



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 é relativo às 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 morfológicamente (Johnson 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 o vetor de 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.

Foi-se considerado um nível de significância de 5% para a tomar decisões quanto aos testes estatísticos aqui apresentados.

2. Análise descritiva

A tabela 1 mostra algumas medidas resumo para as todas variáveis citadas anteriormente, separadas 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.

Na figura 1, 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 as variáveis na espécie *Carteri*, exceto para a variável referente ao comprimento do 13º segmento 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 2 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 verdes 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 variáveis 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).

Tabela 1: Medidas Resumo das variáveis por espécie

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

3. Análise Inferencial

A figura 3 apresenta o gráfico de envelopes para a distância de Mahalanobis (veja mais em Azevedo (2017)) para os dados de ambas as espécies de moscas, note que existem muitos pontos fora das bandas de confiança para ambas as espécies de moscas, indicando que parece não ser razoável a suposição de normalidade multivariada dos dados referentes à ambas as espécies. Foi realizado o teste de Box para igualdade de matrizes de covariâncias dos dados das duas espécies de moscas, ao qual resultou num p-valor <0,001, indicando que existe diferença estatisticamente significativa entre as matrizes de covariâncias dos grupos de moscas, indicando que não parece ser razoável a suposição de igualdade das matrizes de covariâncias entre as espécies de moscas.

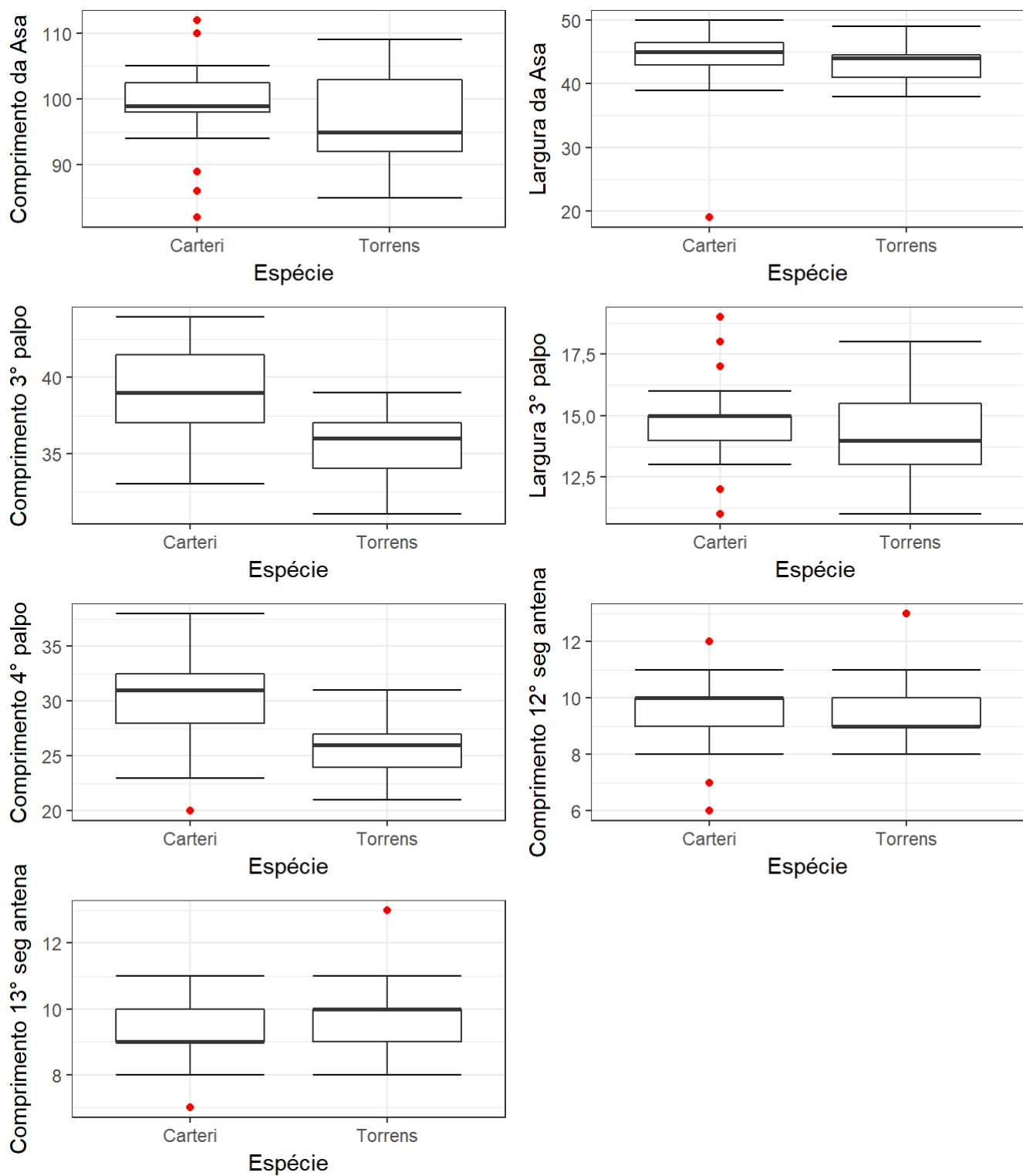


Figura 1: Matriz de diagramas de dispersão entre as variáveis

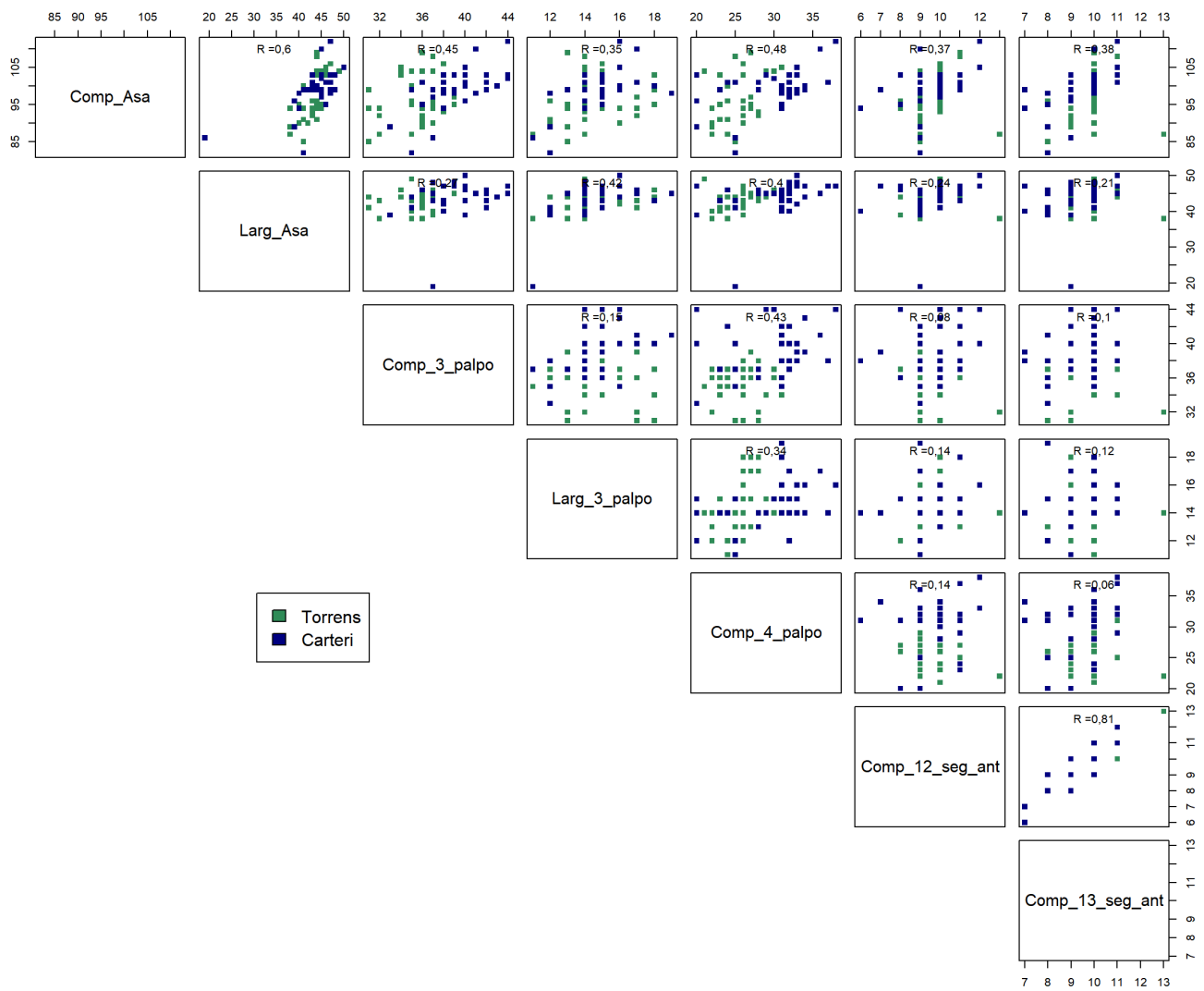


Figura 2: Box-plot das variáveis por espécie

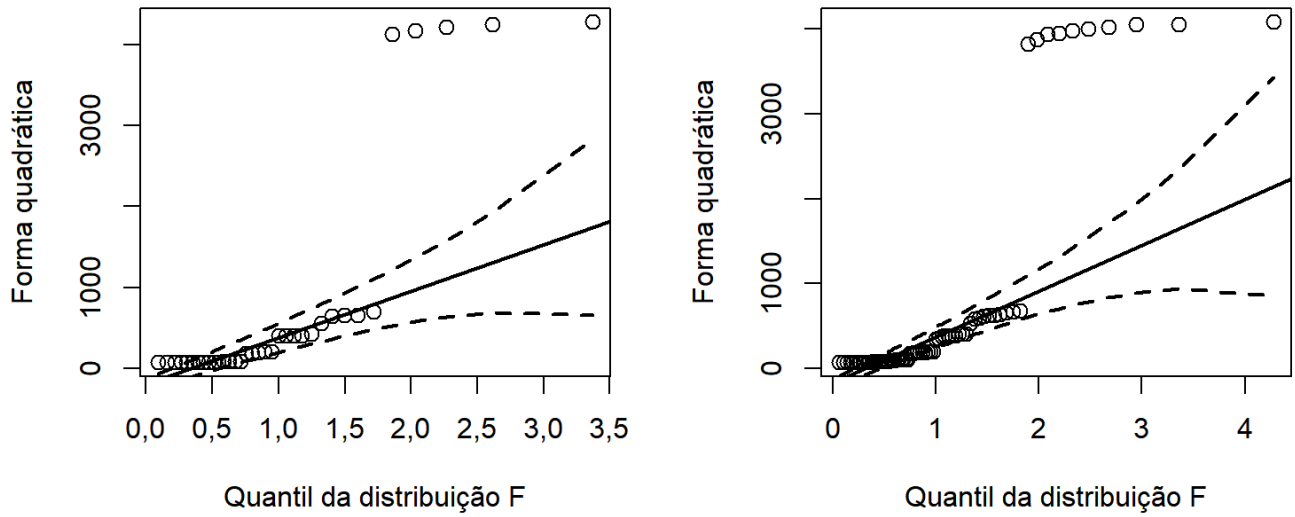


Figura 3: Gráfico de quantil-quantil com envelopes para a distância de Mahalanobis; A) Torrens, B) Carteri

Mesmo não sendo razoáveis as suposições descritas acima, dado o contexto da disciplina, ajustou-se um modelo de regressão normal linear homocedástico multivariado ajustado via mínimos quadrados generalizados (veja mais em Azevedo (2017)), tendo como objetivo identificar as diferenças entre as espécies de moscas:

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

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

$k = 1, \dots, 7$ (variável, 1 - Comprimento da Asa, 2 - Largura da Asa, 3 - Comprimento 3º palpo, 4 - Largura 3º palpo,

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

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

A fim de qualificar o ajuste do modelo proposto, avaliamos as suposições de normalidade multivariada dos erros do modelo considerando as espécies (consequentemente normalidade univariada) e homocedasticidade multivariada entre as espécies (consequentemente homocedasticidade univariada) com base nas figuras 4 a 10, as quais apresentam gráficos para os resíduos studentizados para cada uma das 7 variáveis, e na figura 11 que apresenta o gráfico de envelopes baseado na distância de Mahalanobis (Azevedo, 2017).

A partir da observação destes gráficos, pode-se identificar muitos comportamentos e tendências não esperados, as quais podemos destacar o comportamento apresentado no gráfico 4 das figuras 4, 7, 9 e 10 tendo muitos pontos fora dos limites

das bandas de confiança, nas figuras 5 e 6 onde parece existir uma pequena tendência nos valores dos resíduos e na figura 8 que apresenta muitos pontos com quantis baixos fora das bandas de confiança. Adicionalmente, observa-se comportamento de distribuição com assimetria negativa no gráfico 3 das figuras 4, 5, 6 e 8 e com assimetria positiva na figura 10. Dadas as observações referentes aos gráficos 1 e 4 das figuras 4 a 10 temos fortes indícios de que a suposição de normalidade não é razoável para os erros referentes a nenhum das variáveis presentes no banco de dados.

Observando o gráfico 2 das figuras 4 a 10, identificamos evidências de presença de heterocedasticidade entre os resíduos das espécies de moscas nas figuras 4, 6, 7 de maneira mais leve e nas figuras 8, 9 e 10 de maneira mais acentuada, já para a figura 5 não nota-se, a menos de um valor extremo, a presença de indícios de heterocedasticidade. Portanto, a suposição de igualdade das matrizes de covariância das duas espécies parece não ser razoável.

Não identificamos nenhum comportamento a ser destacado referente ao gráfico 1 das figuras 4 a 10.

Na figura 11 observamos alguns valores fora das bandas de confiança para valores de quantis menores da forma quadrática, além disso, valores de quantis maiores da forma quadrática tendem a se apresentar abaixo da linha de referência que é baseada no quantil da distribuição qui-quadrado, deste modo temos indicações de que a suposição de normalidade multivariada dos erros 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 homocedasticidade dos dados seria a variável “Largura da Asa”, já que todas as variáveis restantes apresentam ao menos um indicio evidente da fuga das suposições. Como não seria razoável supor normalidade e homocedasticidade multivariada neste caso, o modelo de análise de variância multivariada não apresentou um ajuste adequado, 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 conclusão do presente trabalho.

Na tabela 2 estão apresentadas as quatro estatísticas referentes ao teste de análise de variância multivariada realizado com base no modelo proposto.

Tabela 2: Resultados da MANOVA

	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

Note pela tabela 2 que todos os p-valores referentes às quatro estatísticas são menores que o nível de significância adotado, portanto, todas os testes apresentam evidências estatisticamente significativas de que as espécies difiram em ao menos uma das

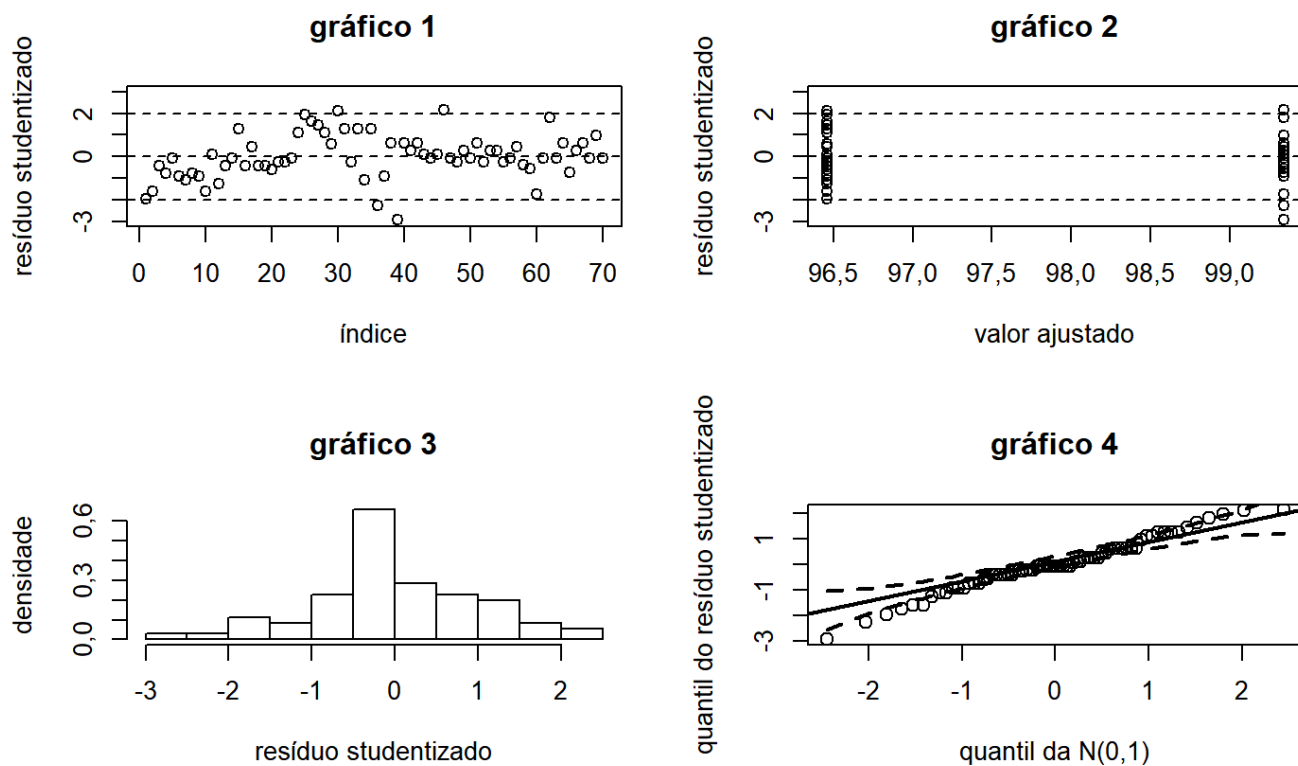


Figura 4: Gráficos para os resíduos referentes à variável Comprimento da Asa

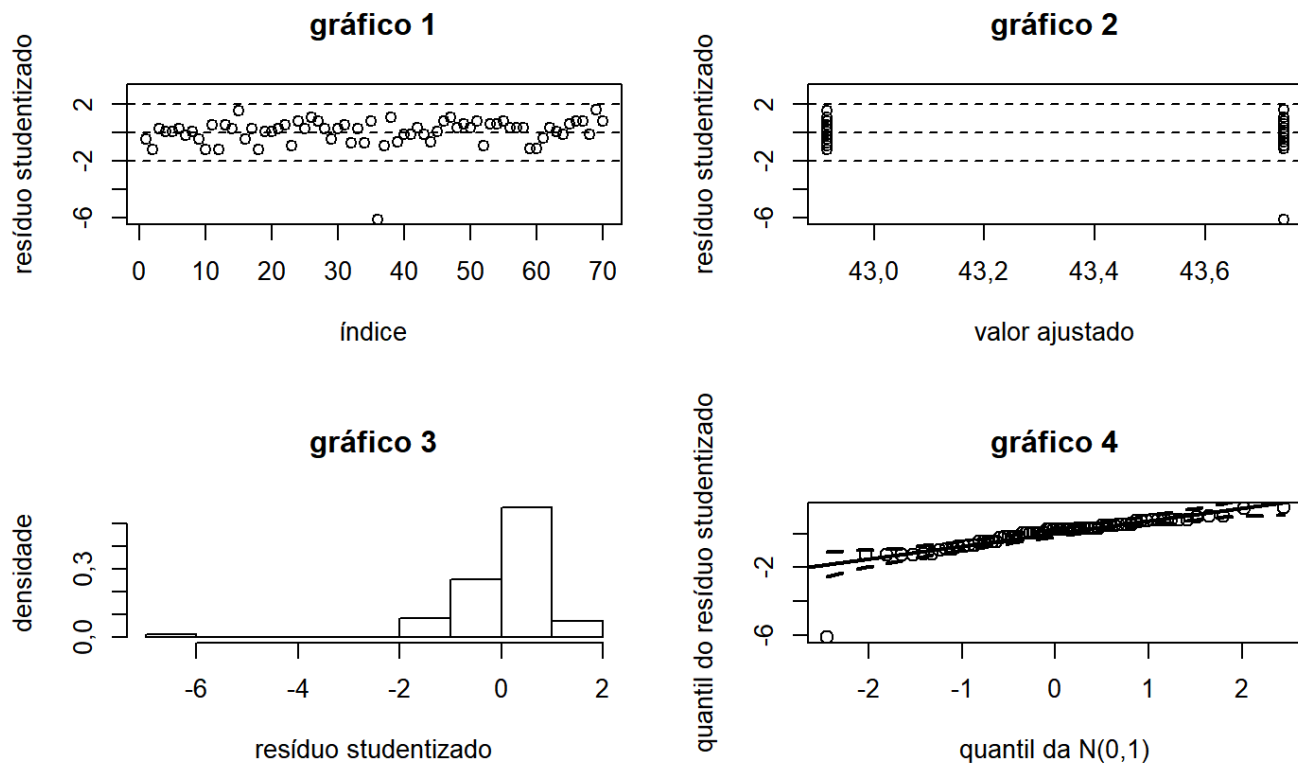


Figura 5: Gráficos para os resíduos referentes à variável Largura da Asa

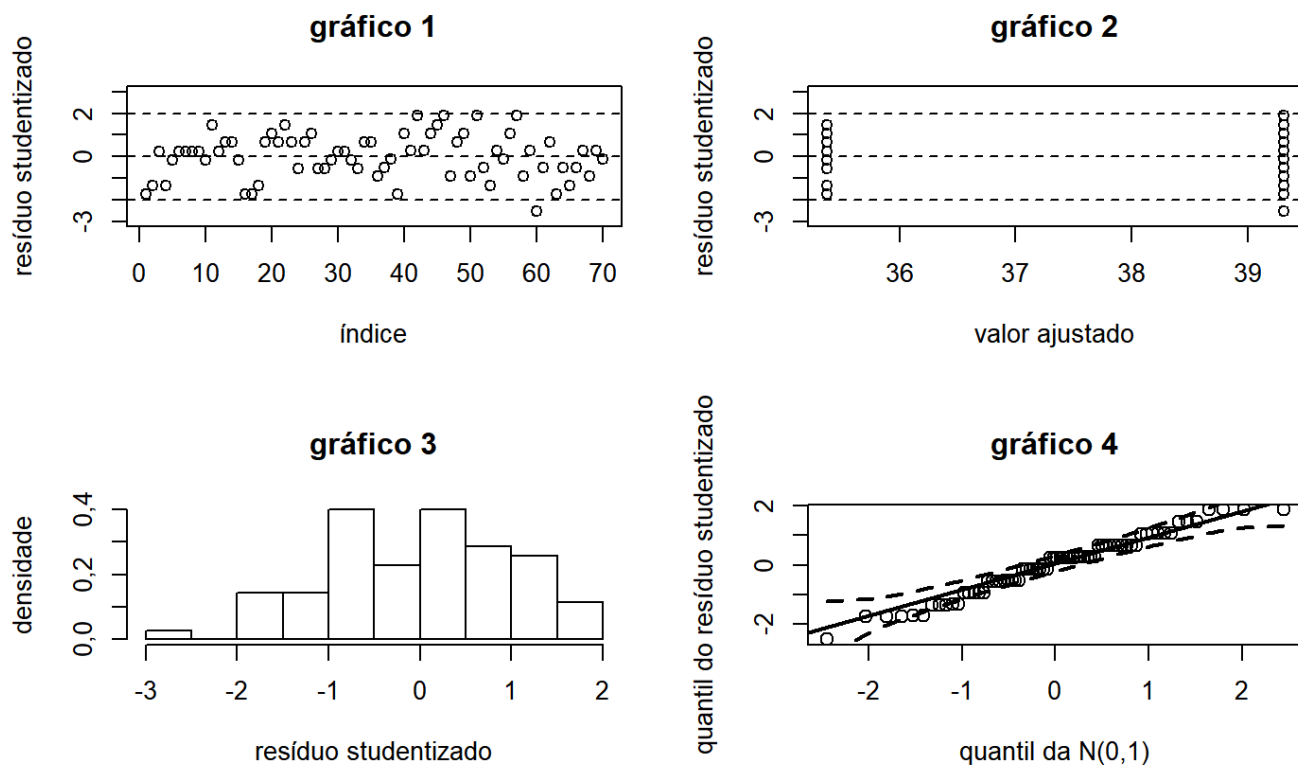


Figura 6: Gráficos para os resíduos referentes à variável Comprimento 3º palpo

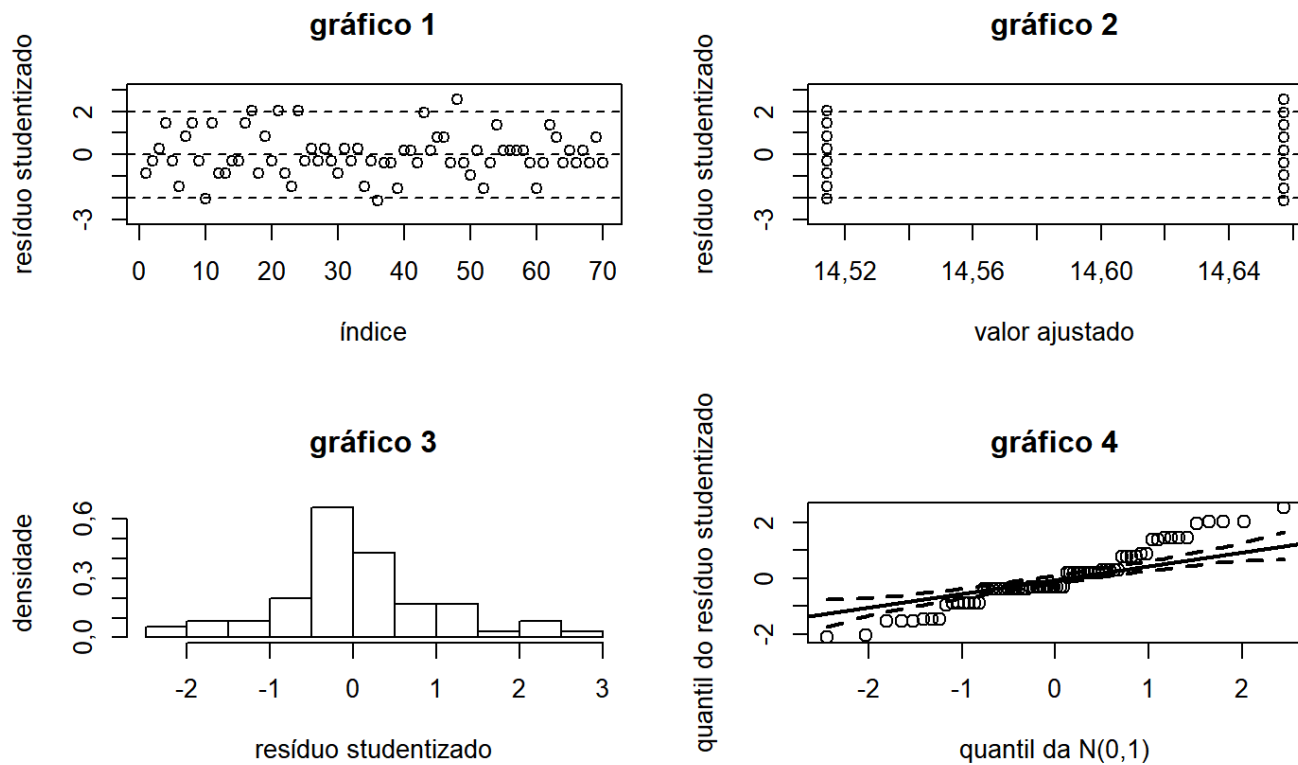


Figura 7: Gráficos para os resíduos referentes à variável Largura 3º palpo

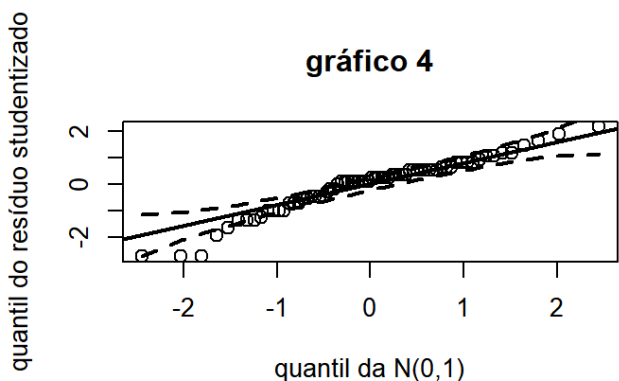
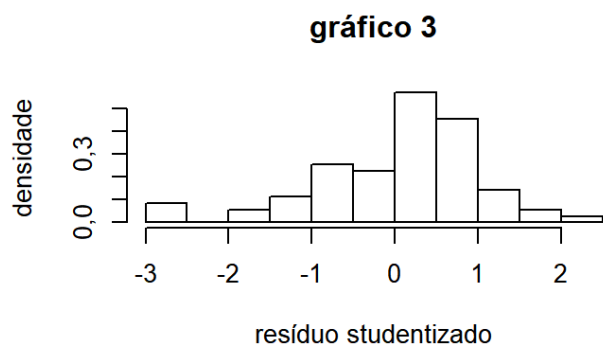
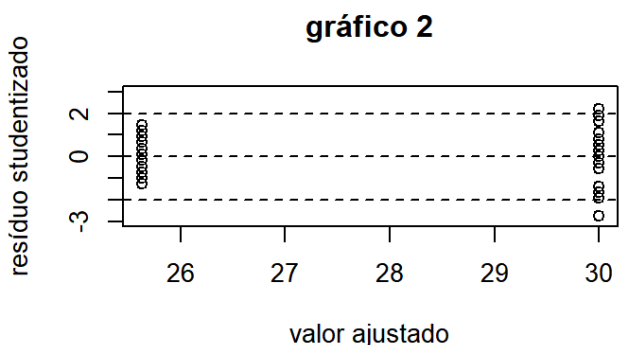
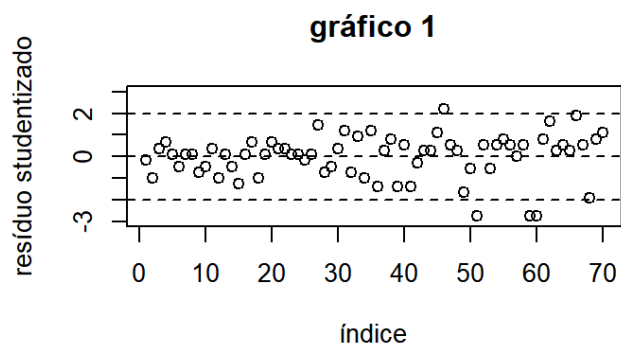


Figura 8: Gráficos para os resíduos referentes à variável Comprimento 4° palpo

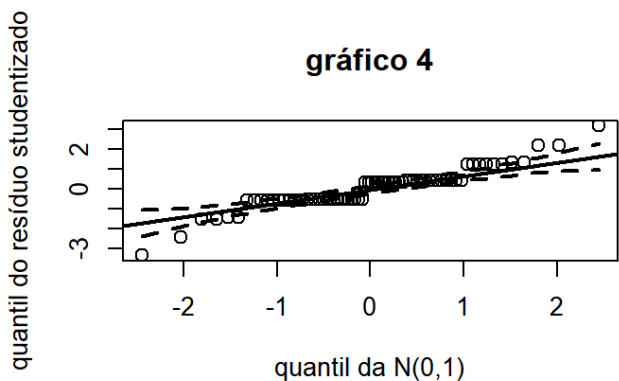
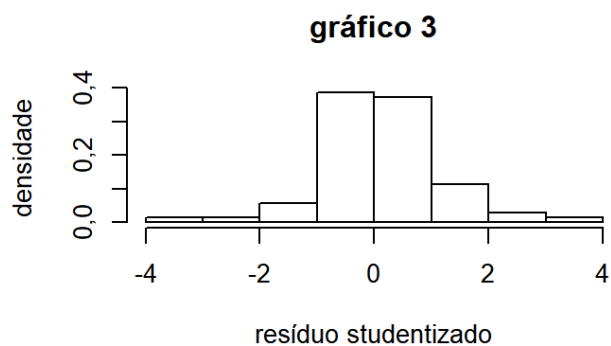
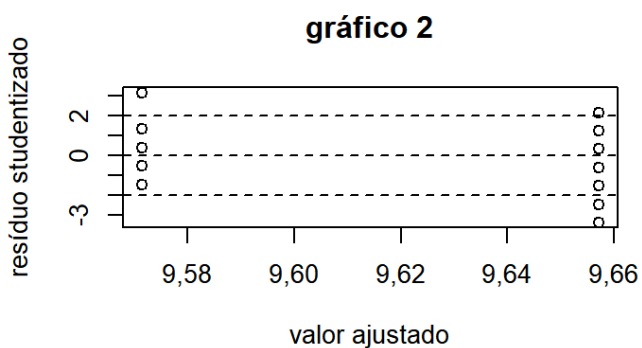
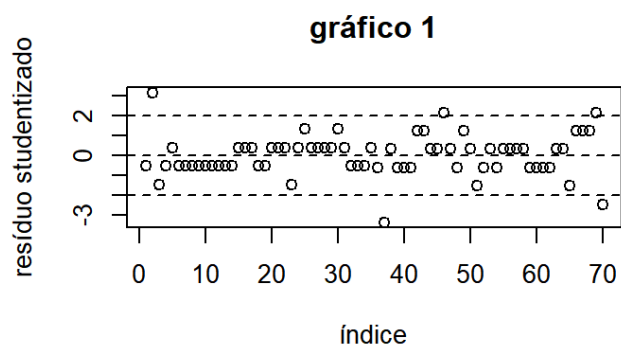


Figura 9: Gráficos para os resíduos referentes à variável Comprimento 12° segmento da antena

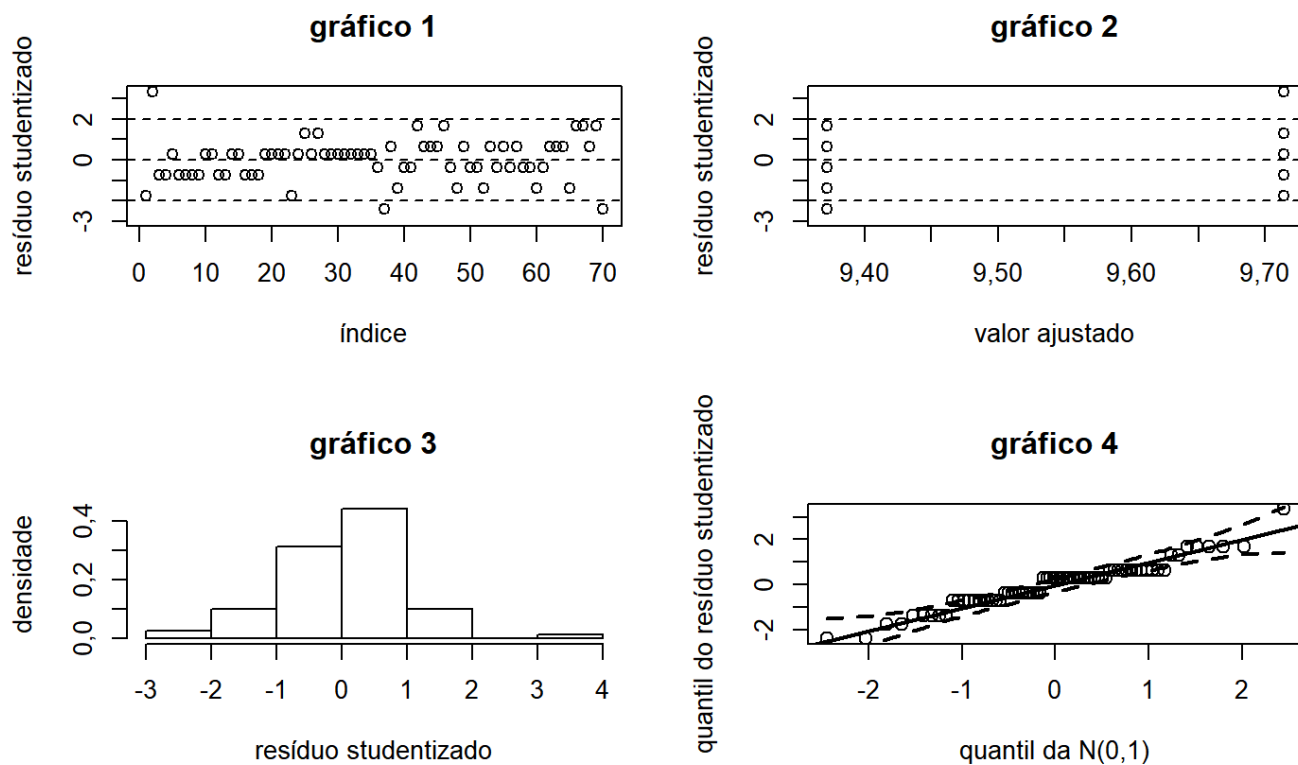


Figura 10: Gráficos para os resíduos referentes à variável Comprimento 13º segmento da antena

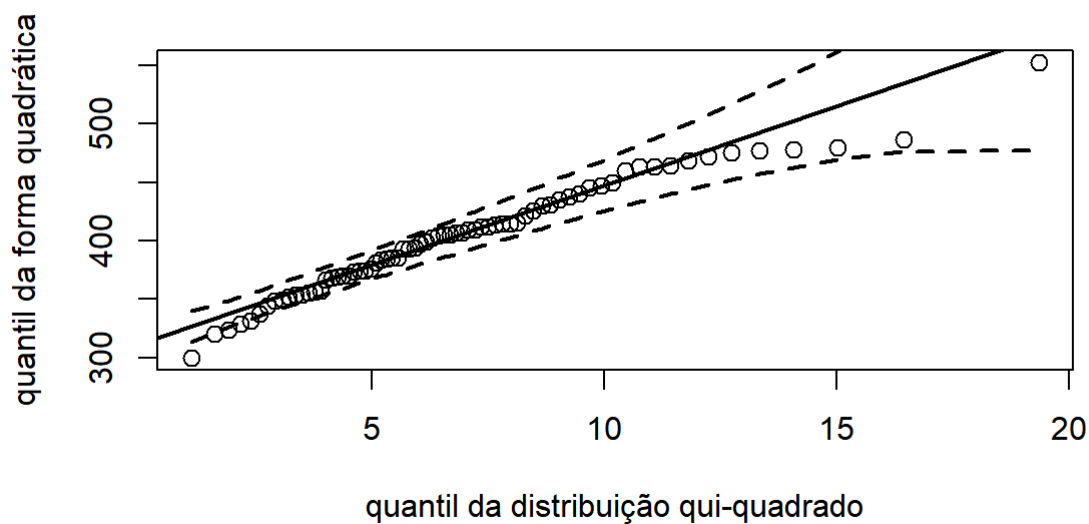


Figura 11: Gráfico de quantil-quantil com envelopes para a distância de Mahalanobis referente aos resíduos do modelo ajustado

variáveis presentes no banco de dados.

A tabela 3 apresenta as estimativas, os erros padrão e os respectivos testes de nulidade para os parâmetros ajustados no modelo.

Tabela 3: Estimativas dos parâmetros do modelo de regressão

Parâmetro	Estimativa	Erro Padrão	Estatística t	p-valor
μ_1	96,457	1,014	95,097	<0,001
α_{21}	2,886	1,434	2,012	0,048
μ_2	42,914	0,689	62,241	<0,001
α_{22}	0,829	0,975	0,85	0,398
μ_3	35,371	0,429	82,479	<0,001
α_{23}	3,943	0,606	6,501	<0,001
μ_4	14,514	0,295	49,259	<0,001
α_{24}	0,143	0,417	0,343	0,733
μ_5	25,629	0,627	40,863	<0,001
α_{25}	4,371	0,887	4,929	<0,001
μ_6	9,571	0,186	51,422	<0,001
α_{26}	0,086	0,263	0,326	0,746
μ_7	9,714	0,168	57,762	<0,001
α_{27}	<0,001	0,238	<0,001	0,15

Note pelos gráficos da figura 12 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 3º palpo e Comprimento 12º 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 evidências estatísticas suficientes para rejeitar a hipótese de igualdade simultânea das médias entre as espécies para as variáveis Largura da Asa, Largura 3º palpo e Comprimento 12º 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 4.

Note pela tabela 4 que os dois primeiros testes indicam a não rejeição da hipótese apresentada, portanto temos evidências estatisticamente significantes de que as espécies tem médias conjuntamente iguais para as variáveis Largura da Asa, Largura 3º palpo, Comprimento 12º segmento da antena e Comprimento da Asa, ou seja, as diferenças parecem residir nas variáveis Comprimento 3º palpo, Comprimento 4º palpo e Comprimento 13º segmento da antena.

4. Conclusões

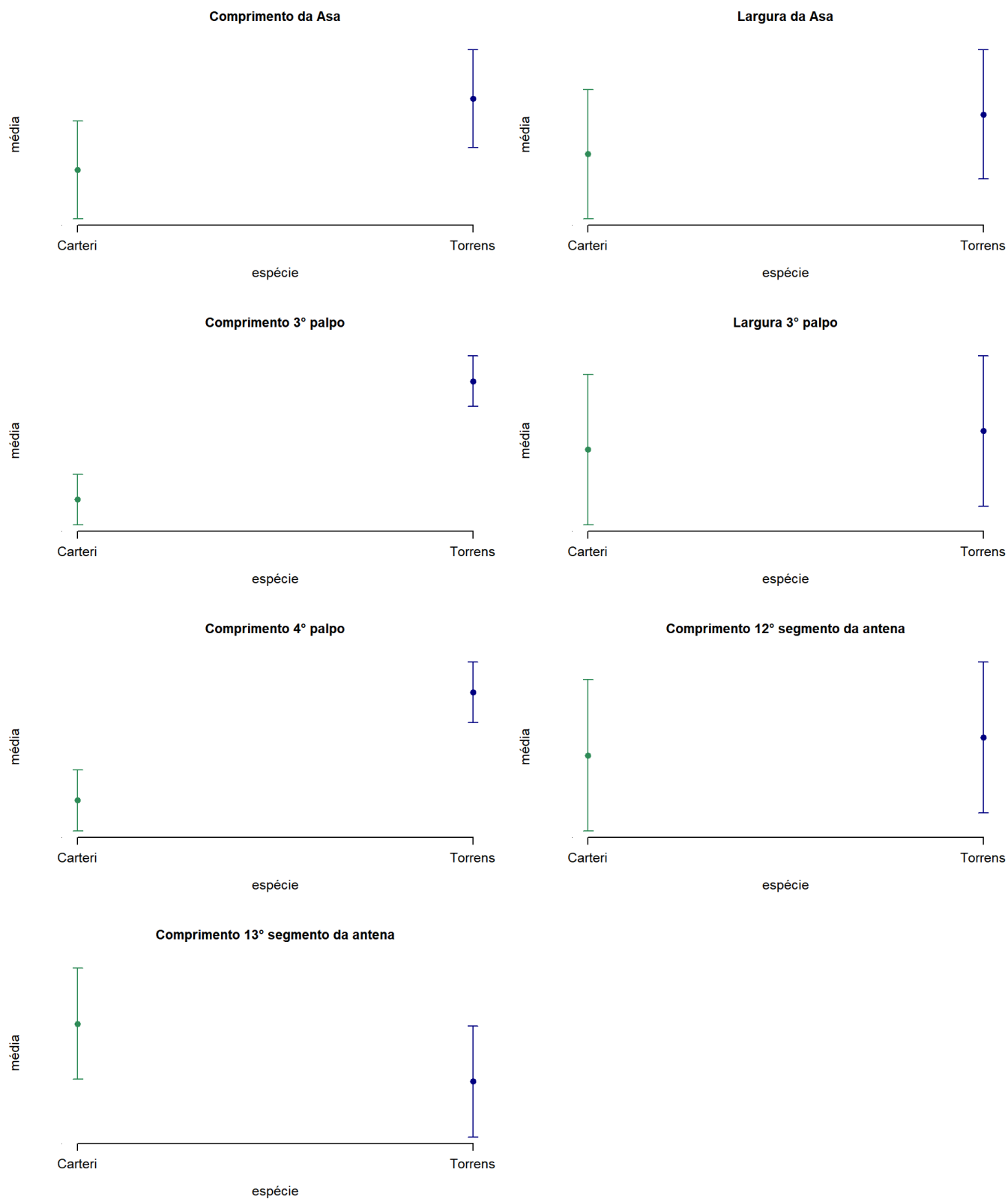


Figura 12: Médias previstas pelo modelo e respectivos intervalos de confiança para cada variável

Tabela 4: Testes CBU = M

Hipótese	p-valor do teste CBU=M
$\alpha_{12} = \alpha_{14} = \alpha_{16} = 0$	0,864
$\alpha_{12} = \alpha_{14} = \alpha_{16} = \alpha_{11} = 0$	0,342
$\alpha_{12} = \alpha_{14} = \alpha_{16} = \alpha_{13} = 0$	< 0,001
$\alpha_{12} = \alpha_{14} = \alpha_{16} = \alpha_{15} = 0$	< 0,001
$\alpha_{12} = \alpha_{14} = \alpha_{16} = \alpha_{17} = 0$	0,029

Com base nas informações descritas neste relatório, podemos concluir que o modelo multivariado ajustado não teve um ajuste adequado, portanto os resultados apresentados referentes ao modelo proposto podem não representar o comportamento dos dados de maneira adequada, porém estes resultados parecem ser razoáveis, uma vez que concordam em boa parte com a análise descritiva realizada. Com a interpretação dos resultados obtidos, temos fortes indicações de que as espécies de moscas diferem quanto às variáveis Comprimento 3º palpo, Comprimento 4º palpo e Comprimento 13º segmento da antena, mas não diferem quanto às variáveis Largura da Asa, Largura 3º palpo, Comprimento 12º segmento da antena e Comprimento da Asa.

5. Bibliografia

- Azevedo, C. L. N. (2017). Notas de aula sobre análise multivariada de dados http://www.ime.unicamp.br/~cnaber/Material_AM_2S_2017.htm
- Johnson, R. A. & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. 6ª edição, Upper Saddle River, NJ: Pearson Prentice Hall.