Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

Trabalho - Parte 1
Relatório - Questão 1

Eliane Ramos de Siqueira RA:155233
Guilherme Pazian RA:160323
Henrique Capatto RA:146406
Murilo Salgado Razoli RA:150987

Disciplina: **ME731 - Análise Multivariada**
Professor: **Caio Lucidius Naberezny Azevedo**

Campinas - SP
18 de Novembro de 2017

1. Introdução

O conjunto de dados utilizado neste relatório são relativos a moscas chamadas, em inglês, de “bitting fly”.Neste conjunto de dados foram consideradas no total 70 moscas, sendo 35 da espécie Leptoconops carteri e 35 da Leptoconops torrens. Tais espécies serão tratadas a partir de agora por Carteri e Torrens, respectivamente.

Para cada uma das espécies, foram medidas sete variáveis referentes ao aspecto morfológico das moscas e para cada unidade amostral (mosca), foram medidas oito variáveis, sendo elas: espécie (0 – Torrens e 1 – Carteri), comprimento da asa (CA), largura da asa(LA), comprimento do 3º palpo (CP3), largura do 3º palpo (LP3), comprimento do 4º palpo (CP4), comprimento do 12º segmento da antena (SA12) e comprimento do 13º segmento da antena (SA13).

Dado que as duas espécies são bastante semelhantes morfologicamente (Johson e Wichern (2007)), o objetivo desta análise é realizar a comparação das médias de todas as variáveis consideradas entre as espécies, com o intuito de verificar se e quais variáveis diferem entre os grupos.

Como abordagem inicial, será utilizada a Análise de Variância Multivariada (MANOVA) (veja mais em Johnson e Wichern (2007)), para verificar a possível existência de diferenças entre as médias. Posteriormente, caso a hipótese de igualdade entre as médias for rejeitada serão realizados testes do tipo $CBU = M$ (veja mais em Azevedo (2017)), com o intuito de descobrir onde essas diferenças se encontram.

Todas as análises serão realizadas com o suporte dos softwares R versão 3.4.2 e R Studio versão 1.1.383

2. Análise descritiva

A tabela 1 mostra algumas medidas resumo para as todas variáveis citadas anteriormente, sepadas por espécie. É possível notar que as médias amostrais para as variáveis comprimento da asa, comprimento do 3o palpo e do 4o palpo são relativamente diferentes entre as espécies, enquanto que para as demais as médias amostrais são relativamente iguais. Pode se notar também, que os desvios padrões para as variáveis largura da asa e comprimento do 4o palpo aparentemente são diferentes entre as espécies, enquanto que para as demais variáveis os desvios padrão são consideravelmente próximos.

A figura 1 apresenta uma matriz de diagramas de dispersão entre as variáveis, separadas por espécie. De um modo geral, podemos notar que a espécie Torrens apresenta valores inferiores à espécie Carteri, como por exemplo, no gráfico de Largura por Comprimento do 3º Palpo, em que os pontos verser se concentram abaixo dos azuis.

Cada diagrama apresenta seu respectivo valor do coeficiente de correlação linear, então pode-se notar que aparentemente as variáveis que mais se destacaram em relação à uma possível associação linear foram as varáveis referentes ao comprimento do 12º palpo e ao comprimento do 13º palpo (0,81) e as referentes ao comprimento e largura da asa (0,6). Além disso, observam-se também possíveis associações entre as demais variáveis, porém com grau menor de correlação, como por exemplo entre as variáveis referentes ao comprimento da asa e comprimento do 3º palpo (0,45) e entre comprimento da asa e comprimento do 4º palpo (0,48).

Na figura 2, temos os boxplots para todas as variáveis separadas por espécie. Algumas distribuições apresentam ligeiras assimetrias, observando de um modo geral. Por exemplo na variável largura da asa para a espécie Torrens, essa assimetria é um pouco mais evidente. Podemos afirmar que as distribuições quando comparadas entre as espécies são diferentes observando de um modo geral. É possível afirmar que as medianas são ligeiramente maiores para todas a variáveis na espécie Carteri, exceto para a variável referente ao comprimento do 13º segmeno da antena. Por fim, nota-se a presença de alguns pontos discrepantes, sendo que em maioria das vezes, estes pontos estão mais presentes na espécie Carteri.

A figura 3 apresenta o gráfico de quantis-quantis com envelopes, para a distância de Mahalanobis (Azevedo(2017)), para cada espécie. Pode-se observar que a suposição de normalidade multivariada dos dados parece não ser razoável, uma vez que há grandes fugas para quantis maiores da distribuição F na parte superior dos gráficos.

Tabela 1: Medidas Resumo das variáveis por espécie

	Espécie	n	Média	Variância	Desvio Padrão	CV(%)	Mínimo	Mediana	Máximo
Comprimento da Asa	Carteri	35	96,457	40,726	6,382	6,616	85	95	109
	Torrens	35	99,343	31,291	5,594	5,631	82	99	112
Largura da Asa	Carteri	35	42,914	7,492	2,737	6,378	38	44	49
	Torrens	35	43,743	25,785	5,078	11,608	19	45	50
Comprimento 3º Palpo	Carteri	35	35,371	4,829	2,197	6,212	31	36	39
	Torrens	35	39,314	8,045	2,836	7,215	33	39	44
Largura 3º Palpo	Carteri	35	14,514	3,375	1,837	12,657	11	14	18
	Torrens	35	14,657	2,703	1,644	11,216	11	15	19
Comprimento 4º Palpo	Carteri	35	25,629	6,24	2,498	9,747	21	26	31
	Torrens	35	30,00	21,294	4,615	15,382	20	31	38
Comprimento 12° Seg antena	Carteri	35	9,571	0,84	0,917	9,577	8	9	13
	Torrens	35	9,657	1,585	1,259	13,036	6	10	12
Comprimento 13° Seg antena	Carteri	35	9,714	0,798	0,893	9,198	8	10	13
	Torrens	35	9,371	1,182	1,087	11,599	7	9	11

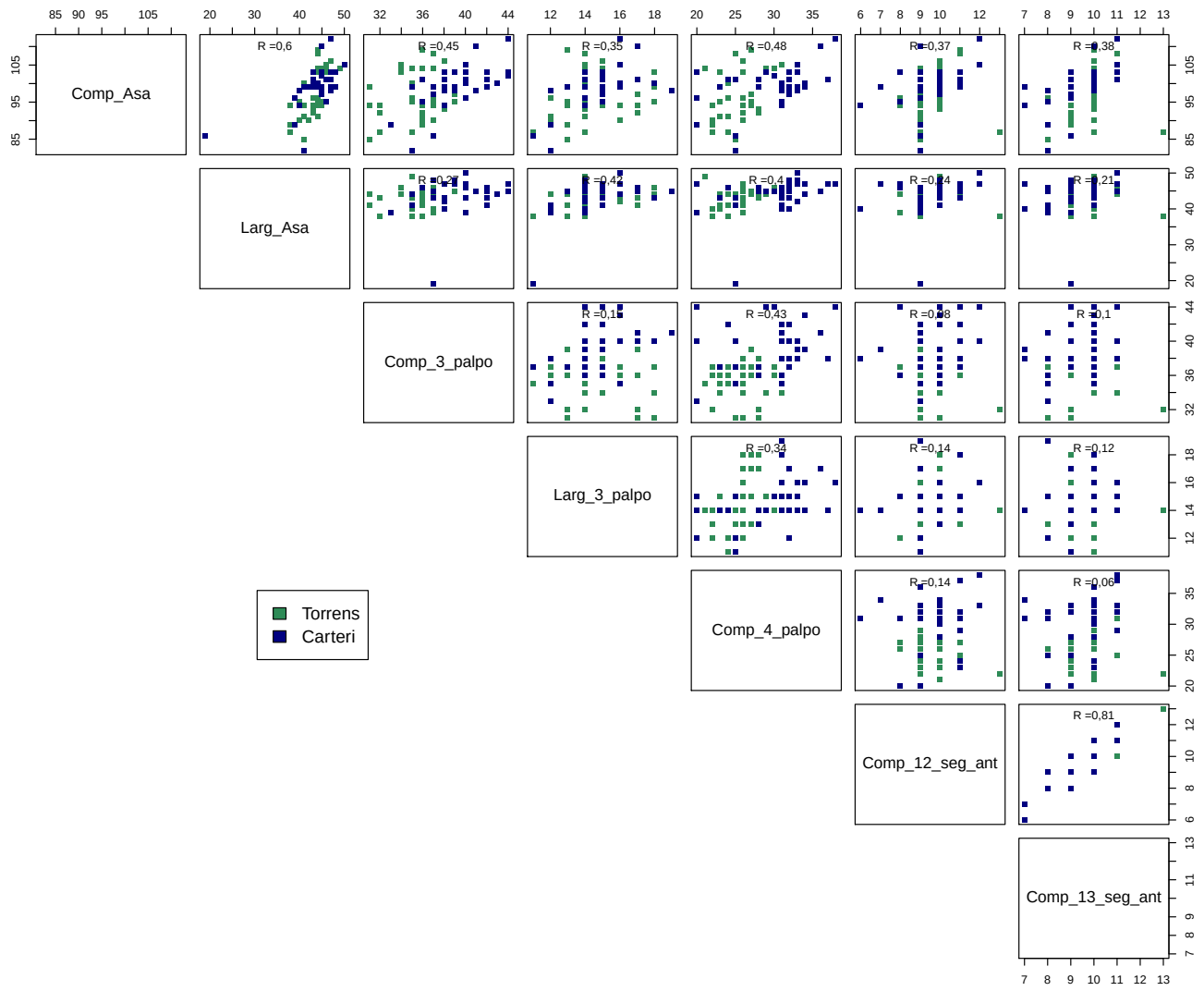


Figura 1: Matriz de diagramas de dispersão entre as variáveis

3. Análise Inferencial

Uma suposição adotada para se realizar um teste de análise de variancia multivariada é a de igualdade das matrizes de covariancia

Com o objetivo de identificar as diferenças entre as espécies de moscas, ajustou um modelo de regressão normal linear homocedástico multivariado ajustado via mínimos quadrados generalizados (veja Azevedo,2017):

$$Y_{ijk} = \mu_k + \alpha_{ik} + \varepsilon_{ijk},\ \varepsilon_{ijk} \sim N_k(0, \Sigma),$$

$i = 1,2$ (espécie, 1 - Leptoconops torrens,2 - Leptoconops carteri), $j = 1,2,...,35$ (moscas) e

$k = 1,...,7$ (variável, 1 - Comprimento da Asa, 2 - Largura da Asa, 3 - Comprimento 3° palpo, 4 - Largura 3° palpo,

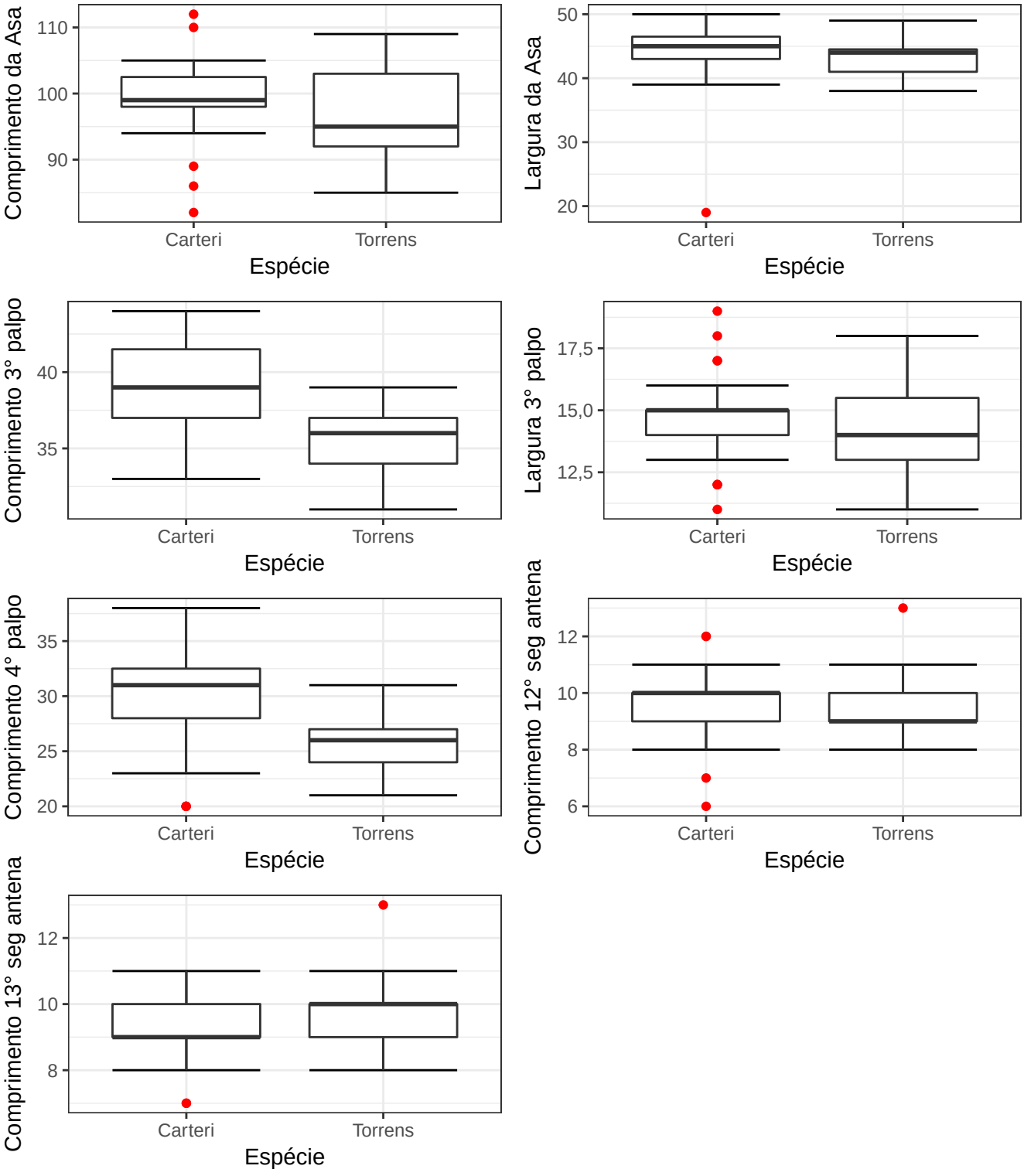


Figura 2: Box-plot das variáveis por espécie

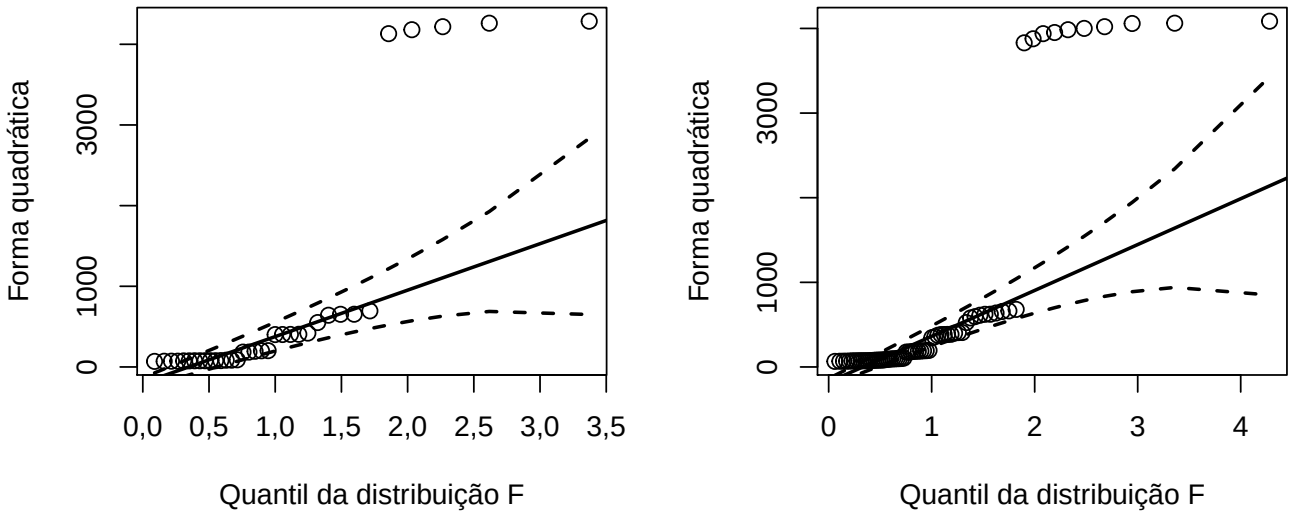


Figura 3: Gráfico de quantil-quantil com envelopes para a distância de Mahalanobis; A) Torrens, B) Carteri

5 - Comprimento 4° palpo, 6 - Comprimento 12° segmento da antena, 7 - Compimento 13° segmento da antena),

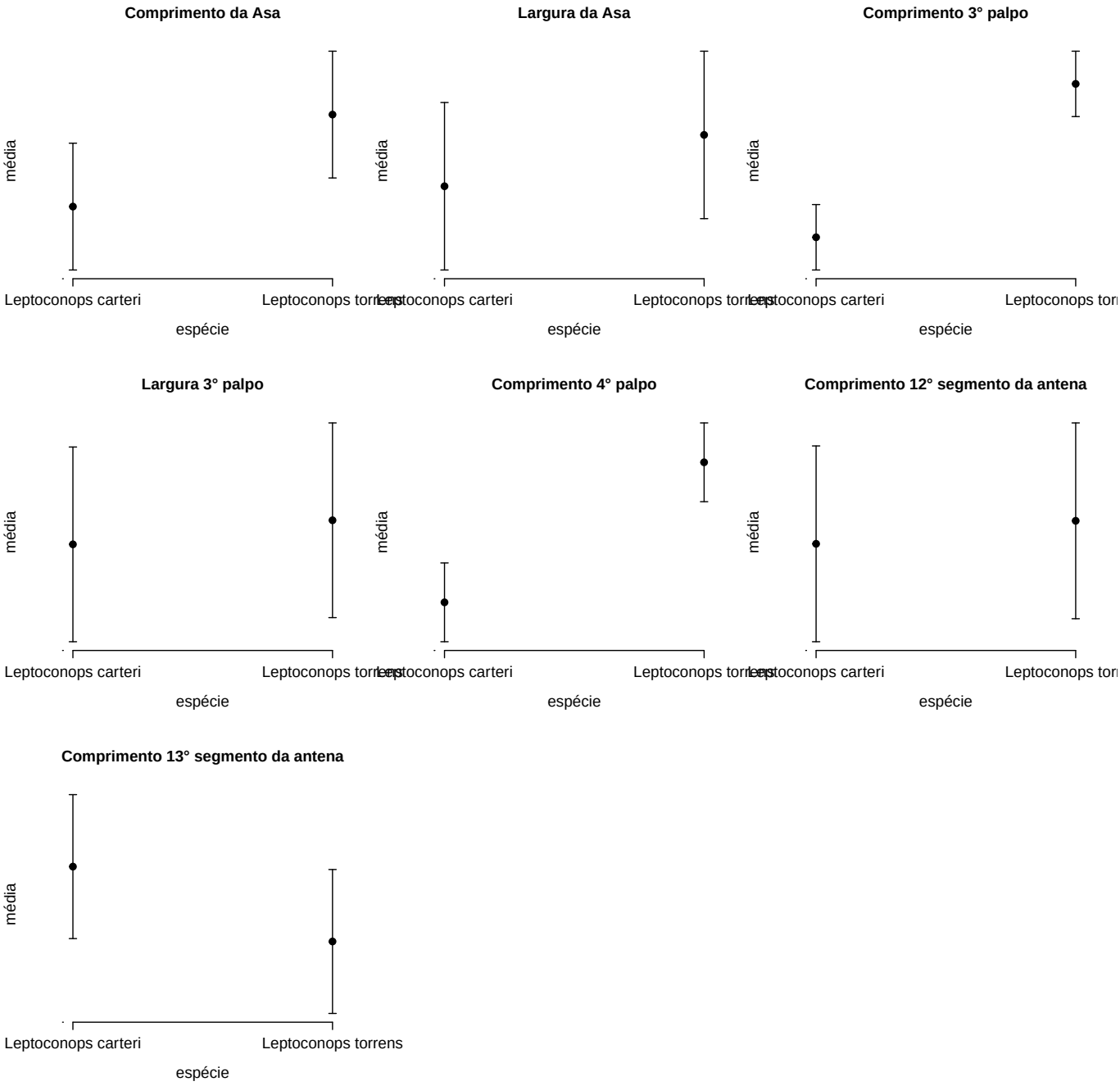
em que $\alpha_{1k} = 0, k = 1, ..., 7$.

	Estatística	Valor	Aprox..distr..F	p.valor
1	Wilks	0,391	13,824	<0,001
2	Pillai	0,609	13,824	<0,001
3	Hotelling-Lawley	1,561	13,824	<0,001
4	Roy	1,561	13,824	<0,001

Na tabela XX(tbl_resultados_MANOVA) estão apresentadas as quatro estatísticas referentes ao teste de análise de variância multivariada:

TABELA tbl_resultados_MANOVA

Note pela tabela XX(tbl_resultados_MANOVA) que todas as estatísticas apresentaram p-valor < 0,05 (Melhorar AQUI) portanto, todas os testes nos apresentam evidencias estatísticamente significativas de que as espécies difiram em ao menos uma das variáveis presentes no banco de dados.



Note pelos gráficos da figura (X-1) que os Intervalos de Confiança para as médias preditas para as espécies de moscas se interceptam num intervalo grande para as variáveis Largura da Asa, Largura 3° palpo e Comprimento 12° segmento da antena, portanto é razoável conjecturar que as espécies de moscas tem médias iguais para estas variáveis. Por meio da metodologia CBU=M (veja Azevedo,2017), testou-se simultaneamente a igualdade das médias destas variáveis entre as espécies de moscas, ao qual resultou num p-valor = 0,864, ou seja, não temos evidencias estatísticas suficientes para rejeitar a hipótese de igualdade simultânea das médias entre as espécies para as variáveis Largura da Asa, Largura 3° palpo e Comprimento 12° segmento da antena. A fim de identificar melhor onde residem as diferenças entre as espécies de moscas, aplicamos esta mesma metodologia acrescentando as demais variáveis na hipótese de igualdade (uma de cada vez), os resultados deste teste constam na tabela (n XX

“tabela_CBU”).

```
##      Hipótese
## [1,] "$\alpha_{12} = \alpha_{14} = \alpha_{16} = 0$"
## [2,] "$\alpha_{12} = \alpha_{14} = \alpha_{16} = \alpha_{11} = 0$"
## [3,] "$\alpha_{12} = \alpha_{14} = \alpha_{16} = \alpha_{13} = 0$"
## [4,] "$\alpha_{12} = \alpha_{14} = \alpha_{16} = \alpha_{15} = 0$"
## [5,] "$\alpha_{12} = \alpha_{14} = \alpha_{16} = \alpha_{17} = 0$"
##      p-valor do teste CBU=M
## [1,] "0,864"
## [2,] "0,342"
## [3,] "< 0,001"
## [4,] "< 0,001"
## [5,] "0,029"
```

Colocar aqui a tabela CBU

Note pela tabela (n XX “tabela_CBU”) que os dois primeiros testes indicam a não rejeição da hipótese apresentada, portanto temos evidencias estatísticamente significantes de que as espécies tem médias conjuntamente iguais para as variáveis Largura da Asa,Largura 3° palpo,Comprimento 12° segmento da antena e Comprimento da Asa, ou seja, as diferenças parecem residir nas variáveis Comprimento 3° palpo, Comprimento 4° palpo e Comprimento 13° segmento da antena.

3.1 Análise dos resíduos

X <- é o número do primeiro gráfico de resíduos

A fim de avaliar a validade das suposições de normalidade multivariada dos dados considerando as espécies (consequentemente normalidade univariada) e homocedásticidade multivariada entre as espécies (consequentemente homocedasticidade univariada), podemos observar as figuras (X) a (X+7) que apresentam gráficos para os resíduos studentizados para cada uma das 7 variáveis, assim como a figura (X+8) que apresenta o gráfico de envelopes baseado na distância de Mahalanobis (veja Azevedo, 2017). A partir da observação destes gráficos, pode-se identificar muitos comportamentos e tendencias não esperadas, as quais podemos destacar o comportamento apresentado no gráfico 4 das figuras (X), (X+3), (X+5) e (X+6) tendo muitos pontos fora dos limites das bandas de confiança, nas figuras (X+1), (X+2) parece existira uma pequena tendência nos valores dos resíduos e na figura (X+4) apresenta muitos pontos com quantis baixos fora das bandas de confiança. Adicionalmente, identificamos comportamento assimétrico negativo no gráfico 3 das figuras (X),(X+1),(X+2) e (X+4) e assimetrico positivo apresentando na figura (X+6). Dadas as observações referentes aos gráficos 1 e 4 das figuras (X) a (X+6) temos um forte indício de que a suposição de normalidade não é razoável para nenhuma das variáveis presentes no banco de dados. Observando o gráfico 2 das figuras (X) a (X+7), identificamos evidencias de presença de heterocedásticidade dos nos dados nas figuras (X), (X+2), (X+3) de maneira

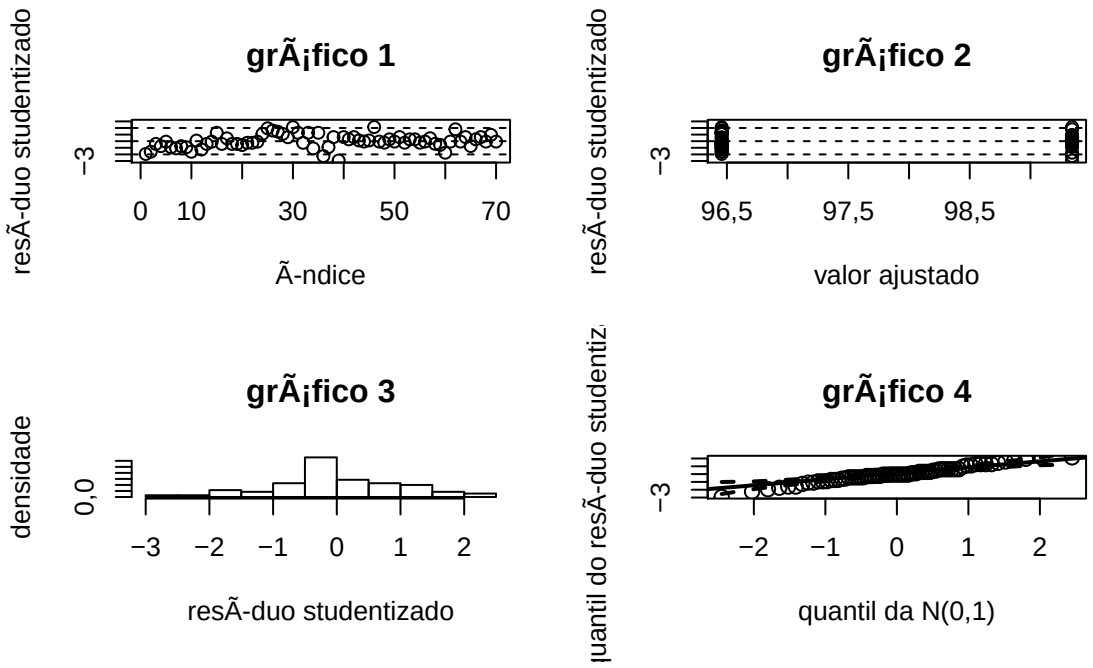


Figura 4: Gráficos para os resíduos referentes à variável Comprimento da Asa

mais leve e nas figuras (X+4),(X+5) e (X+6) de maneira mais acentuada, já para a figura (X+1) não nota-se, a menos de um valor extremo, a presença de indicios de heterocedásticidade. Não identificamos nenhum comportamento a ser destacado referente ao gráfico 1 das figuras (X) a (X+7). Na figura (X+8) observamos alguns valores fora das bandas de confiança para valores menores de quantis da forma quadrática, além disso, valores maiores de quantis da forma quadrática tendem a se apresentar abaixo da linh de referência baseada no quantil da distribuição qui-quadrado, deste modo temos indicações de que a suposição de normalidade multivariada dos dados não parece ser uma suposição razoável neste caso. Contudo, dadas as observações destacadas, temos que a única variável do banco de dados a qual não seria irrasuável supor normalidade e homocedásticidade dos dados seria a variável "Largura da Asa", e todas as restantes apresentam ao menos um indicio evidente da fuga destas suposições portanto não seria razoável supor normalide e homocedásticidade multivariada neste caso, o que também fica evidente na figura (X+8), sendo assim, o modelo de análise de variância multivariada não apresentou um ajuste adequado aos dados aqui analisados e se é necessário procurar técnicas alternativas para realizar uma análise adequada ao banco de dados. Dado o nosso contexto acadêmico, iremos continuar com as análises dos resultados para elaborar a conclução do presente trabalho.

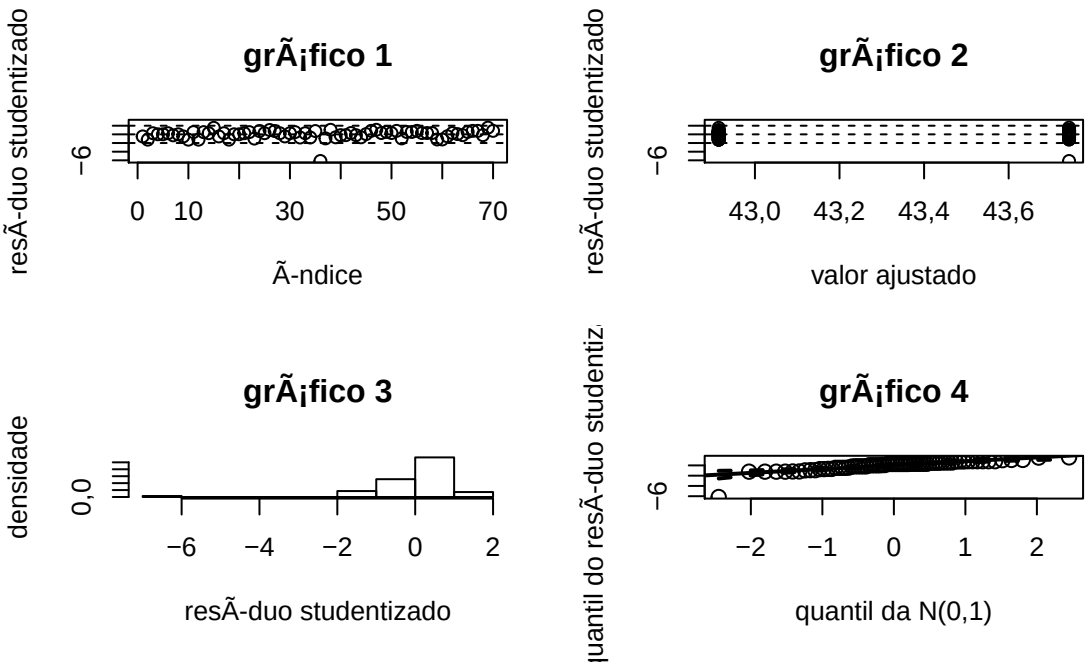


Figura 5: Gráficos para os resíduos referentes à variável Largura da Asa

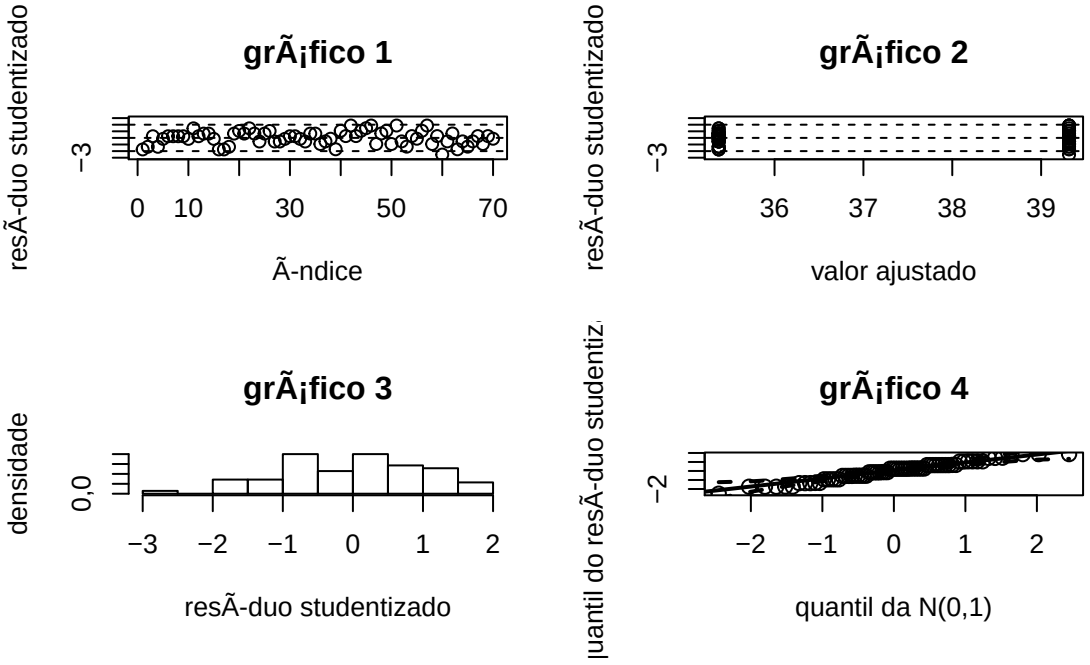


Figura 6: Gráficos para os resíduos referentes à variável Comprimento 3° palpo

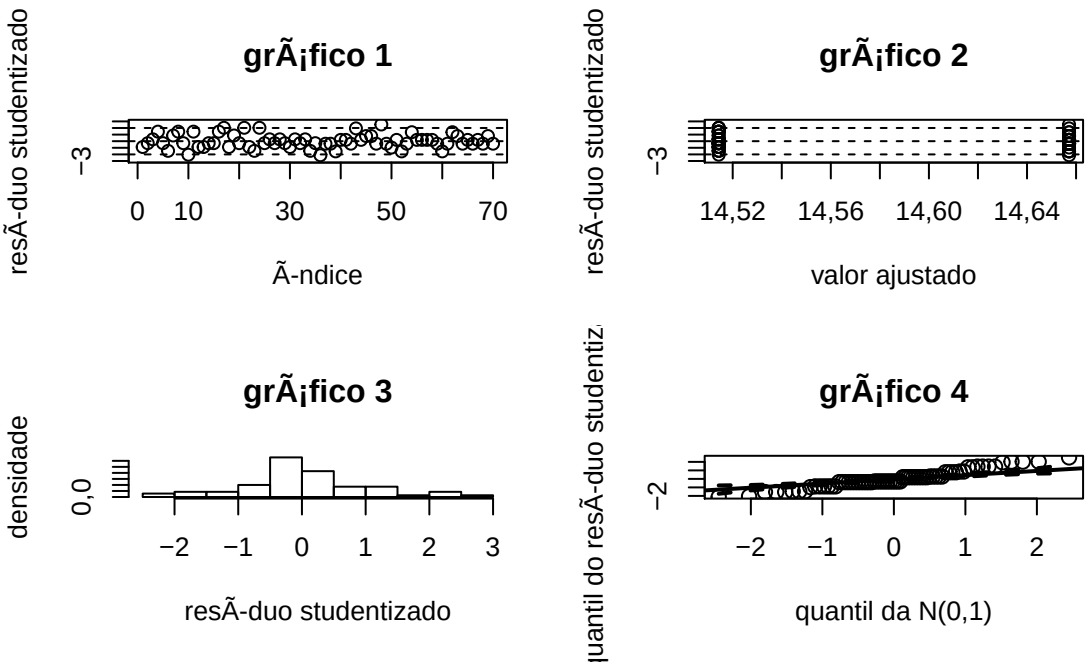


Figura 7: Gráficos para os resíduos referentes à variável Largura 3° palpo

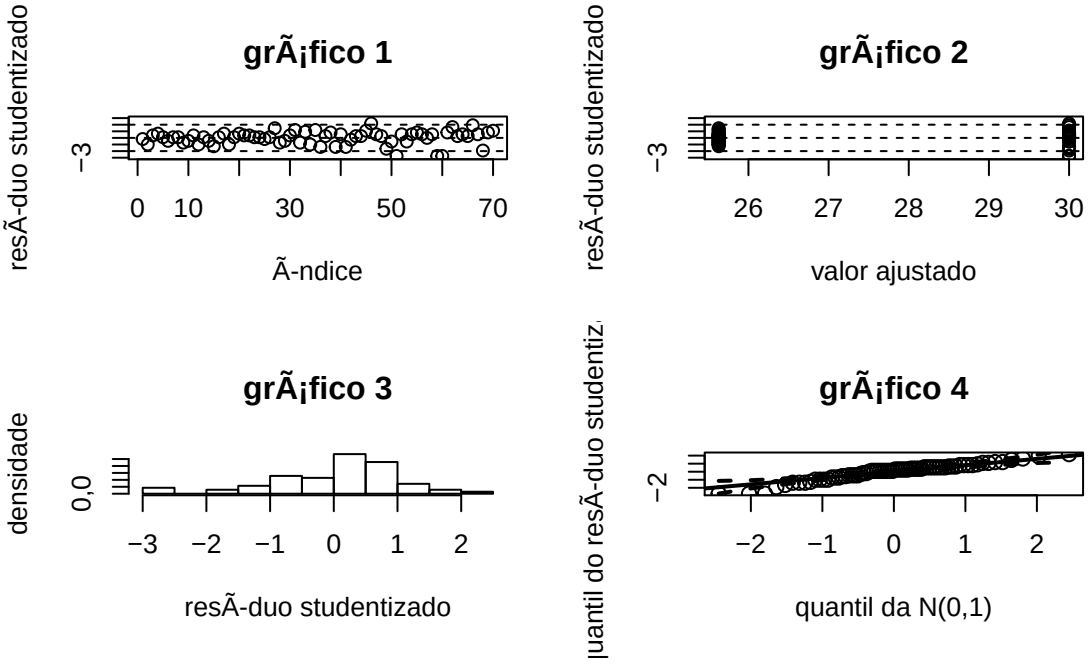


Figura 8: Gráficos para os resíduos referentes à variável Comprimento 4° palpo

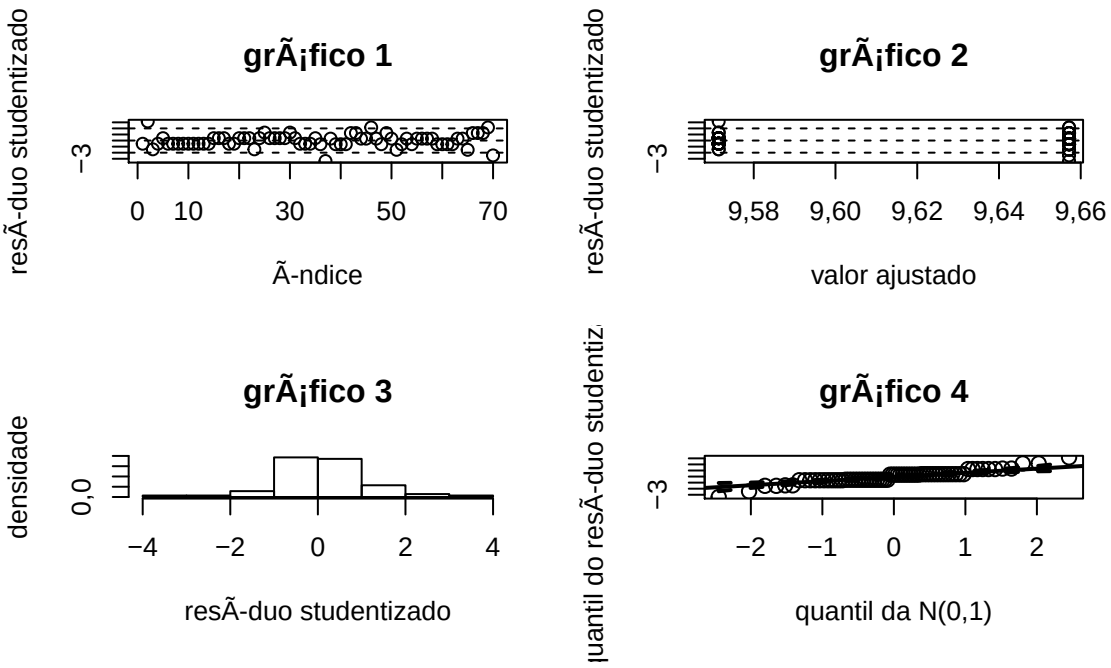


Figura 9: Gráficos para os resíduos referentes à variável Comprimento 12° segmento da antena

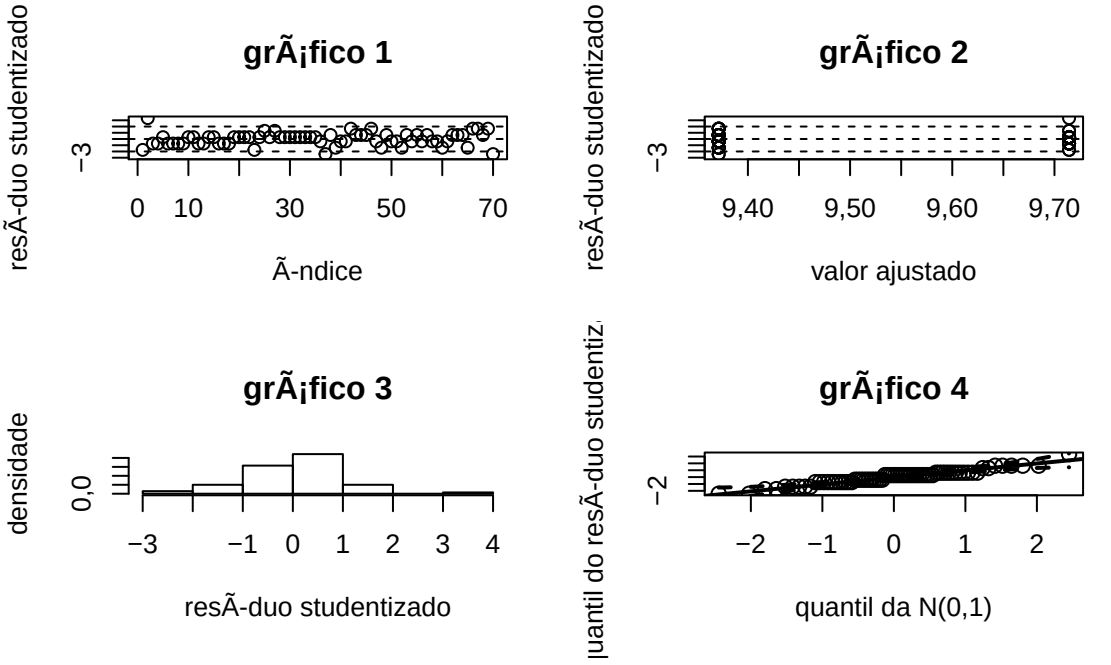
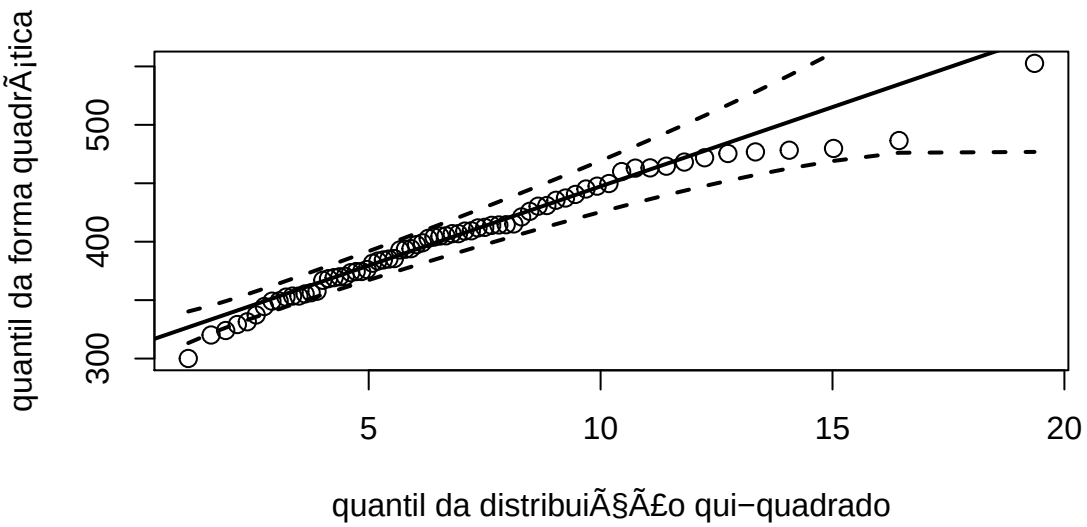


Figura 10: Gráficos para os resíduos referentes à variável Comprimento 13° segmento da antena


quantil da forma quadrÃ¡tica
300
400
500
5
10
15
20
quantil da distribuiÃ§Ã£o qui−quadrado

4. Conclusões

5. Bibliografia