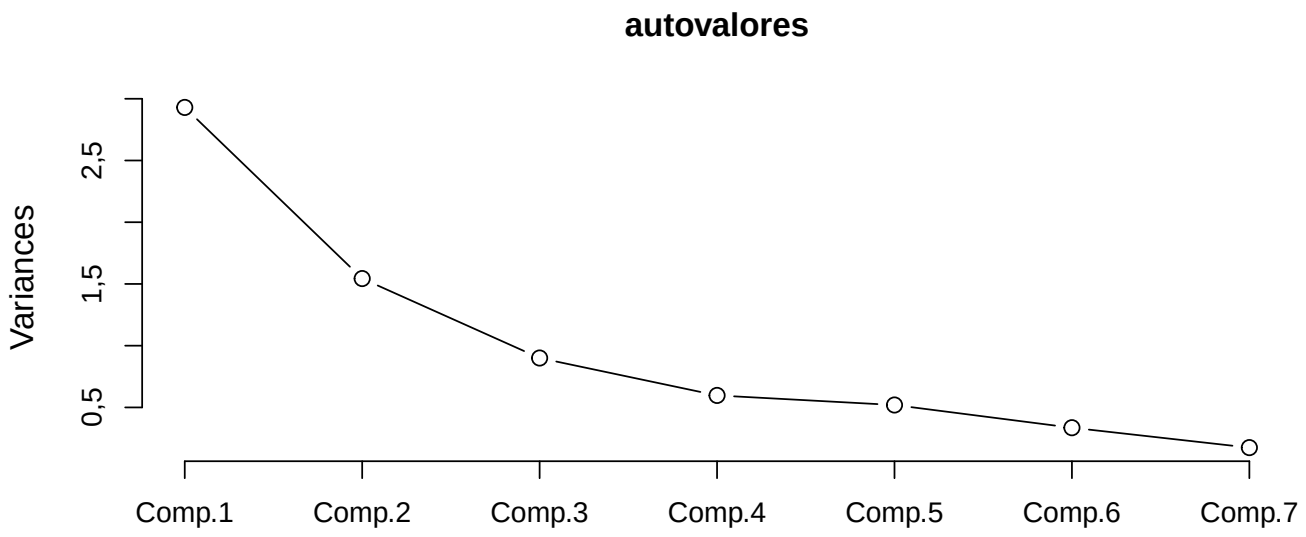
Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

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. A variáveis 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 foram consideradas morfologicamente similares 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 como métodoa análise de componentes principais (Principal Component Analisis (PCA) em inglês) para identificar tais distinções, via matriz de correlação para realizar tal tarefa. Faremos também uma análise de regressão utilizando a primeira componente.

Observação: Para facilidade de interpretação deste presente trabalho assumimos que as variaveis 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

Podemos na Tabela XX, composta pelos valores do desvio padrão(DP), proporção da variabilidade(PVE) e proporção

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

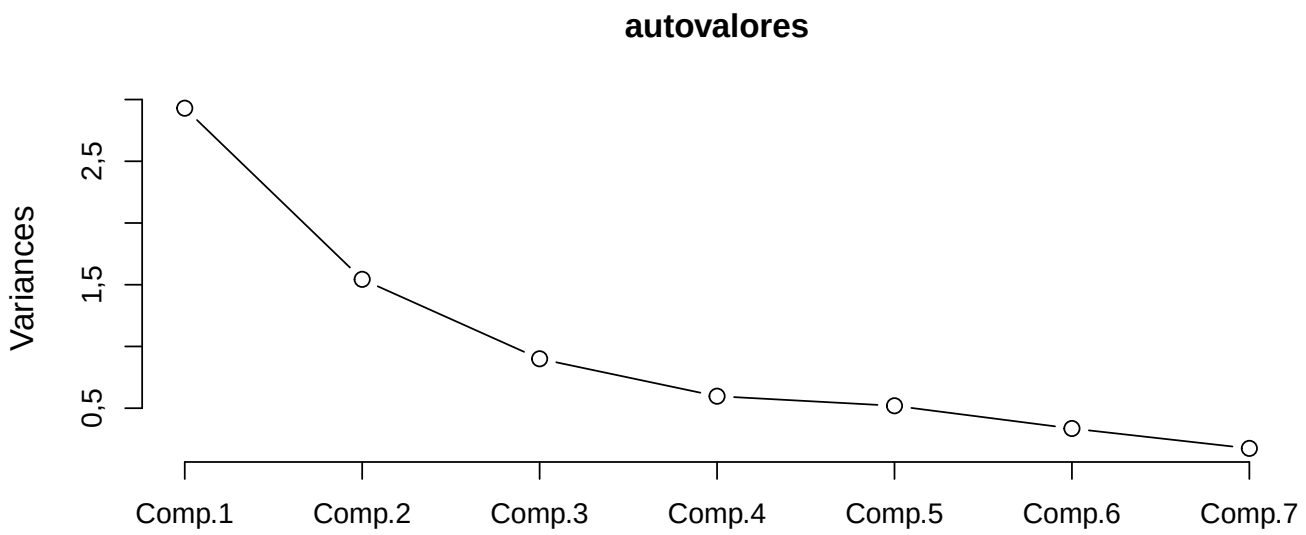


Figura 1: Screeplot legenda aqui

da varaibilidade acumulada(PVEA), vemos, na PVEA, que as três componentes principais conjuntamente explicam ‘77,00% da variabilidade dos dados e portanto, considerano este número razoável, vamos utilizar apenas três componentes para analisarmos a estrutura da variabilidade dos dados. pode-se observar também que as outras variâncias, da quarta a sétima componente, não trazem muita informação acerca da variabilidade, e podemos observar pelo *screeplot*, que as três primeiras componentes explicam boa parte da variabilidade dos dados e que a partir da quarta, a explicação já não é tão significativa.

Tabela 1: LEGENDA Aqui

	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 XX, 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 A primeira componente pode ser vista como o escore ponderado entre cas sete componentes. A segunda, como um constraste entre os escores das variáveis largura da asa, comprimento do terceiro palpo, largura do terceiro palpo, comprimento do quarto palpo e das variáveis comprimento

do décimo segundo segmento da antena e comprimento do décimo terceiro segmento da antena. A terceira componente, pode ser interpretada como o como constraste entre as variáveis largura da asa e largura do terceiro palpo, com os comprimentos do terceiro e quarto palpo.

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

Variável	Componente 1	Componente 2	Componente 3
CP_ASA	-0,49	-0,08	0,09
LG_ASA	-0,42	-0,18	-0,30
CP_3P	-0,32	-0,30	0,65
LG_3ASA	-0,32	-0,21	-0,67
CP_4P	-0,37	-0,36	0,15
CP_12ANT	-0,35	0,58	0,04
CP_13ANT	-0,34	0,60	0,07

Na Figura XX. 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 dos dados. No primeiro e segundo gráfico, podemos perceber que a variabilidade de Carteri parece ser maior do que a de Torrens. No Terceiro parecem ter mesma variabilidae. Na Figura XX, vemos pelos boxplot 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.

A partir da Figura X, podemos observar que para o Componente 1, a distribuição estimada apresenta uma assimetria positiva para a Espécie Torrens e Carteri, também é possivel observar que para o 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 componente 3, vemos que para a Espécie torrens temos uma distribuição assimetrica negativa e para Carteri temos uma assimetria levemente positiva.

A verificação de normalidade para os componentes pode ser de suma importância da inferência estatística, podemos ver na Figura X, que para a Componenete 1 não é razoavél a suposição de normalidade, já que em ambos os gráficos de probabilidade, os pontos se comportam de forma sistematica, é notavel que para a componente 1 Carteri, mais especificamente, existem pontos fora do envelope, tendo a fuga de normalidade para ambas as especies.

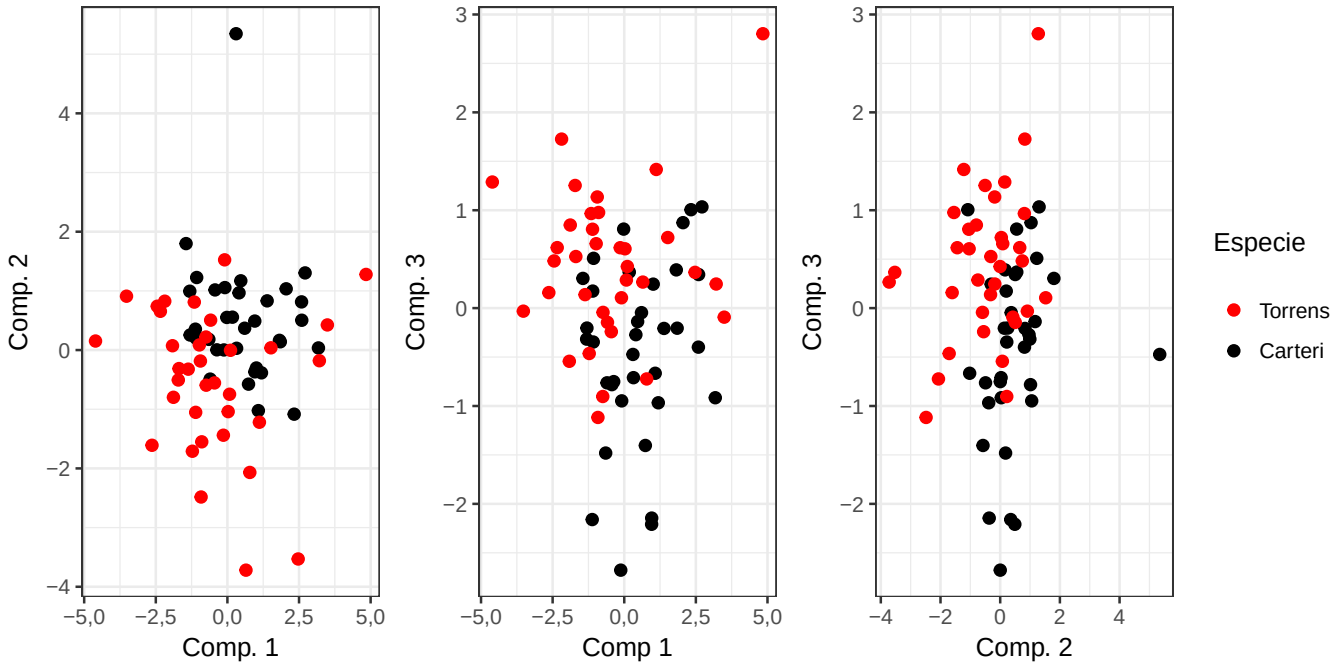


Figura 2: colocar legenda aqui

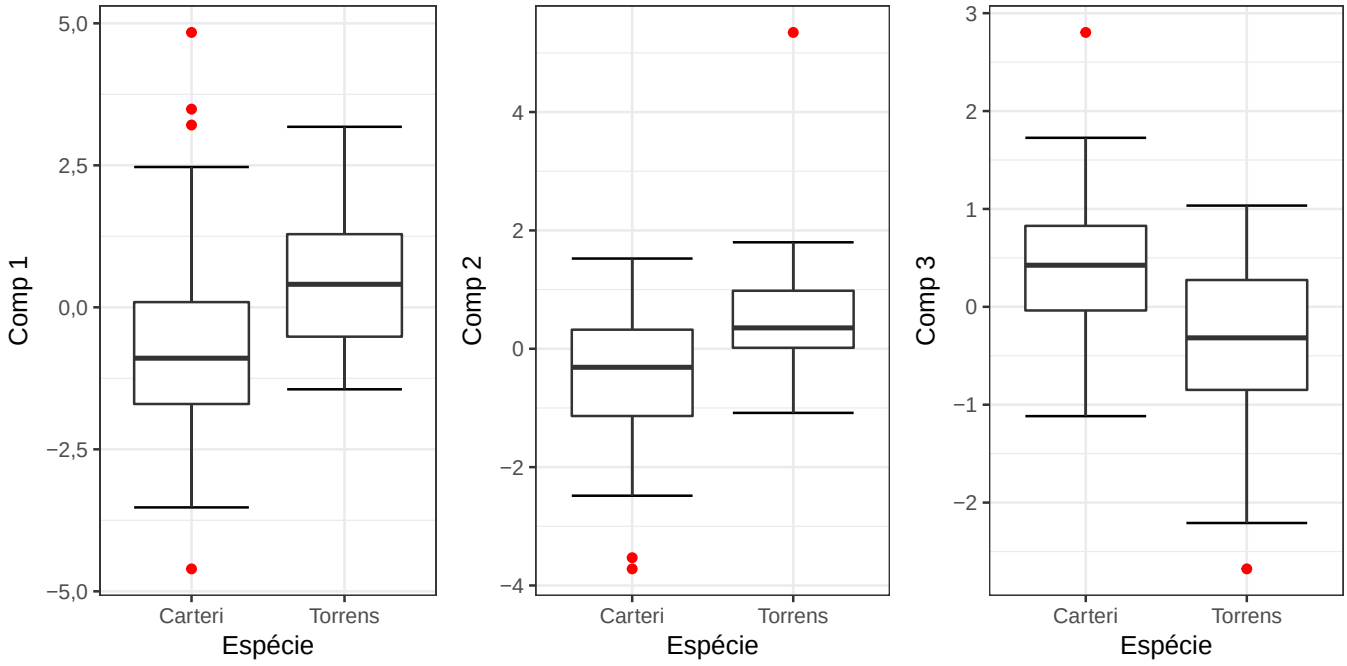


Figura 3: colocar legenda aqui

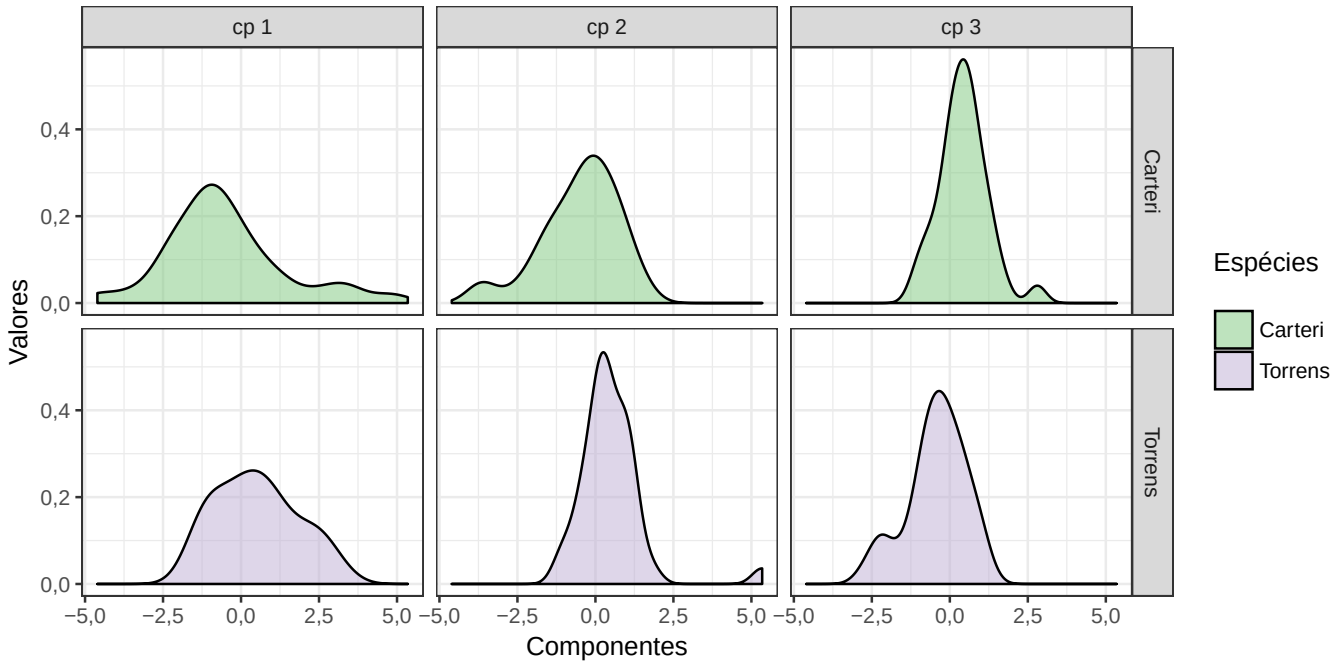


Figura 4: colocar legenda aqui

Já no componenete 2, podemos ver uma forma concava em torno da linha de referencia para a especie Carteria, além para do ponto atipico, que entra em concordancia com o que foi dito anterioremente para Torrens.

Para o componente 3 é notavel que existem poucos pontos fora do gráfico de envelope.

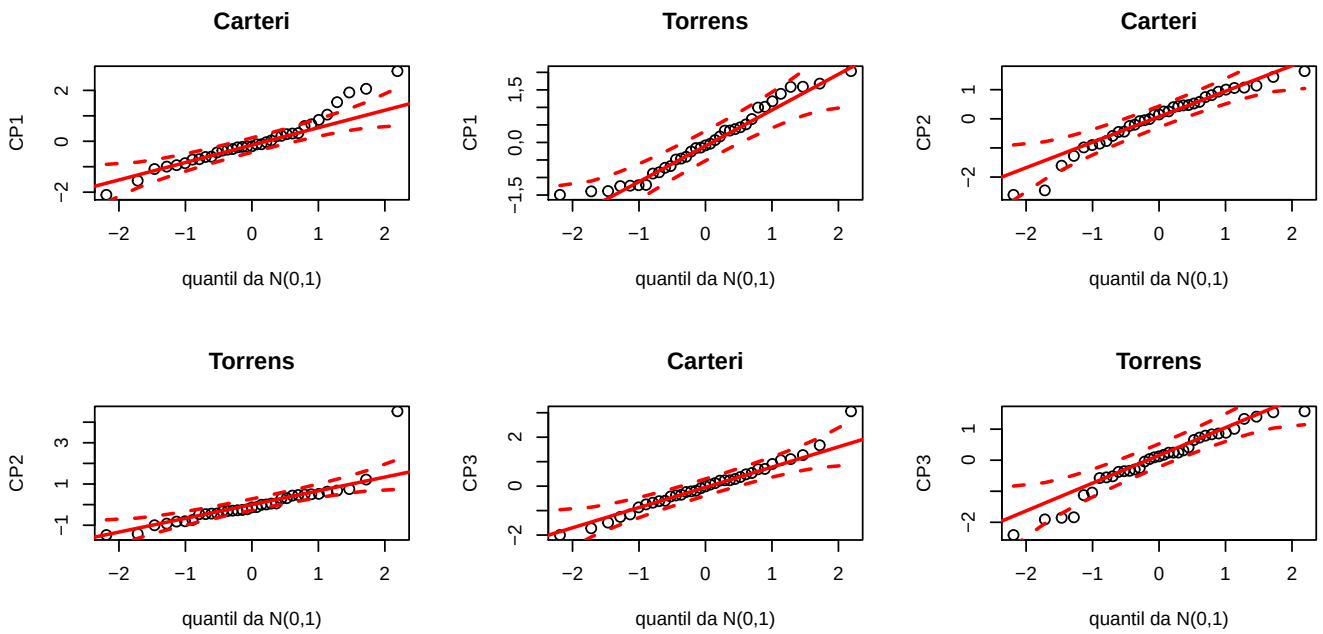


Figura 5: legenda aqui

3. Análise Inferencial

Podemos observar na Figura Componente 1 vs Componente 2 que as variaveis CP_ASA, LG_ASA , CP_3P, LG_3ASA e CP_4P há uma concentração dos individuos Carteri ao redor dos scores que a especie Torrens, mostrando que a disposição das setas (Escores) estão apontando para este grupo, além dessas variaveis já citadas apresentarem mais peso na componente 1, além disso isso indica que estas variáveis apresentam valores acima da média no grupo Carteri, Já no grupo Torrens, observando a disposição da dispersão deste grupo sobre as variaveis, supostamente essas cinco variaveis possuem valores a abaxio da média de acordo com a dispersão sobrepostas a ela.

Os componentes 1 e 3 (Figura 2.6B) e entre componentes 2 e 3 (Figura 2.6C), vemos que as mesmas conclusões são válidas para CA, CP3 e CP4, isto é, estas variáveis parecem ter valores acima da média para Carteri e abaixo para Torrens. No entanto, nestes dois biplots já não é possível chegar às mesmas conclusões para LA e LP3: não parece haver maior concentração de indivíduos de nenhum dos dois grupos na direção das setas, ou opostos a elas.

Para pobservarmos uma relação entre as espécies, vamos utilizar a 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:

$$Y_{ij} = \mu_1 + \alpha i + \varepsilon_{ij}$$

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

O Ajuste dos parâmetros foram realizados pela forma usual de m

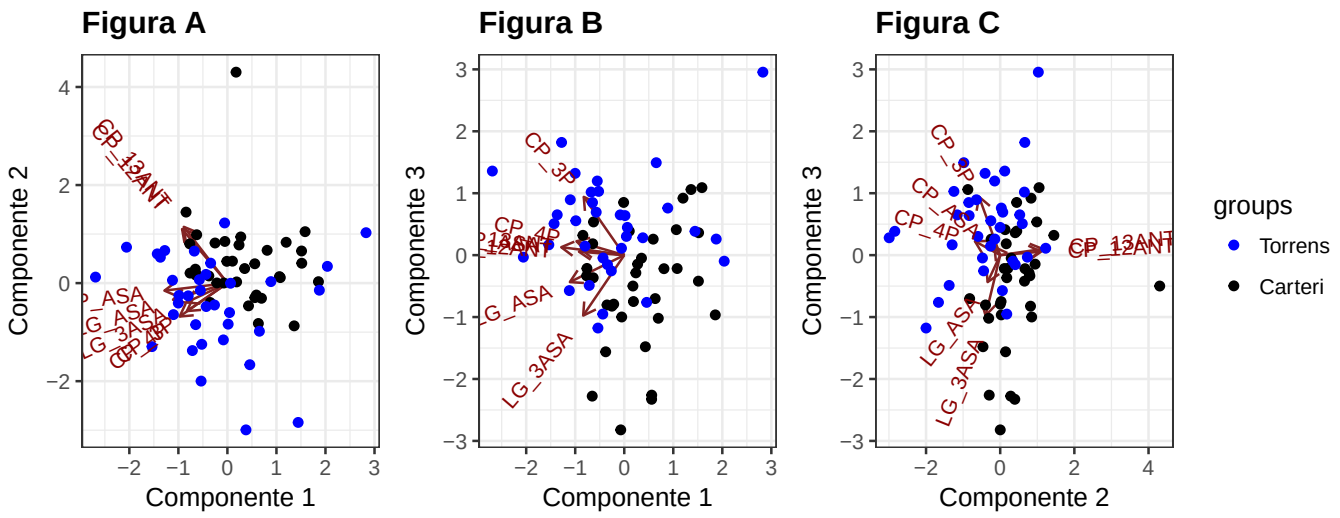


Figura 6: legenda aqui

A Análise de resíduos mostra que o modelo proposto não está bem ajustado porque no gráfico A que há alguns pontos,4 destes, que estão fora da região tracejada em 2.5 e -2.5, que podem indicar que estes resíduos não estão bem ajustados. No gráfico B, há indícios de heterocedasticidade no re´siduos pois a variabilidade muda de um grupo para o outro. No gráfico C, vemos que possívelmente os resíduos não seguem uma distribuição normal, e possívelmente possuem uma distribuição assimétrica positiva. No gráfico D, vemos que há alguns pontos fora das bandas de confiança e nas caudas o modelo não está bem ajustado, com caudas pesadas.

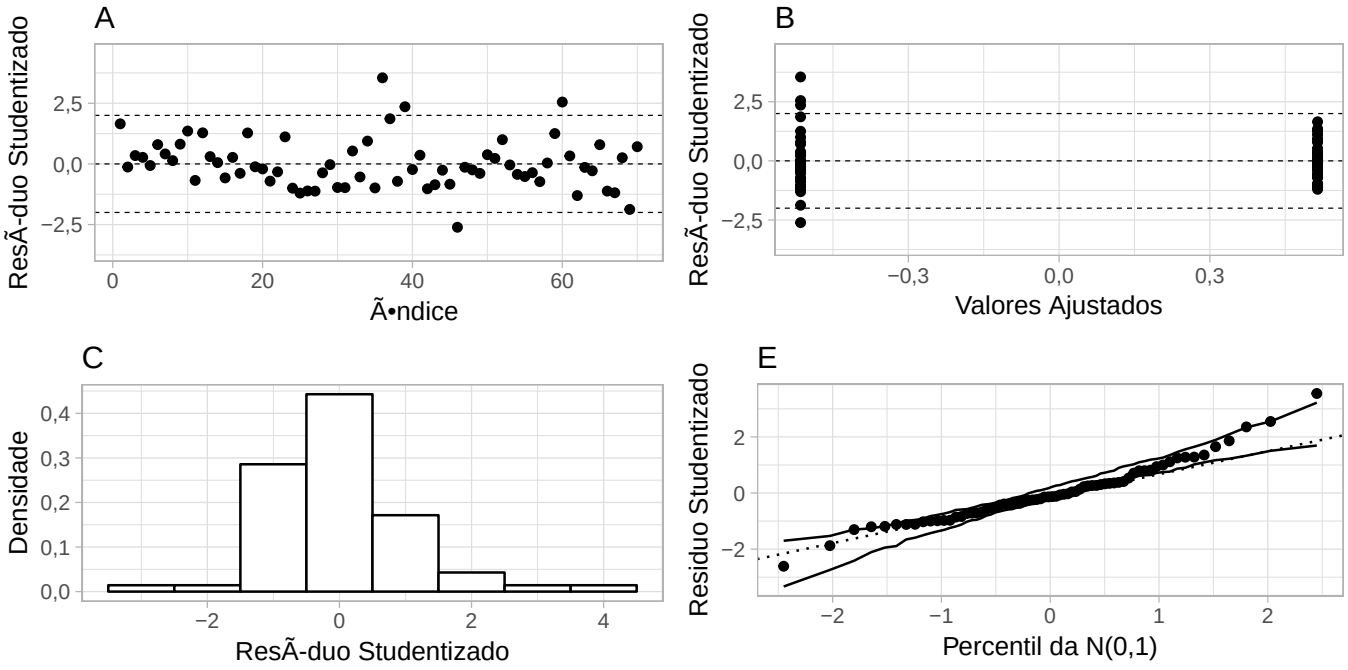


Figura 7: legenda aqui

4. Conclusões

A partir das análises realizadas, a análise de componentes principais nos trouxe a informação que algumas variaveis se comportam de forma diferente para cada grupo, mas não nos possibilitou ter uma visualização clara da separação dos dois grupos. Além diso a análise da componente 1, usando um modelo de regressão, possibilitou ver que existe diferenças enre as Especies, 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 viavel possivel.

5. Bibliografia