimento do terceiro palpo, largura do terceiro palpo, comprimento do quarto palpo, comprimento do 12º segmento da antena e comprimento do 13º segmento da antena. Para ser mais eficiente, essas variáveis foram renomeadas como sendo: Especie (0 - torrens e 1- carteri), C.Asa, L.Asa, C3p, L3p, C4p, C12a e C13a, respectivamente.

Estas duas espécies são tão semelhantes (Johson e Wichern (2007)) que chegaram a ser consideradas, pelos pesquisadores, como uma uníca espécie. Sendo assim, objetivo do estudo é comparar as duas espécies de moscas em relação as variáveis citadas acima para saber se há diferença entre esses dois grupos e, se houver, em que variáveis reside(m) essa(s) diferença(s). Os principais métodos utilizados para este fim foram a aplicação da análise de variância multivariada(MANOVA) e testes de significâncias individuais para os parâmetros.

2 Análise Descritiva

Na tabela 1, são apresentadas as medidas resumos das variáveis presentes no banco de dados, separadas por espécie. É possível observar que as diferenças entre as médias dos grupos para as variáveis L.Asa, L3p, C12a e C13a diferem em menos de 1 ponto, enquanto para as variáveis C.Asa, C3p e C4p essa diferença é de 2,88, 3,94 e 3,37 respectivamente . Para os desvios padrões também observamos que para as variáveis C.Asa, C3p, L3p C12a e C13a a diferença entre os grupos é de menos de 1 ponto enquanto para L.Asa e C4p essa diferença é de 2,32 e 2,11 respectivamente.

Tabela 1: Medidas resumo

Variável	Especie	Média	DP	Var	CV	Mínimo	Mediana	Máximo	n
C.Asa	Carteri	99,34	5,59	31,29	5,63	82	99,00	112	35
	Torrens	96,46	6,38	40,73	6,62	85	95,00	109	35
L.Asa	Carteri	43,74	5,08	25,78	11,61	19	45,00	50	35
	Torrens	42,91	2,74	7,49	6,38	38	44,00	49	35
C3p	Carteri	39,31	2,84	8,05	7,21	33	39,00	44	35
	Torrens	35,37	2,20	4,83	6,21	31	36,00	39	35
L3p	Carteri	14,66	1,64	2,70	11,22	11	15,00	19	35
	Torrens	14,51	1,84	3,37	12,66	11	14,00	18	35
C4p	Carteri	30,00	4,61	21,29	15,38	20	31,00	38	35
	Torrens	25,63	2,50	6,24	9,75	21	26,00	31	35
C12a	Carteri	9,66	1,26	1,58	13,04	6	10,00	12	35
	Torrens	9,57	0,92	0,84	9,58	8	9,00	13	35
C13a	Carteri	9,37	1,09	1,18	11,60	7	9,00	11	35
	Torrens	9,71	0,89	0,80	9,20	8	10,00	13	35

Na figura 1, é apresentado o gráfico de dispersão entre as variáveis de interesse para ambos os grupos. Pode-se notar que, em geral, as variáveis apresentam valores menores para a espécie Torrens em comparação com a espécie Carteri. Dois exemplos disso são os gráfico entre as variáveis C4p e L.Asa e C.Asa e Cp3, que concentram quase todos os pontos referentes à espécie Carteri em uma região densa acima da concentração de pontos referentes à espécie Torrens. Os gráficos também nos dão indícios de associações, por grupo, entre as variáveis os quais podem sem comprovados ou refutados pela tabela 2, a qual apresenta uma matriz de covariancias (parte triangular inferior) e correlações (parte triangular superior) entre todas as características de interesse, para cada grupo. Podemos notar, por exemplo, que para as variáveis C.Asa e C3p o gráfico mostra que para a espécie Carteri os dados apresentam uma determinada associação linear que é comprovada pela tabela 2, onde vemos que a correlação entre elas é de 0,62, enquanto para Torrens a dispersão não parece ter uma associação significativa, o que é novamente comprovado pela tabela 2 que nos mostra que a correlação dessas características para essa espécie é de apenas 0,17. As dispersões entre C.Asa e L.Asa e C12a e C13a aparentam ter alta associação linear para ambos os grupos, o que é verificado quando, na tabela 2, vemos que os valores das correlações entre tais variáveis para Torrens e Carteri são, respectivamente, 0,67, 0,61, 0,78 e 0,87.

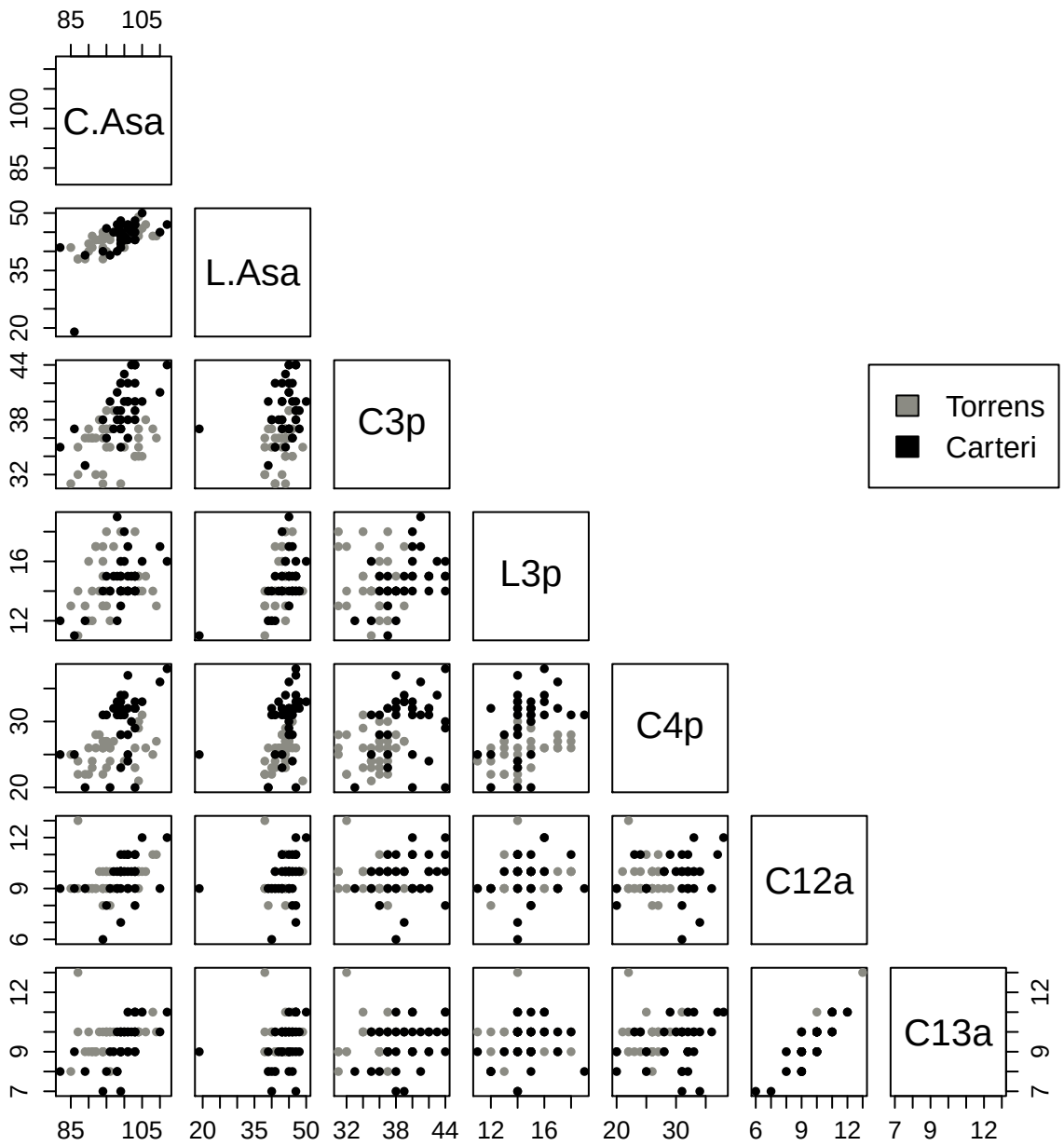


Figura 1: Gráficos de Dispersão

Tabela 2: Tabela de covariância e correlação

	C.Asa	L.Asa	C3p	L3p	C4p	C12a	C13a
TORRENS							
C.Asa	-	**0,67**	**0,17**	0,19	0,39	0,32	0,29
L.Asa	11,72	-	0,30	0,37	0,48	0,13	0,12
C3p	2,33	1,83	-	-0,19	0,11	-0,18	0,09
L3p	2,20	1,84	-0,78	-	0,37	0,10	0,00
C4p	6,26	3,26	0,61	1,70	-	-0,01	-0,02
C12a	1,88	0,32	-0,37	0,17	-0,02	-	**0,78**
C13a	1,66	0,30	0,17	0,00	-0,05	0,64	-
CARTERI							
C.Asa	-	**0,61**	**0,62**	0,56	0,50	0,42	0,60
L.Asa	17,47	-	0,26	0,50	0,38	0,28	0,28
C3p	9,83	3,70	-	0,46	0,20	0,22	0,38
L3p	5,15	4,14	2,17	-	0,41	0,18	0,25
C4p	12,88	8,94	2,62	3,12	-	0,20	0,26
C12a	2,97	1,79	0,79	0,38	1,15	-	**0,87**
C13a	3,63	1,57	1,17	0,46	1,32	1,19	-

Na figura 2, estão disponibilizados os boxplots referentes às variáveis de interesse para ambos os grupos. A princípio, nota-se a grande presença de outliers para a espécie Carteri em seis dos sete gráficos, enquanto que a espécie Torrens apresenta outliers em apenas um deles. Além disso, também para seis das sete variáveis, a espécie Carteri apresenta valores superiores da mediana em relação à espécie Torrens, não ocorrendo tal evento apenas para a variável C13a. Ainda utilizando os boxplots (figura 2), pode-se identificar, pelas distâncias entre os primeiros e terceiros quartis em relação à mediana, que há bastante assimetria nas distribuições, e que a distribuições das variáveis aparentam ser diferentes entre as espécies, sendo que as menores diferenças parecem estar nas caracterísicas C3p e C4p.

A figura 3 mostra os gráficos de quantis-quantis com envelopes para a distância de Mahalanobis para cada espécie. É perceptível que a suposição de normalidade multivariada não aparenta ser razoável para cada uma das espécies. Nota-se, nestes gráficos, que há diversas fugas para quantis menores na parte inferior do gráfico, e de quantis maiores na parte superior do gráfico da distribuição F, o que indica que a suposição de normalidade multivariada pode não ser adequada aos dados.

Diante destas análises pode-se conjecturar que exista, entre as espécies, diferença na média de variáveis como C.Asa, C3p e C4p. Para verificar a validade ou não de tais conjecturas foi feita uma análise inferencial dos dados.

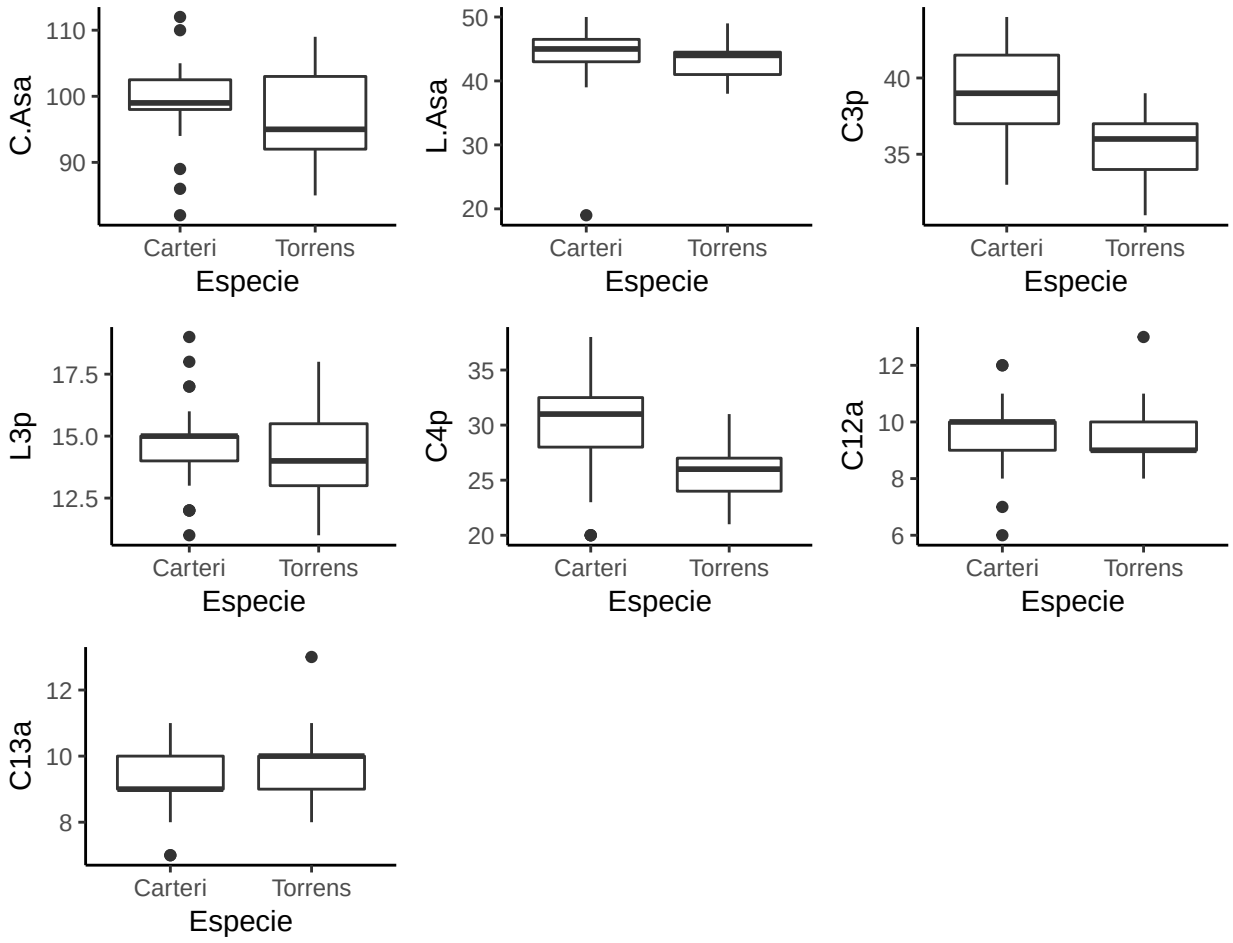


Figura 2: BoxPlots

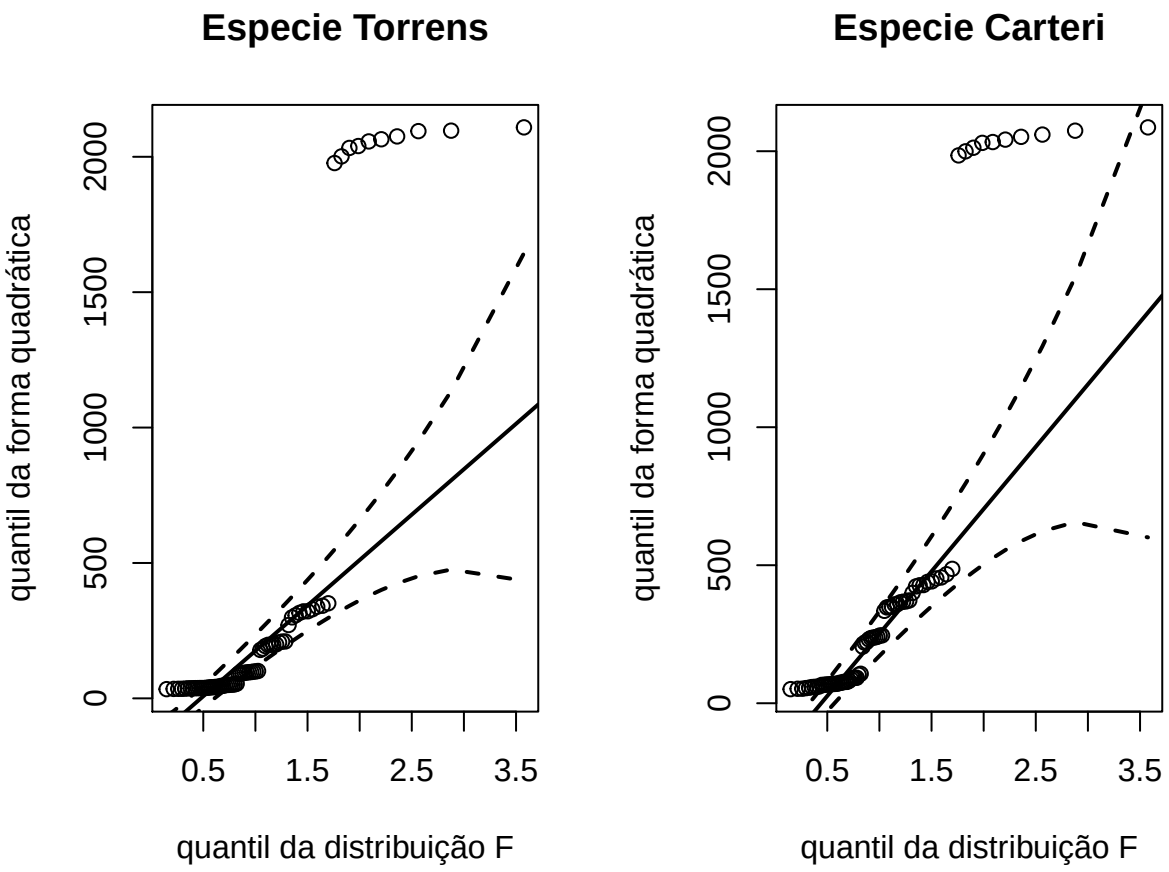


Figura 3: Gráfico de quantil-quantil com envelopes para a distância de Mahalanobis

3 Análise inferencial

Para comparar os grupos Torrens e Carteri com relação às variáveis de interesse, o seguinte modelo foi ajustado:

$$Y_{ijk} = \mu_k + \alpha_{ik} + \varepsilon_{ijk} \; com \; \alpha_{1k} = 0 \; e \; \boldsymbol{\varepsilon_{ij}} \sim N_7(\mathbf{0}, \blacksquare)$$

onde:

- i = 1,2 (grupo, 1-Torrens, 2-Carteri)
- j = 1, 2,..., 35 (indivíduo)
- k = 1, 2,..., 7 (variáveis: 1-C.Asa, 2–L.Asa, 3–C3p, 4–L3p, 5–C4p, 6–C12a, 7–C13a).

O modelo foi ajustado de acordo com as metodologias vistas em sala de aula (ver Azevedo(2017)).

A tabela 3 mostra as estimativas e significância dos parâmetros do modelo para cada variável resposta. Através dela é possivel notar que as médias para as características C.Asa, C3p e C4p diferem entre as espécies e que o grupo Carteri apresenta a maior média. Para as demais variáveis, conclui-se que não existe diferença quanto a média das espécies Torrens e Carteri.

Tabela 3: Estimativas dos parâmetros do modelo de regressão

Variável	Parâmetro	Estimativa	Erro Padrão	Valor t	p-valor
C.Asa	μ_1	96,46	1,01	95,10	<0,01
	α_1	2,89	1,43	2,01	0,05
L.Asa	μ_1	42,91	0,69	62,24	<0,01
	α_1	0,83	0,98	0,85	0,40
C3p	μ_1	35,37	0,43	82,48	<0,01
	α_1	3,94	0,61	6,50	<0,01
L3p	μ_1	14,51	0,29	49,26	<0,01
	α_1	0,14	0,42	0,34	0,73
C4p	μ_1	25,63	0,63	40,86	<0,01
	α_1	4,37	0,89	4,93	< 0,01
C12a	μ_1	9,57	0,19	51,42	<0,01
	α_1	0,09	0,26	0,33	0,75
C13a	μ_1	9,71	0,17	57,76	<0,01
	α_1	-0,34	0,24	-1,44	0,15

As constatações da tabela acima são comprovadas com a realização da metodologia MANOVA, utilizada para testar a igualdade simultanea de médias entre os grupos Torrens e Carteri. Nessa metodologia quatro testes multivariados foram utilizados: Wilks, Pillai, Hotelling-Lawley e Roy, os quais têm aproximação pela distribuição F e seus resultados são mostrados na tabela 3.

Tabela 4: MANOVA

Estatística	Valor	Aprox. Dist. F	P-Valor
Wilks	0,39	13,82	<0,01
Pillai	0,61	13,82	<0,01
Hotteling-Lawley	1,56	13,82	<0,01
Roy	1,56	13,82	<0,01

Para todos os testes obteve-se o resultado de que não há evidências para afirmar que as médias são iguais. Tendo isso, o interesse, agora, está em testar cada variável separadamente a fim de identificar onde reside a diferença.

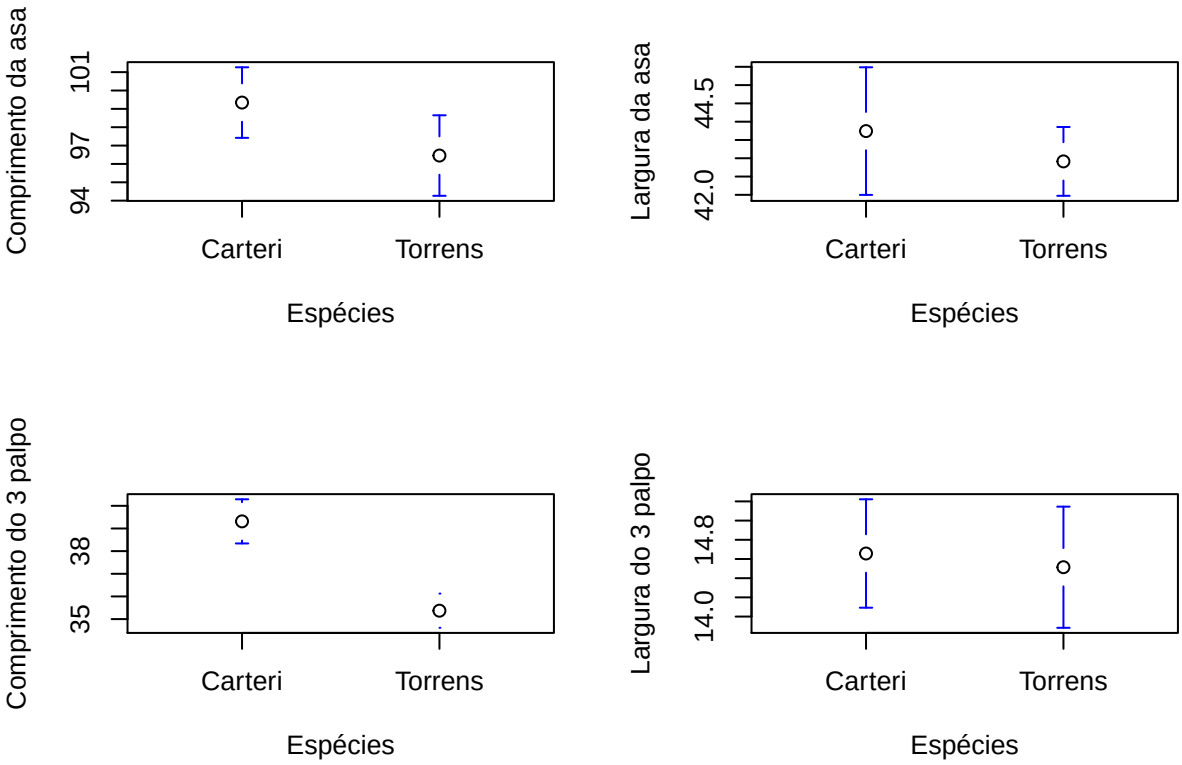
Para realizar a supracitada análise, foram utilizados testes do tipo CBU=M (ver Azevedo(2017)) considerando um nível de significância $\alpha = 0,05$ cujos resultados são mostrados na tabela 5, e por meio destes conclui-se que quatro das sete variáveis possuem as médias estatisticamente iguais em relação as duas espécies, sendo estas as mesmas indicadas pelos testes de

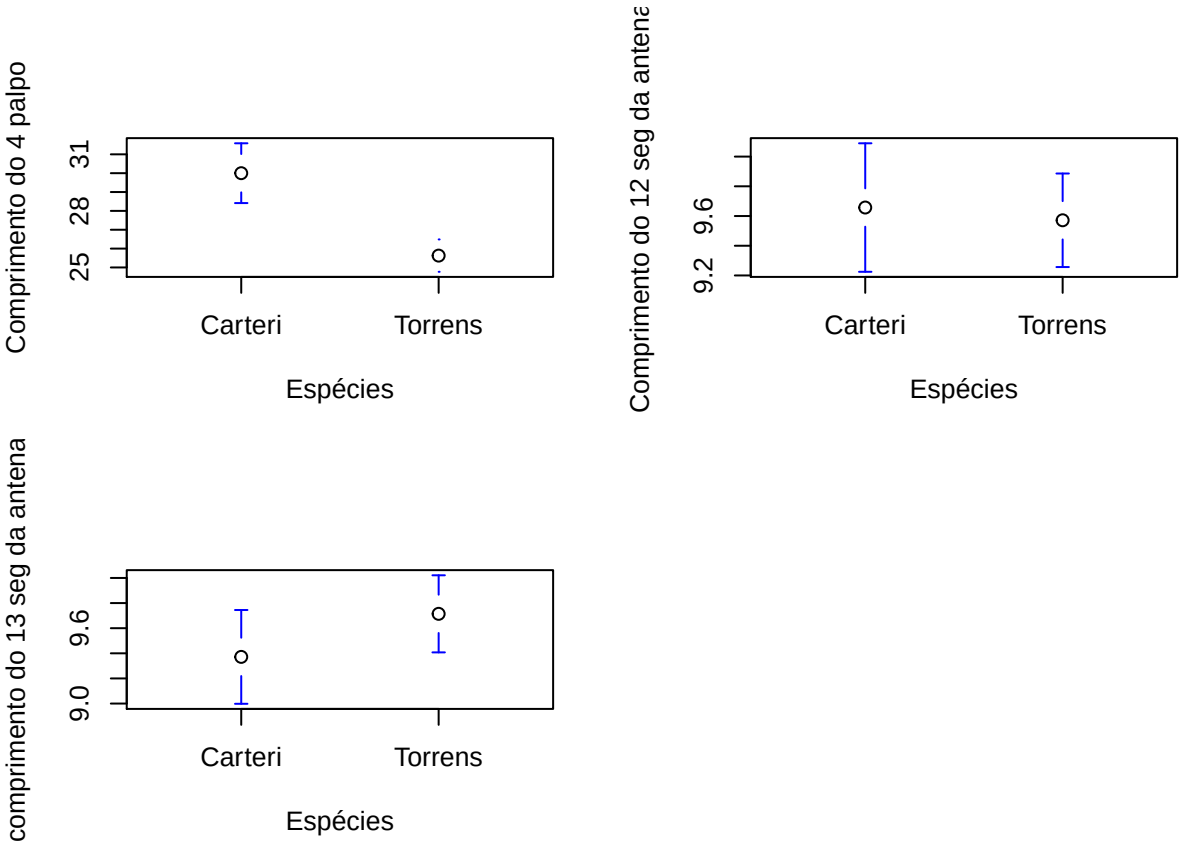
significancia dos parâmetros, ou seja, L.Asa, L3p, C12a e C13a. Para as demais (C.Asa, C3p, C4p), tem-se evidência que existe diferença entre os grupos estudados.

Tabela 5: Teste individual de nulidade das médias entre as espécies

Variável	Estatística	p-valor
C.Asa	4,05	0,00
L.Asa	0,72	0,40
C3p	42,26	0,00
L3P	0,12	0,73
C4p	24,29	0,00
C12a	0,11	0,75
C13a	2,08	0,15

Na figura 4 estão os gráficos de médias preditas pelo modelo completo para cada variável, por grupo. Os intervalos de confiança assintóticos apresentados foram calculados utilizando a metodologia estudada em sala e descrita em (Azevedo(2017)). Para as variáveis C3p e C4p pode-se notar que a espécie Carteri tem uma maior média predita e que seus intervalos de confiança não se interceptam, indicando que a distribuição destas variáveis é consideravelmente diferente entre os grupos. Apesar de o teste de igualdade de médias ter rejeitado, a um nível de 5% de confiança, a hipótese de igualdade para a variável C.Asa, no gráfico essa diferença não é tão explícita e os intervalos de confiança para os dois grupos chegam a se interseptarem. Para as outras variáveis as médias preditas são muito próximas e os intervalos de confiança têm grandes intersecções, o que está de acordo com os resultados dos testes de igualdade de médias feitos acima.


Comprimento da asa
94
97
101
Carteri
Torrens
Espécies
Largura da asa
42.0
44.5
Carteri
Torrens
Espécies
Comprimento do 3 palpo
35
38
Carteri
Torrens
Espécies
Largura do 3 palpo
14.0
14.8
Carteri
Torrens
Espécies



As figuras de 6 a 12 apresentam os gráficos de resíduo do modelo para cada variável, onde temo, da esqueda para a direita, de cima para baixo, os seguintes gráficos (A)Resíduos Studentizado x Índice, (B)Resíduos Studentizado x Valor Ajustado, (C)Hístograma dos Resíduos Studentizado e (D)Q-Qplot da distríbuição normal. A partir de agora usaremos as respectivas letras (A, B, C e D) para nos referirmos a cada um dos gráficos. Análisando os gráficos A e B nota-se que para algumas variáveis o comportamento destes gráficos indicam a presença de heterocedasticidade e dependencia dos resíduos, como por exemplo para C4p para a qual o gráfico A aprensenta um aumento da dispersão dos pontos com o aumento dos indíces, ficando com um formato de “cone” e no gráfico B a variabilidade dos resíduos para um dos grupos é bem maior que para o outro. Nas variáveis C.Asa e c3p também é observado um aumento da variábilidade entre os grupos no gráfico B e, mais especificamente para a variável C3p uma tendência em A parecida com a que foi observada em C4p. As variáveis L3p, C12a e C13a, apesar de não apresentarem tendência em A, apresentam tendências em B, indicando heterocedásticidade Para o gráfico C tem-se em toda as variáveis, exceto C12a, existe assimetria no histograma de todas as variáveis, podendo ser levemente positiva (como para L3p e C13a), levemente

negativa (como para C.Asa e C3p) ou fortemente negativa (como para L.Asa e C4p). Isso tras uma evidencia contrária a suposição de normalidade dos resíduos. No grafico D observa-se padrões no comportamento dos resíduos de todas as variáveis, e para as variáveis C.Asa, L3p, C4p e C12a também é possível notar diversos pontos fora das bandas de confiança. Estes resultados evidênciam a não validade da suposição de normalidade dos resíduos. Portanto, ao final da análise resídual conclui-se que o modelo escolhido nao se ajustou satisfatóriamente aos dados por conta da contatação de não normalidade, heterocedasticidade e, em alguns casos, possível dependência dos resíduos. Seria razoavel a tentativa de ajustar um modelo que nao suponha a homocedasticidade e, caso este ainda nao funcione, a de ajustar um modelo que nao suponha normalidade.

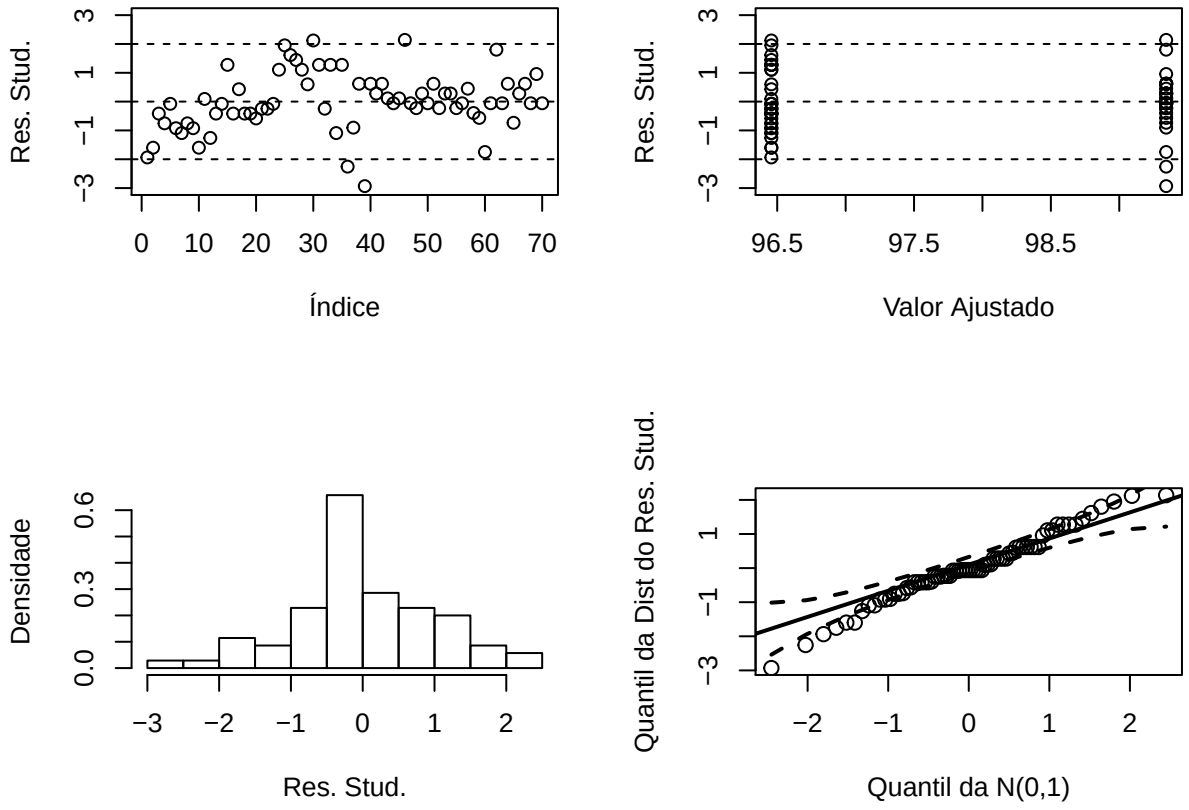


Figura 4: Gráfico de Resíduos para a variável C.Asa

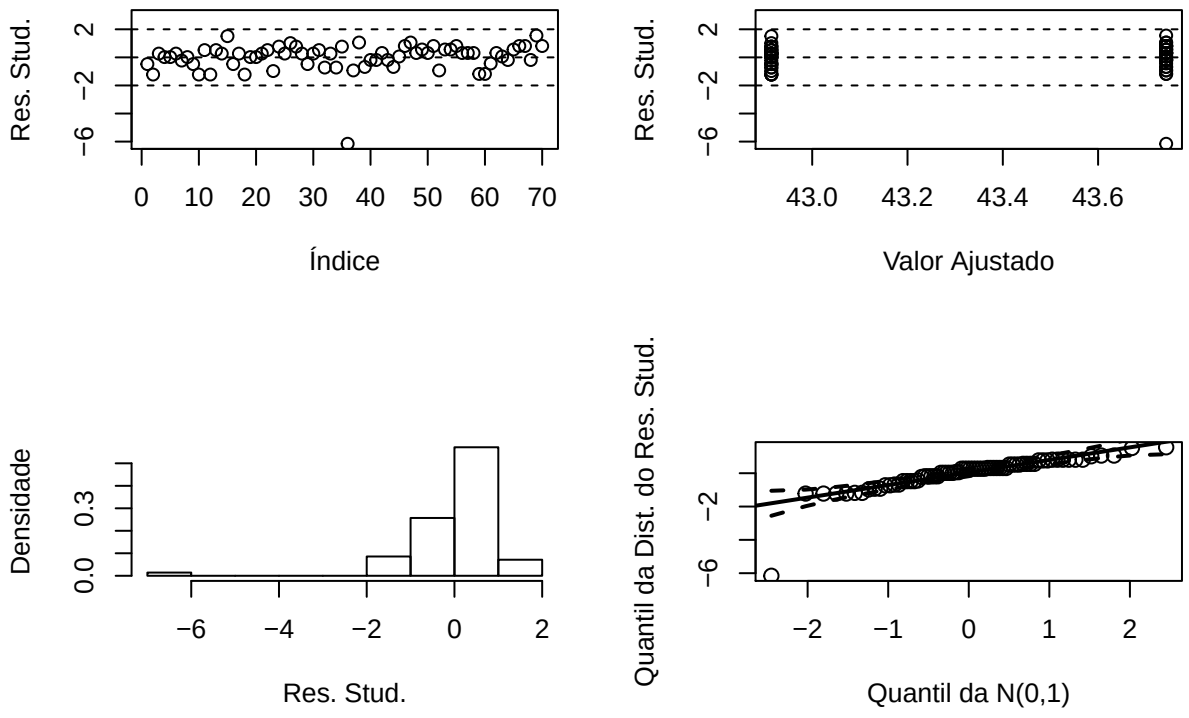


Figura 5: Gráfico de Resíduos para a variável L.Asa

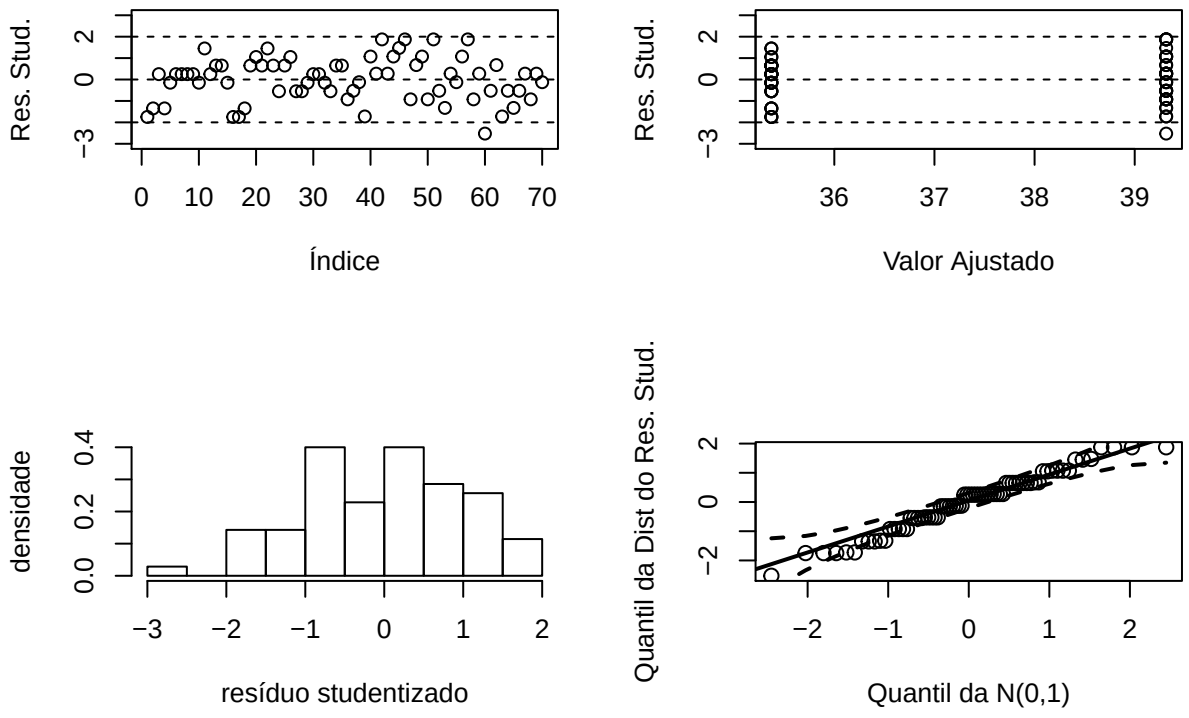


Figura 6: Gráfico de Resíduos para a variável C3p

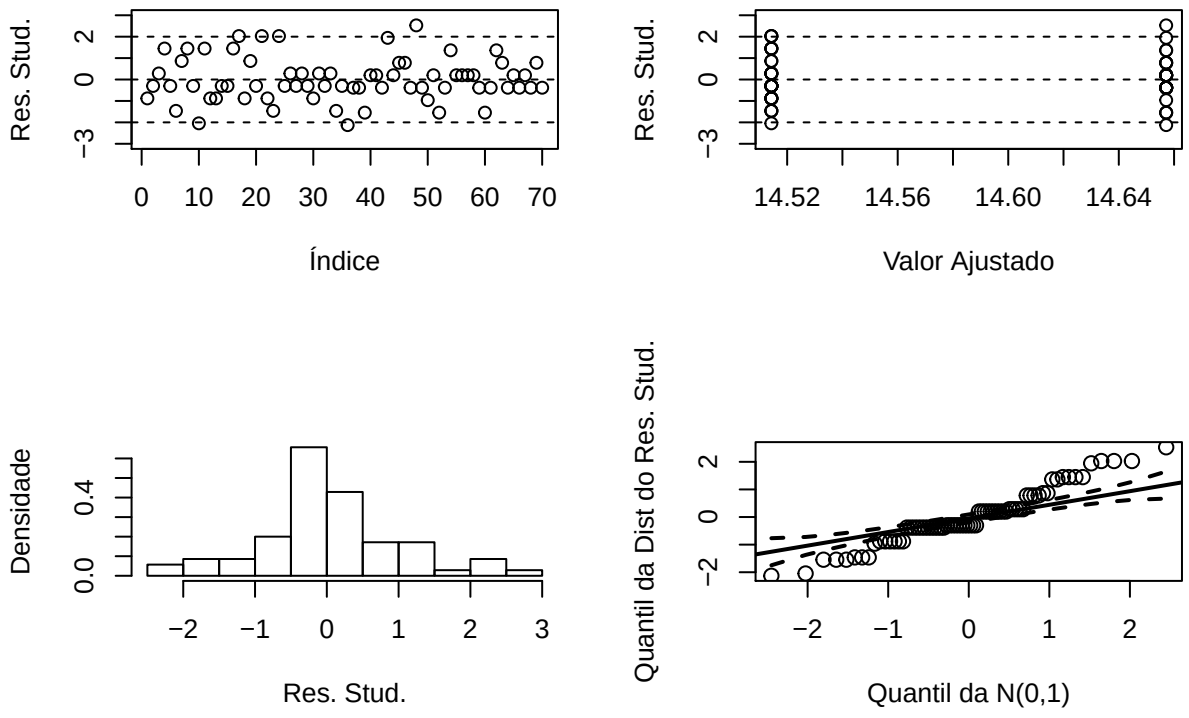


Figura 7: Gráfico de Resíduos para a variável L3p

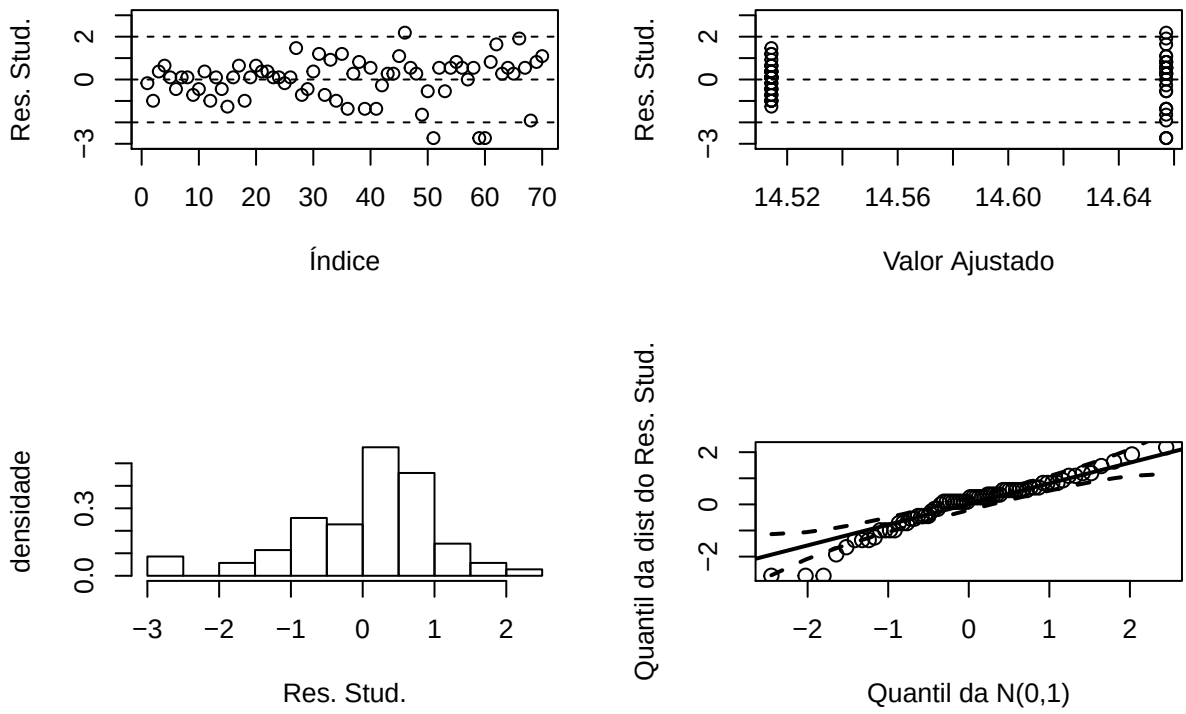


Figura 8: Gráfico de Resíduos para a variável C4p

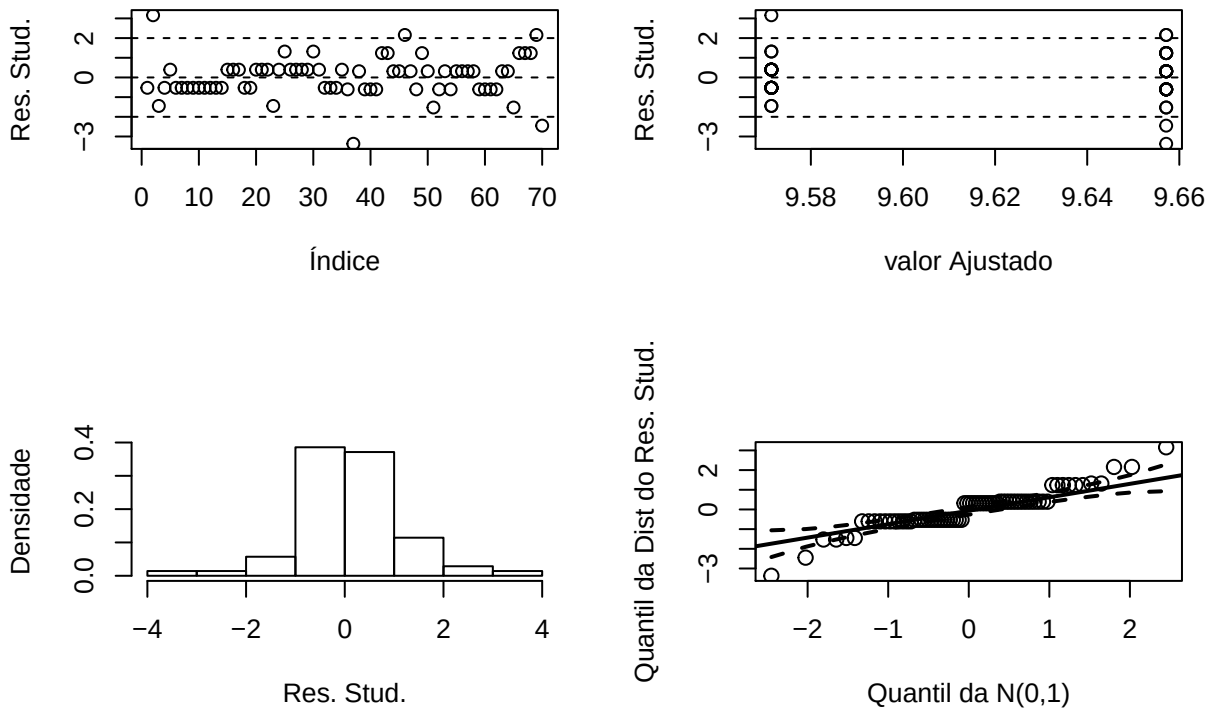


Figura 9: Gráfico de Resíduos para a variável C12a

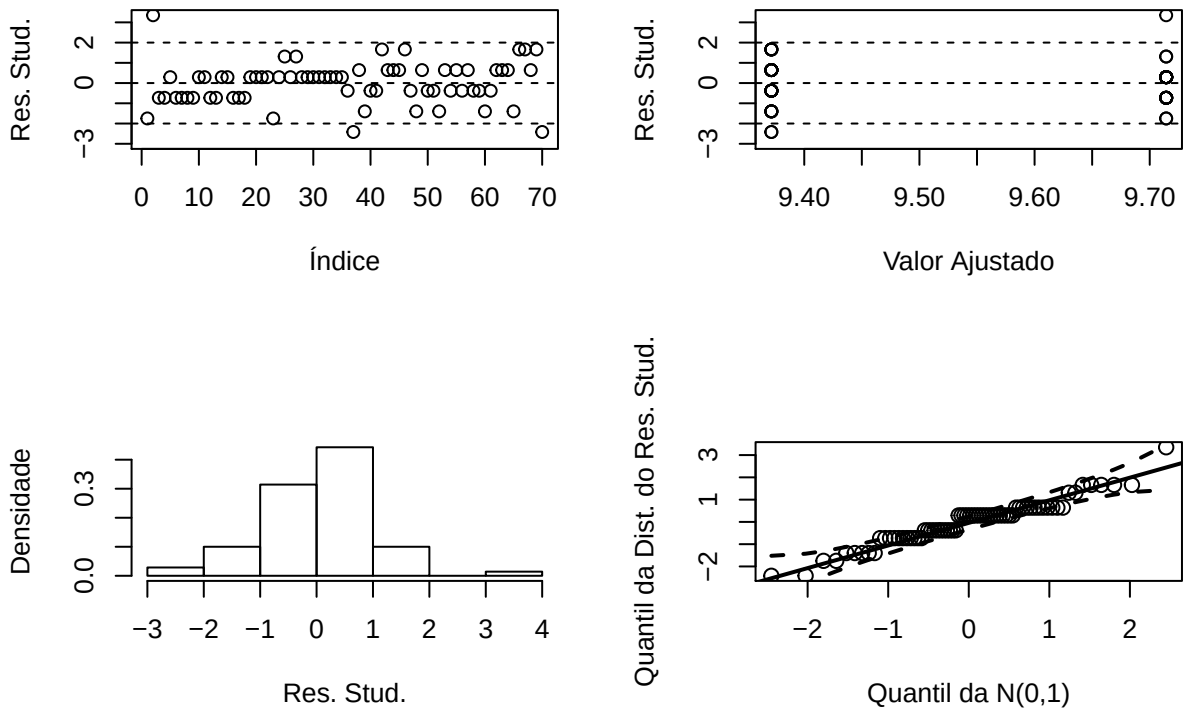


Figura 10: Gráfico de Resíduos para a variável C13a

4 Conclusão

Feitas a devida ressalva de que o modelo escolhido como base para as análises nao se ajustou bem aos dados, e considerando validos os resultados através dele obtidos, concluimos, por meio dos testes de igualdade de médias, que as espécies Torrens e Carteri diferem, em média, nas características Comprimento da Asa, Comprimento do 3º palpo e Comprimento do 4º palpo. Esta conclusão está de acordo com a conjectura feita através da análise descritiva dos dados.

5 Referências

1. AZEVEDO, C. L. N. (2017). **Notas de Aula - Métodos em Análise Multivariada**. Disponível em < http://www.ime.unicamp.br/~cnaber/Material_AM_2S_2017.htm >.
2. JOHNSON, R. A., WICHERN, D. W. (2002). **Applied Multivariate Statistical Analysis**, 5^a edição, Upper Saddle River, NJ: Prentice-Hall.
3. R CORE TEAM (2017). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Viena, Áustria. Disponível em < https://www.R-project.org/ >.
4. **gg_qq.r** (2017). Repositório GitHub, disponível em < https://gist.github.com/rentrop/d39a8406ad8af2a1066c >.