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

Trabalho - Parte 1
Relatório - Questão 2

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 banco de dados consiste em 70 observações vindas da medição de sete variáveis em duas espécies de moscas *Leptoconops carteri* e *Leptoconops torrens*, com 35 observações cada. As variáveis constituintes do banco são: espécie (0 - torrens e 1- carteri), comprimento da asa, largura da asa, comprimento do terceiro palpo, largura do terceiro palpo, comprimento do quarto palpo, comprimento do décimo segundo segmento da antena e comprimento do décimo terceiro segmento da antena. Para análises foram utilizados os softwares *R* [1], versão 3.4.2 e *Rstudio* [2], versão 1.0.1.

As duas espécies têm semelhanças morfológicas e por um período de tempo foram consideradas como uma única espécie. O objetivo desta análise é verificar as possíveis distinções entre espécies e para atingirmos tal tarefa utlizaremos como método a análise de componentes principais via matriz de correlação (Principal Component Analisys (PCA) em inglês) para identificar tais distinções. Faremos também uma análise de regressão utilizando a primeira componente para identificarmos diferenças entre as médias da primeira componente para cada espécie.

Observação: Para facilitar a interpretação deste relatório. assumimos que além da variável espécie (0 - torrens e 1- carteri), as variáveis foram consideradas com os seguintes nomes, comprimento da asa (CP_ASA), largura da asa (LG_ASA), comprimento do terceiro palpo (CP_3P), largura do terceiro palpo (LG_3ASA), comprimento do quarto palpo (CP_4P), comprimento do décimo segundo segmento da antena (CP_12ANT) e comprimento do décimo terceiro segmento da antena (CP_13ANT).

2. Análise descritiva

A partir da Figura 1, podemos observar no screeplot que as variâncias (autovalores) associadas a cada componente nos trazem informações relevantes sobre a proporção da variância explicada (PVE), onde se observa que as variâncias das componentes de 1 a 3 trazem uma contribuição maior no valor da PVE do que as outras componentes de acordo com Tabela 1. Na mesma tabela podemos observar que a proporção da variância explicada acumulada (PVEA), indicam que é adequado o uso das três primeiras componentes, já que estas explicam conjuntamente ‘77,00% da variância total. Pelo fato das componentes 4 a 7 não estarem contribuindo significantemente no PVEA e suas variâncias estarem bem próximas, consideramos apenas os três primeiros componentes para as análises.

[1] https://cran.r-project.org/
[2] https://www.rstudio.com/

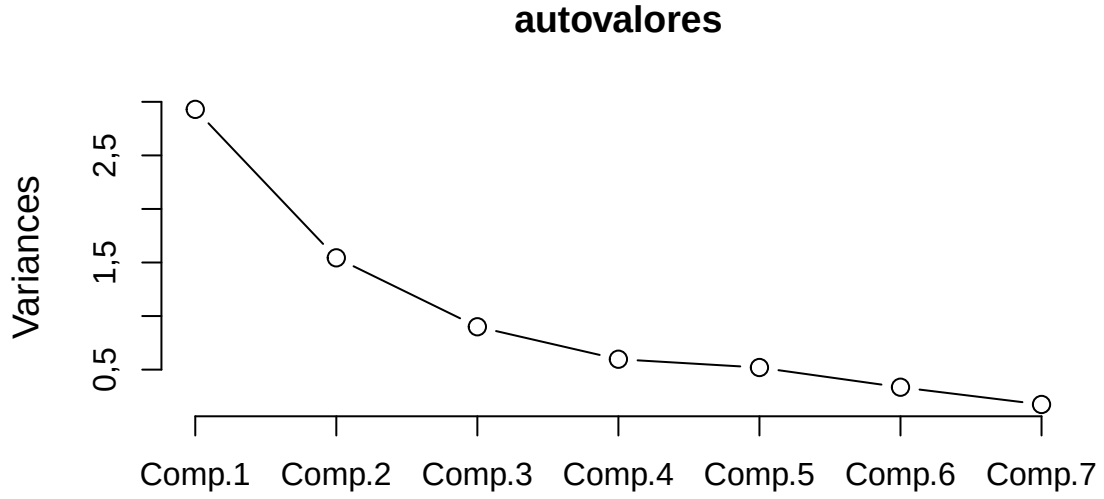


Figura 1: Screeplot Variâncias associadas a cada componente principal

Tabela 1: Sumário das Componentes Principais

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7
DP	1,71	1,24	0,95	0,77	0,72	0,58	0,42
PVE	0,42	0,22	0,13	0,09	0,07	0,05	0,02
PVEA	0,42	0,64	0,77	0,85	0,93	0,98	1,00

Na Tabela 2, vemos os escores das três componentes e podemos interpretá-las de forma a termos um sentido relacionado ao problema. Vale ressaltar que os escores com valores menores que 0.10 serão descartas das análises. Interpretando esta tabela vemos que primeira componente pode ser vista como o escore ponderado entre as sete variáveis. A segunda componente observa-se um constraste entre os escores das variáveis CP_12ANT e CP_13ANT com as outras variaveis. A terceira componente, pode ser interpretada como o como constraste entre as variáveis LG_ASA e LG_3ASA, com CP_3P e CP_4P. Além disso, observamos que para as componentes 1 a 3, as variáveis estão bem correlacionadas com pelo menos uma dessas componentes.

Tabela 2: Coeficientes das três primeiras componentes principais e correlações com cada variável

Variável	Componente 1	Componente 2	Componente 3
CP_ASA	-0,49(-0,84)	**-0,08(-0,10)**	**0,09(0,08)**
LG_ASA	-0,42(-0,72)	-0,18(-0,22)	-0,30(-0,28)
CP_3P	-0,32(-0,54)	-0,30(-0,37)	0,65(0,61)
LG_3ASA	-0,32(-0,55)	-0,21(-0,26)	-0,67(-0.64)
CP_4P	-0,37(-0,64)	-0,36(-0,45)	0,15(0.15)
CP_12ANT	-0,35(-0,60)	0,58(0,72)	**0,04(0,04)**
CP_13ANT	-0,34(-0,58)	0,60(0,75)	**0,07(0,07)**

Na Figura 2. vemos os gráficos de dispersão dois a dois para entre cada componente. Observa-se que nos três gráficos não temos uma separação clara entre as duas espécies, havendo uma sobreposição entre dados. No primeiro e segundo gráfico, podemos perceber que a variabilidade de Carteri parece ser maior do que a de Torrens, já no terceiro parecem ter mesma variabilidade.

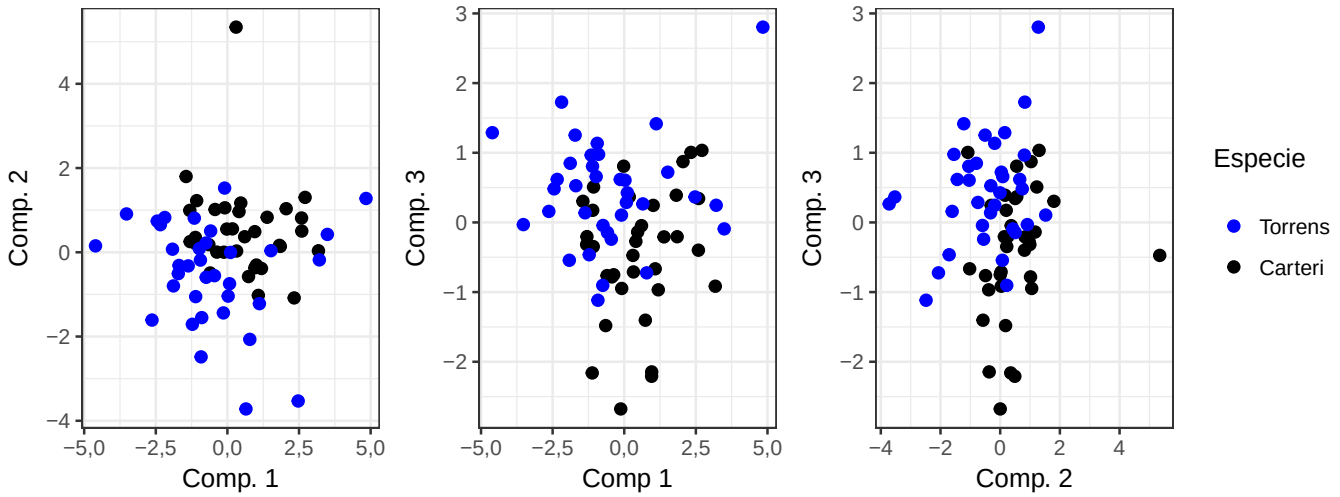


Figura 2: Gráfico de dispersão entre as componentes principais

Na Figura 3, vemos pelos Boxplots das componentes que para a espécie Torrents, as duas primeiras componentes tem maiores valores da distribuição, porém para terceira componente o contrário ocorre, com a espécie Cartieri tendo maior valor.

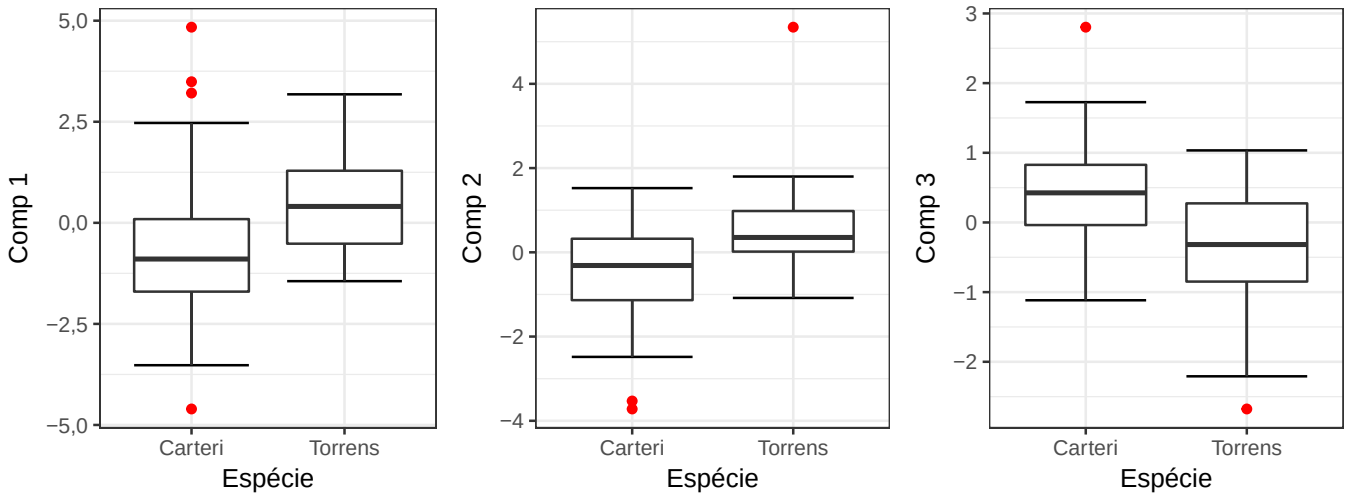


Figura 3: Box-plots das componentes (Comp) por espécie

A partir da Figura 4, podemos observar que para a Componente 1, a distribuição estimada apresenta uma assimetria positiva para a Espécie Torrens e Carteri, também é possivel observar que para a componente 2, supostamente a distribuição apresenta uma assimetria negativa, mais visivel para a Especie Carteri, para este componente , Torrens apresenta visivelmente um ponto atipico, tanto na sua distribuição, quanto nos outros gráficos que foram apresentados. Além disso podemos observar que para a terceira componente a Espécie torrens tem distribuição assimétrica negativa e para Carteri, uma assimétria levemente positiva.

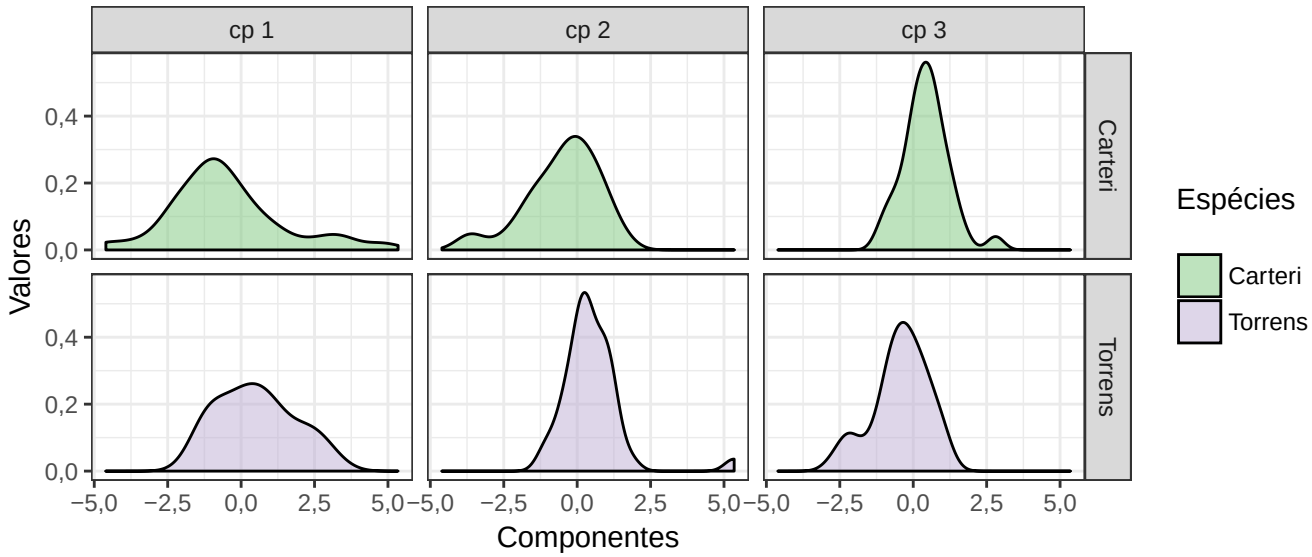


Figura 4: Densidades estimadas das componentes (Comp) por espécie

Podemos ver na Figura 5. o gráfico qqPlot (quantis-quantis com envelopes) mostra que para a Componente 1, para a espécie Carteri, existem cinco pontos fora do envelope, além de um comportamento sistematico em torno da linha de refêrencia, o mesmo acontece com o componente 2, tendo uma forma concova, e o componente 3, reforçando o as densidades vistas acima, ou seja, não é razoavél a suposição de normalidade.

Para a espécie Torrens vemos que para a Componente 1 e 3, existe um comportamento sistematico em torno da linha de referência, já o componente 2 existe um ponto atipico, que entra em concordância com o que foi dito anterioremente para esta espécie.

Logo, para ambas as espécies não há concordancia de normalidade, já que nenhum dos gráficos de quantis-quantis com envelopes entraram em concordância com a normalidade, no qual já era esperado, já que as variaveis também não aparentam ter distribuição normal (Página 10 - Relatório referente a questão 1 - Figura 10).

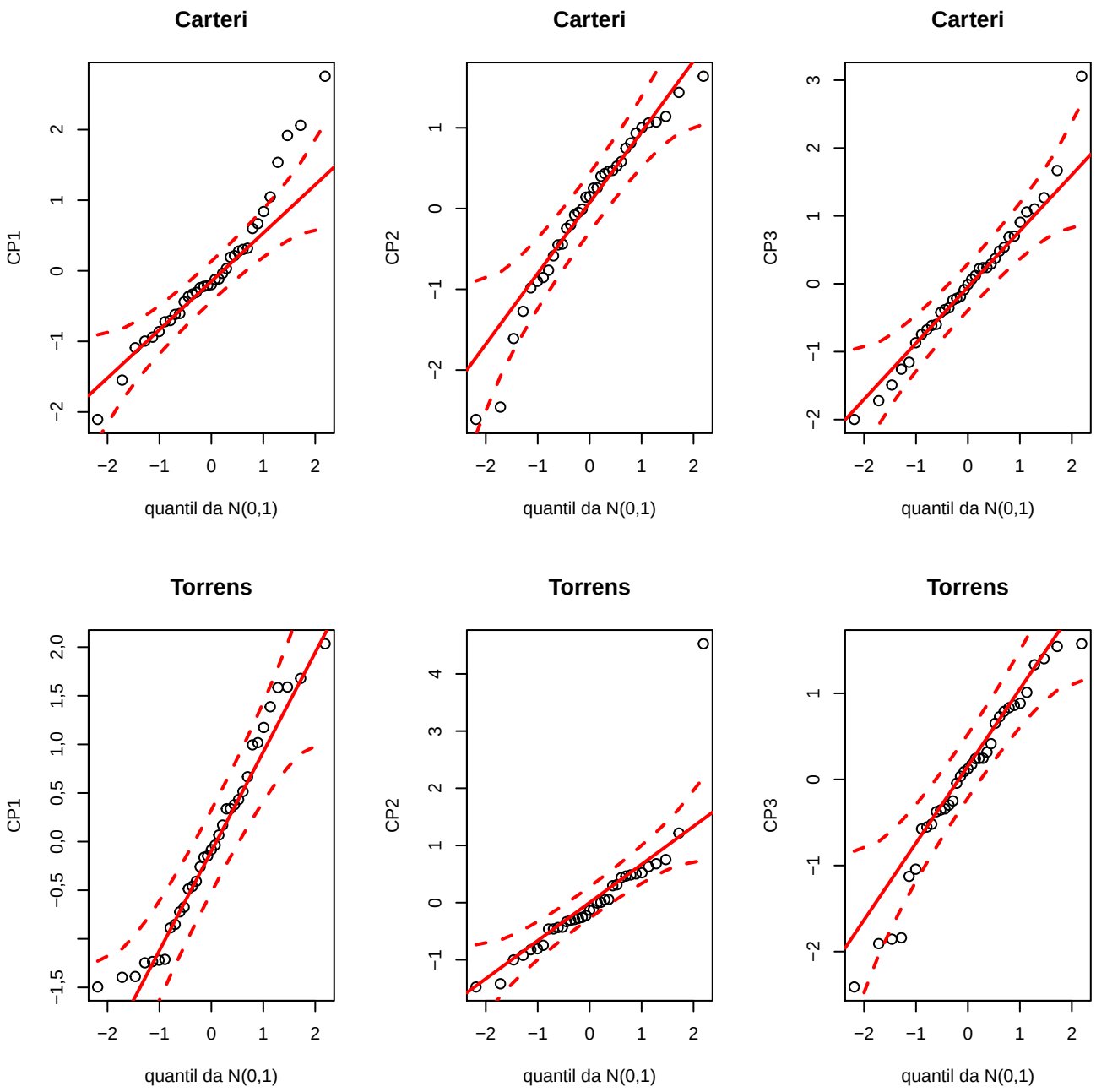


Figura 5: Quantis-quantis com envelopes dos componentes por espécie

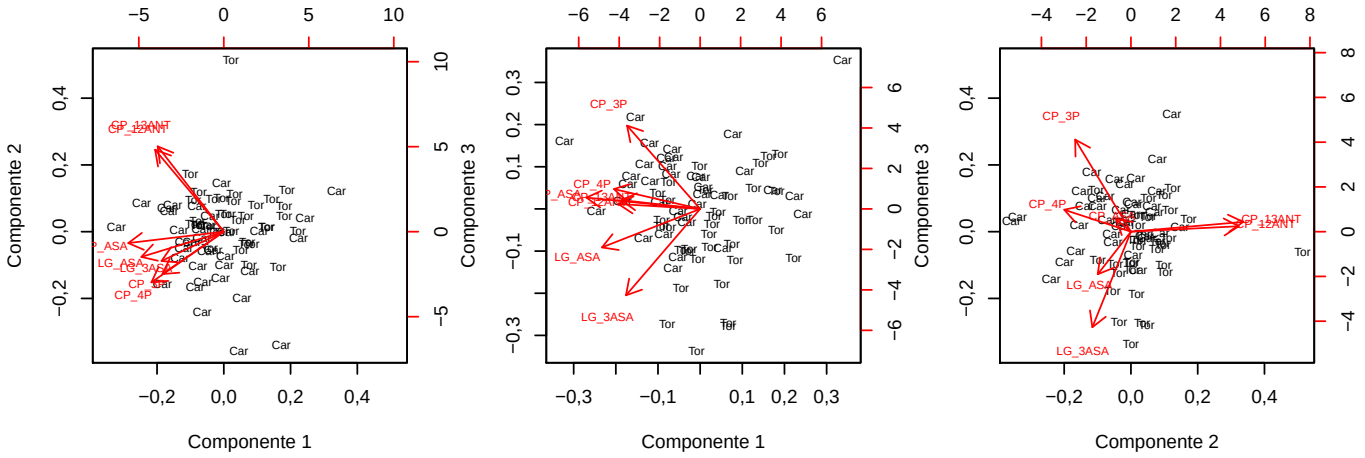
3. Análise Inferencial

Podemos ver no biplot (componente 1 vs componente 2) que para as as variáveis CP_ASA, LG_ASA, LG_3ASA, CP_3P e CP_4P, os pesos dela influênciam mais na componente 1, além disso observamos que há uma tendência maior de concentração dos indivíduos de Carteri na direção das setas, isso mostra que para esse grupo, as variáveis apresentam valores acima da média. Podemos perceber que para a Espécie Torrens, essas cinco variaveis possuem valores abaixo da média.

Para o biplot (componente 1 vs componente 3) vemos que as CA, CP3 e CP4 tem grande influência no componente 1 e além disso, estas variáveis parecem ter valores acima da média para Carteri e abaixo da média para Torrens, mas não é possível tirar alguma conclusão sobre a dispersão dos individuos sobre as setas, já que não parece haver maior concentração de indivíduos de nenhum dos dois grupos na direção das setas.

Para o biplot (componente 2 vs componente 3), vemos que as mesmas suposições do biplot B são válidas,mas também não parece haver maior concentração de individuos de nenhuma das duas spécies sobre as setas.

Assim, concluimos que para as variaveis CP_12ANT e CP_13ANT, não há possibilidade de inferência sobre cada grupo , já que não foi possivel identificar a concentração de indivíduos nessas setas, além disso para as variaveis CP_ASA, CP_3P , CP_4P, a espécie Carteri apresenta valores acima da média ou em torno dela, já Torrens apresenta mais valores em torno ou abaixo da média.



Queremos observar uma relação entre as duas espécies e podermos compará-las, vamos ajustar um modelo de regressão linear utilizando a primeira componente principal obtida. As vantagens deste método são: A redução de dimensionalidade via PCA, evitar multicolinearidade entre preditores e mitigação do *overfitting*. O Modelo ajustado foi:

$$Y_{ij} = \mu_1 + \alpha_i + \varepsilon_{ij}$$

assim:

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

onde: α_1=0 e i = (1=Torrens,2=Cartieri), $j = 1, 2, \ldots, 35$.

Observação: O Ajuste dos parâmetros foram realizados pela forma usual de mínimos quadrados.

Podemos ver que todas as informações sobre os parâmetros se encontram na Tabela 3 abaixo. Foi realizado um teste do tipo $CBU = M$, para verificar as possíveis diferenças entre as médias entre as espécies **Carteri** e **Torrens**.

Podemos observar que o teste $CBU = M$ (Azevedo (2015)) equivale a testar $H_0 : \alpha_2 = 0$, diante deste contexto, na Tabela 3, é possível notar que para a componente 1, Torrens e Carteri são espécies diferentes, já que o resultado do teste abaixo, rejeita a hipótese nula.

As estimativas para as médias estão na Tabela 4 e vemos que para Carteri o valor da média é significativamente menor que a espécie Torrens, tal fato também é notado no biplot do componente 1 e além disso é possível notar que os Intervalos de confiança para as duas espécies estão sobrepostos, tal fato, não necessariamente será preciso, como já foi dito, a estimação intervalar não é muito precisa pois resíduos apontam que o modelo não seria apropriado.

Tabela 3: Estimativas dos parâmetros do modelo de regressão

	Parâmetro	Estimativa	EP	Estatística t	p-valor
1	μ_1	0,51	0,28	1,84	0,07
2	α_2	-1,03	0,40	-2,60	0,01

Tabela 4: Médias preditas pelo modelo

	Especie	Estimativa	EP	IC inf	IC sup
1	Torrens	0,51	0,28	-0,04	1,06
2	Carteri	-0,52	0,40	-1,30	0,26

A análise de resíduos mostrou que o modelo proposto não está bem ajustado, podemos observar no gráfico A que os resíduos supostamente são independentes, pois parecem se distribuir aleatoriamente em torno do zero. No gráfico B, há indícios de heterocedasticidade nos resíduos pois a variabilidade muda de uma espécie para outra. No gráfico C, pelo histograma vemos que a distribuição aparenta ter uma leve assimetria positiva, já gráfico D vemos que supostamente os resíduos não seguem uma distribuição normal, já que existem pontos fora do envelope, e possivelmente há uma leve sistematização dos resíduos em torno da linha de referência, além disso parece haver caudas pesadas.

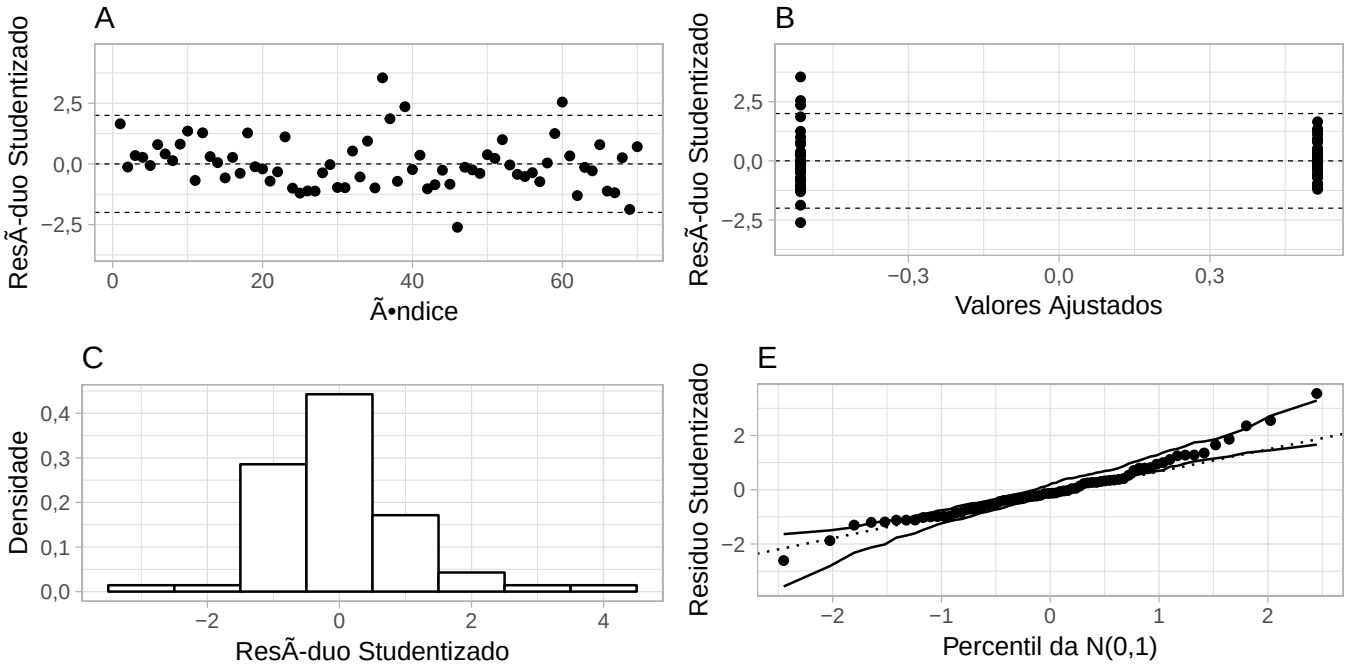


Figura 6: Gráficos para análise de resíduos do modelo ajustado utilizando a primeira componente

4. Conclusões

A partir das análises realizadas, a análise de componentes principais nos trouxe a informação que algumas variáveis se comportam de forma diferente para cada grupo, mas não nos possibilitou uma visualização clara da separação dos dois grupos. Além disso, a análise da componente 1, usando um modelo de regressão, possibilitou ver que existe diferenças entre as espécies, mas conforme a análise de resíduos, o modelo não tem um bom ajuste já que seus resíduos indicam que as suposições de homocedasticidade e normalidade não foram satisfeitas, uma das formas de contornam este problema, talvez seja ajustar um modelo que nos possibilite analisar os dados de maneira mais assertiva.

5. Bibliografia

- Azevedo, C. L. N (2017). Notas de aula sobre análise multivariada de dados
- Johson, R. A. and Wichern, D. W. (2007). Applied Multivariate Statistical Analysis, 7a edição, Upper Saddle River, NJ : Prentice-Hall.