UNICAMP - Universidade Estadual de Campinas
IMECC - Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística
ME731 - Métodos em Análise Multivariada

Trabalho Final - ME731

Caio Henrique de Sousa Lima RA 214144
Carlos Eduardo Lima Kamioka RA 168624
Giovanni Torres Chaves RA 198105
Júlio Mendes Pazelli RA 219494

Campinas - 2021

1 Questão 1

1.1 Introdução

O banco de dados é composto pelo perfil de tamanho e formato de 48 tartarugas, sendo 24 machos e 24 fêmeas. As informações obtidas foram largura, comprimento e altura. O estudo tinha como objetivo observar e comparar o sexo ao qual o animal pertencia e sua relação com essas medidas físicas. Para realizar a análise desse banco, foi utilizado o software estatístico R (R Core Team (2020)), realizando uma análise descritiva com o auxílio do pacote *Tidyverse* (Wickham et al. (2019)) e aplicado algumas metodologias como a Análise de Variância Multivariada (MANOVA).

1.2 Análise descritiva

A Tabela 1 apresenta as medidas resumo das tartarugas, dividas por sexo. Observa-se que as médias das medidas das fêmeas foram maiores que as dos machos, e do mesmo modo as variâncias são bem maiores para o sexo feminino. As assimetrias estão em torno de zero e as curtoses entre 2 e 3, para ambos os sexos.

Tabela 1: Medidas resumo

	Média	Var.	DP	CV(%)	Min.	Med.	Max.	CA	Cur.
Feminino									
Comprimento	136,00	451,39	21,25	15,62	98,00	136,50	177,00	-0,23	2,26
Largura	102,58	171,73	13,10	12,77	81,00	102,00	132,00	0,31	2,55
Altura	51,96	66,65	8,16	15,71	38,00	51,00	67,00	-0,03	2,19
Masculino									
Comprimento	113,38	138,77	11,78	10,39	93,00	115,00	135,00	-0,08	2,10
Largura	88,29	50,04	7,07	8,01	74,00	89,00	106,00	0,20	3,15
Altura	40,71	11,26	3,36	8,24	35,00	40,00	47,00	0,18	2,20

A Figura 1 mostra que, para ambos os sexos, as medidas de altura, largura e comprimento das tartarugas estão muito correlacionadas e com tendência positiva.

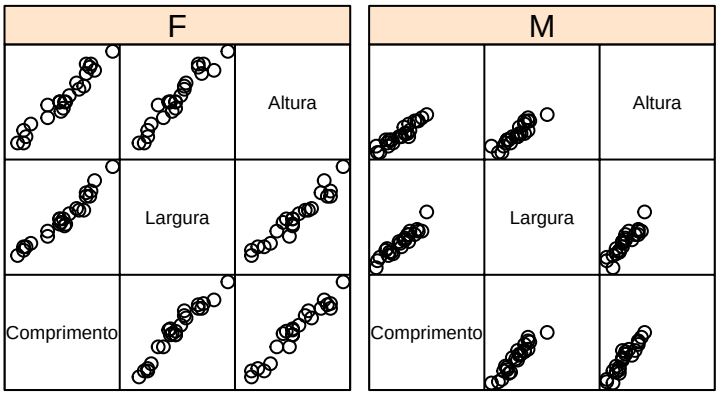

Figura 1: Matriz de gráfico de dispersões sexo feminino.

A matriz de covariância para ambos os sexos, vista na Tabela 2, parecem ser diferentes, na qual as medidas do sexo masculino aparentam ter variância bem menor se comparado ao sexo feminino. Pode ser visto também na Tabela 3 que as matrizes de correlação são aparentemente diferentes, em que as medidas são mais correlacionadas para as fêmeas.

Tabela 2: Matriz de Covariâncias.

	Feminino			Masculino		
	Comprimento	Largura	Altura	Comprimento	Largura	Altura
Comprimento	451,39	271,17	168,70	138,77	79,15	37,38
Largura	271,17	171,73	103,29	79,15	50,04	21,65
Altura	168,70	103,29	66,65	37,38	21,65	11,26

Tabela 3: Matriz de Correlação

	Feminino			Masculino		
	Comprimento	Largura	Altura	Comprimento	Largura	Altura
Comprimento	1,00	0,97	0,97	1,00	0,95	0,95
Largura	0,97	1,00	0,97	0,95	1,00	0,91
Altura	0,97	0,97	1,00	0,95	0,91	1,00

Os boxplots da Figura 5 (Apêndice) mostram que há uma aparente diferença entre os sexos nas três medidas e nota-se que no Comprimento e na Altura dos machos há uma assimetria em torno da mediana, o que pode indicar uma não normalidade. Da mesma forma, os histogramas vistos na Figura 6 (Apêndice) e os gráficos quantil-quantil da Figura 7 (Apêndice) aparentam seguir uma distribuição diferente da normal, visto que não são simétricas e alguns possuem caudas pesadas. Mesmo com estas observações, iremos continuar com a análise inferencial.

1.3 Análise inferencial

Como visto na Análise Descritiva, as matrizes de covariâncias aparentam serem diferentes entre os sexos, e assim, para observar inferencialmente essa diferença foi realizado um teste de Box. Obteve-se uma estatística de 24,04 e um p-valor de 0.0005, logo temos que as matrizes realmente são diferentes. Após isso, foi realizada uma MANOVA com o objetivo de visualizar se existe diferença entre os vetores de médias para ambos os sexos, e como visto na Tabela 4, rejeitamos a hipótese dos vetores serem iguais.

Tabela 4: Resultados da MANOVA

	Valor	Aproximação pela distribuição F	p-valor
Wilks	0,41	21,28	$<$ 0,0001
Pillai	0,59	21,28	$<$ 0,0001
Hotelling-Lawley	1,45	21,28	$<$ 0,0001
Roy	1,45	21,28	$<$ 0,0001

Foi estimado então os parâmetros do Modelo de Regressão Normal Linear Multivariada (MRNLM), dado por $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, onde $\mathbf{Y}$ é a matriz de dados (supõe que $\mathbf{Y}$ segue uma normal 3-variada), $\mathbf{X}$ a matriz de planejamento, $\mathbf{B}$ a matriz parâmetros de regressão e $\mathbf{E}$ a matriz de erros. Como visto

na Tabela 5, todas as estimativas foram significativas. Dessa forma, é interessante verificar quais medidas se diferem entre os sexos, e portanto aplicamos um teste CBU.

O teste CBU tem as seguintes hipóteses: H_0 : **CBU** $=$ **M** versus H_1 : **CBU** $\neq$ **M**. A Tabela 6, mostra que todas as medidas são estatisticamente diferentes entre os sexos.

Tabela 5: Estimativas dos parâmetros do modelo

	Estimativa	EP	Estatística t	p-valor
μ_1	136,00	3,51	38,79	$<$ 0,0001
μ_2	102,58	2,15	47,72	$<$ 0,0001
μ_3	51,96	1,27	40,78	$<$ 0,0001
α_{21}	-22,63	4,96	-4,56	$<$ 0,0001
α_{22}	-14,29	3,04	-4,70	$<$ 0,0001
α_{23}	-11,25	1,80	-6,24	$<$ 0,0001

Tabela 6: Testes CBU ($\alpha_{2i} = 0$)

Parâmetro	Estatística Qui-quadrado	p-valor
α_{21} (Length)	20,82	$<$ 0,0001
α_{22} (Width)	22,10	$<$ 0,0001
α_{23} (Height)	38,99	$<$ 0,0001

Avaliando os resíduos do modelo ajustado nas Figuras 2, 3 e 4, referentes ao Comprimento, Largura e Altura, respectivamente, podemos observar que a suposição de homocedasticidade e de normalidade não parecem ser razoáveis.

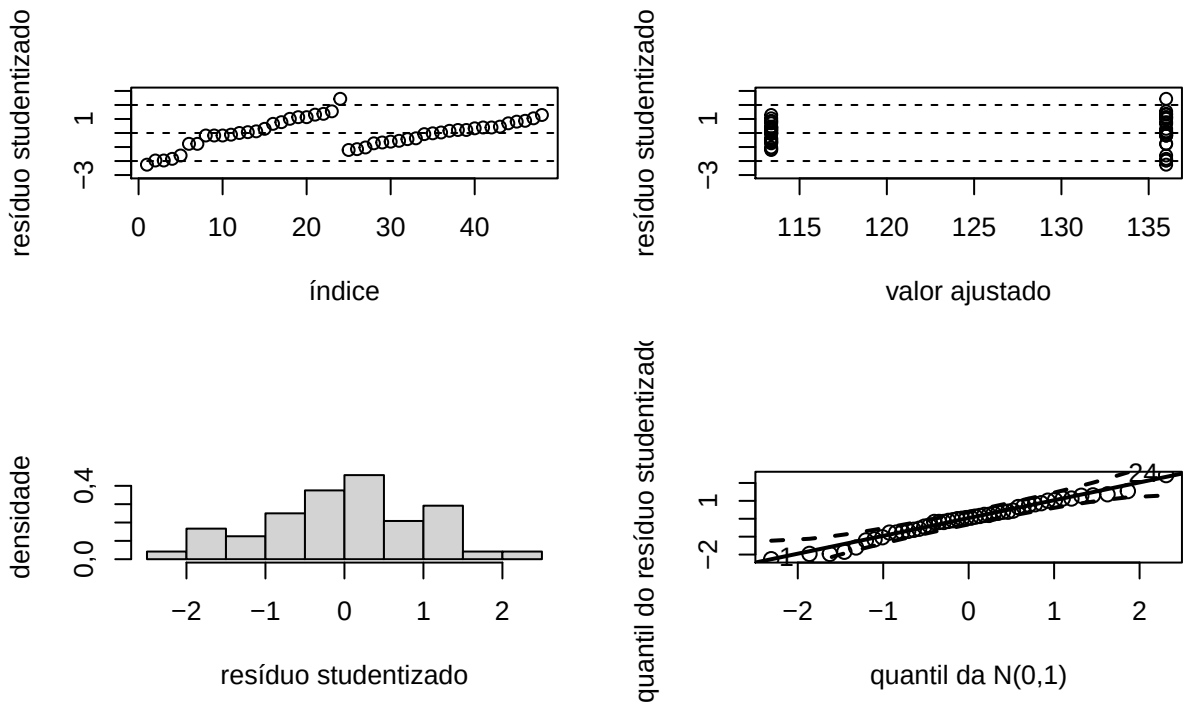

Figura 2: Análise de Resíduos - Comprimento

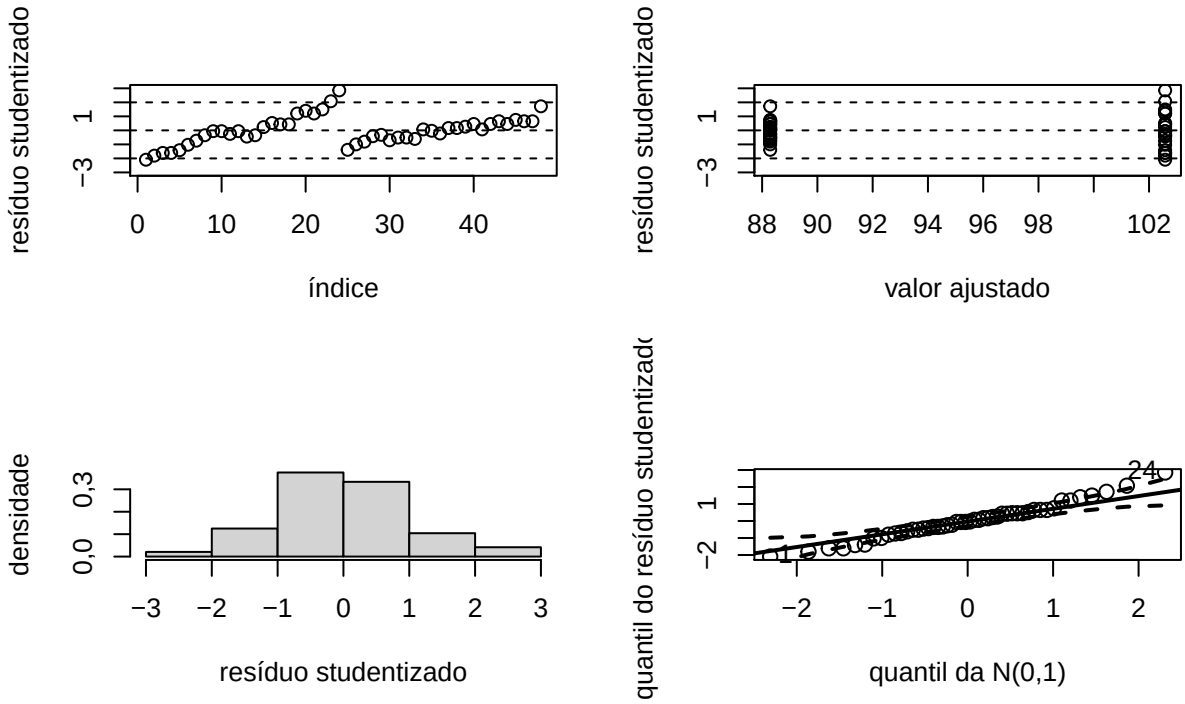

Figura 3: Análise de Resíduos - Largura

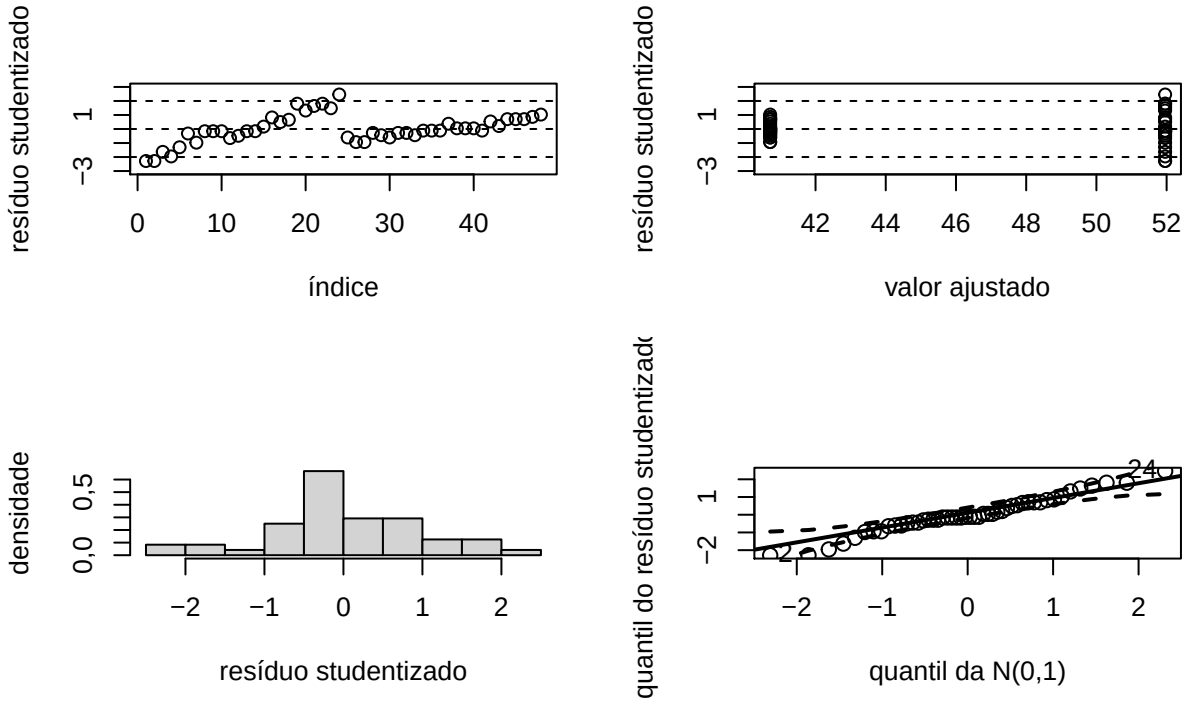

Figura 4: Análise de Resíduos - Altura

1.4 Conclusão

Utilizando a metodologia MRNLM e o teste CBU, concluímos que as médias entre os sexos para todas as variáveis são diferentes. Porém, como as suposições do modelo não foram satisfeitas, essa conclusão pode ser falsa. É necessário, então, um outro tipo de análise para melhores conclusões.

Referências

Azevedo, C. L. N. (2020). Notas de aula sobre análise multivariada de dados. https://www.ime.unicamp.br/~cnaber/Material_AM_2S_2020.htm.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686.

Apêndice

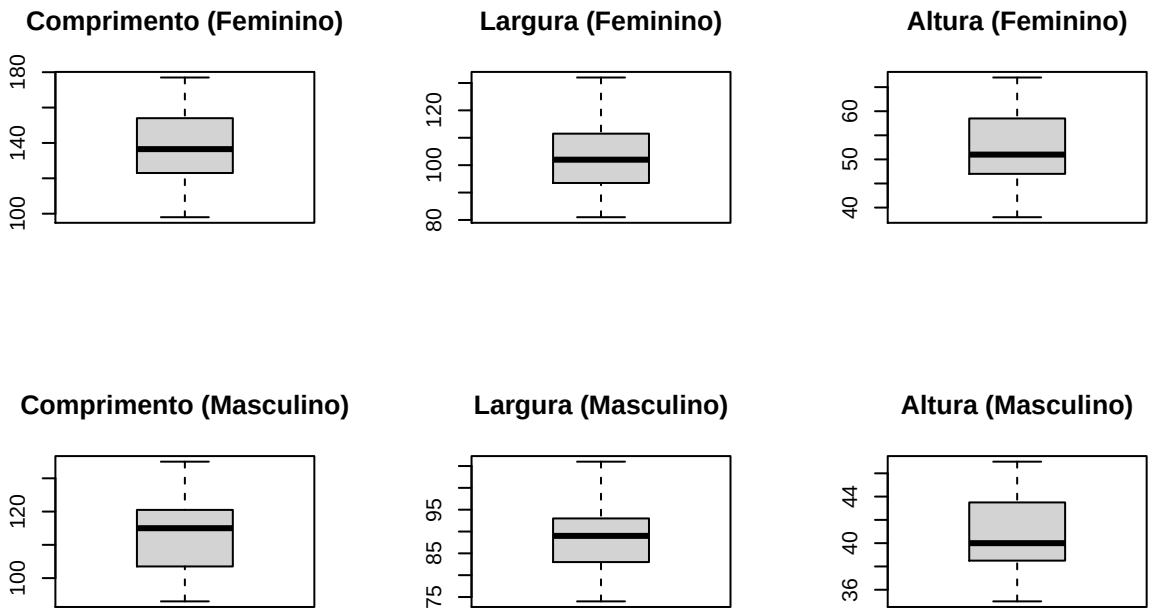

Figura 5: Boxplots das variáveis por sexo.

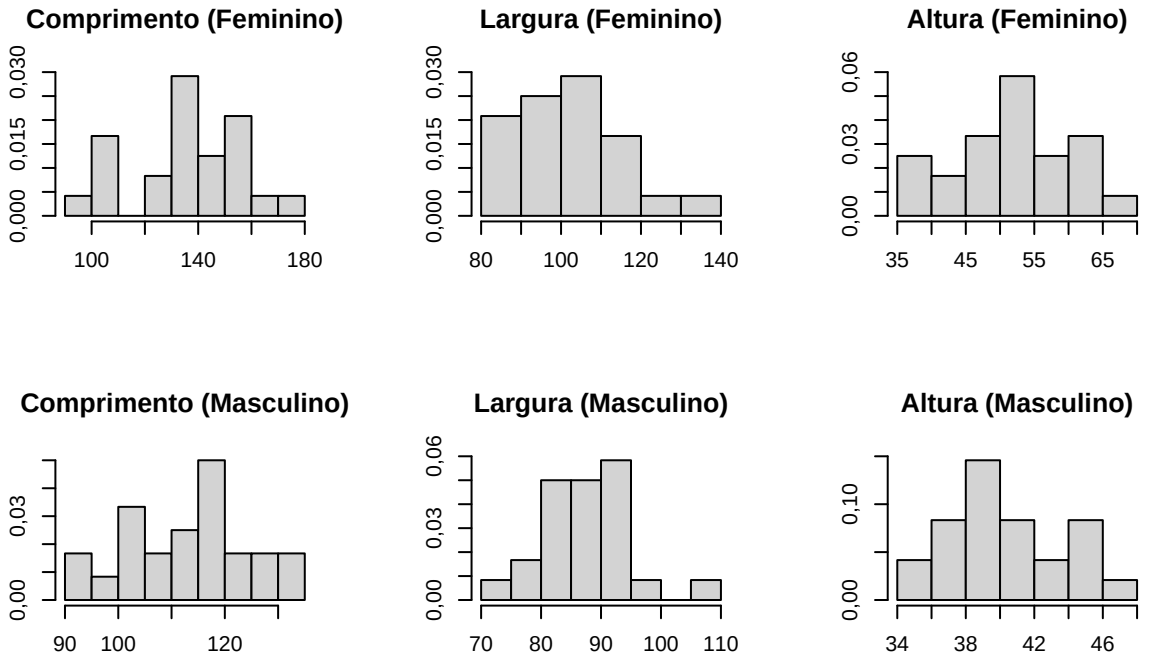

Figura 6: Histogramas das variáveis por sexo.

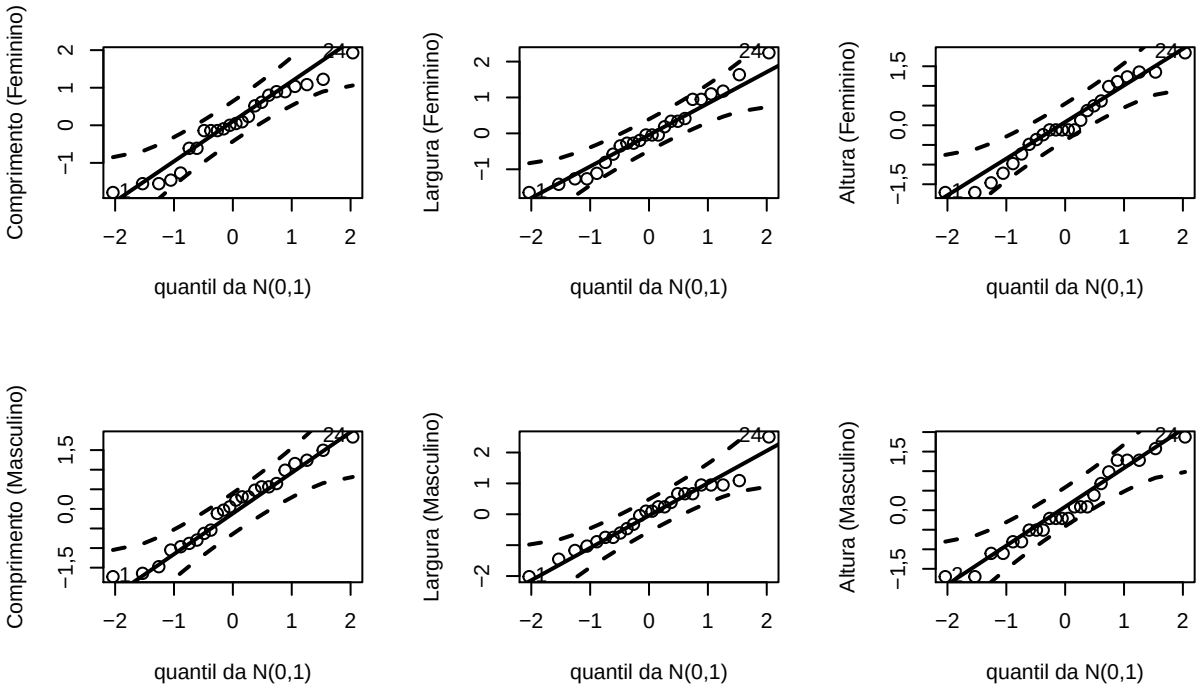

Figura 7: QQplots das variáveis por sexo.

2 Questão 2

2.1 Introdução

O banco de dados é composto pelos resultados de 25 atletas em provas que compõem o heptatlo, nas olimpíadas de 1988, em Seul. São 7 resultados de cada prova do heptatlo e o escore total na competição. O estudo tem como objetivo utilizar a metodologia das componentes principais para caracterizar as atletas, com relação às variáveis medidas. Para realizar a análise desse banco, foi utilizado o software estatístico R (R Core Team (2020)), realizando uma análise descritiva com o auxílio dos pacotes *Tidyverse* (Wickham et al. (2019)) e *Corrplot* (Taiyun Wei and Viliam Simko (2017)), e aplicando as metodologias necessárias.

2.2 Análise descritiva

Antes de iniciar nossa análise, foi retirado a variável escore total. O motivo da remoção foi devido ao fato de que a variável é, de certa forma, uma combinação linear das outras variáveis. Podendo trazer viés para o futuro rank gerado.

A Tabela 7 contém as medidas resumo dos resultados das provas. Podemos notar que a maioria das variáveis não tem uma curtose ideal (próxima de 3) e que os coeficientes de assimetria também não foram ideais. Também é possível notar que para as provas shot, javelin e run800m, os resultados das atletas foram mais dispersos comparado com as outras variáveis.

Tabela 7: Medidas resumo dos resultados das provas

	Média	Var.	DP	CV(%)	Min.	Med.	Max.	CA	Cur.
hurdles	13,84	0,54	0,74	5,32	12,69	13,75	16,42	1,65	7,22
highjump	1,78	0,01	0,08	4,37	1,50	1,80	1,86	-1,99	7,87
shot	13,12	2,23	1,49	11,37	10,00	12,88	16,23	0,18	2,78
run200m	24,65	0,94	0,97	3,93	22,56	24,83	26,61	-0,17	2,62
longjump	6,15	0,22	0,47	7,71	4,88	6,25	7,27	-0,48	4,28
javelin	41,48	12,57	3,55	8,55	35,68	40,28	47,50	0,16	1,89
run800m	136,05	68,74	8,29	6,09	124,20	134,74	163,43	1,40	5,89

A Figura 8 contém a matriz de gráfico de dispersão dos resultados de cada prova. Podemos notar correlações positivas e correlações negativas entre as variáveis. As Tabelas 8 e 9 contém a matriz de covariâncias e a matriz de correlações dos resultados de cada prova. Podemos notar que as provas de corrida são negativamente correlacionadas com as provas de desempenho medidas em metros. A mesma coisa acontece comparando a covariância entre as provas de desempenho medidas em metros com as provas de corrida.

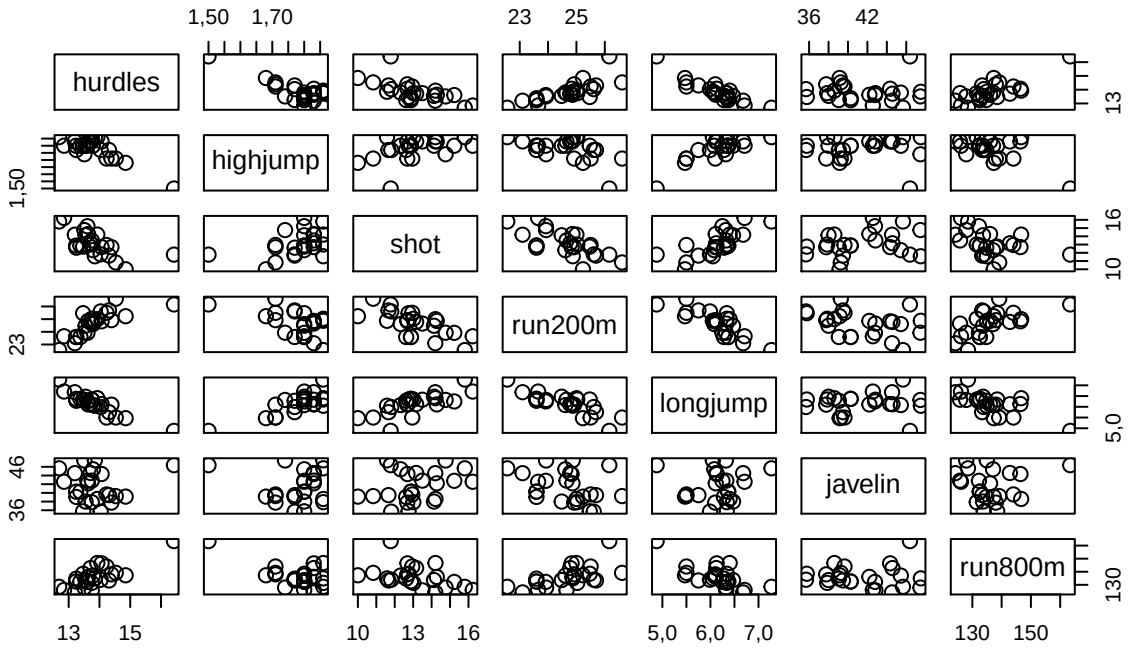

Figura 8: Matriz de gráfico de dispersão entre as variáveis.

Tabela 8: Matriz de covariâncias

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
hurdles	0,54	-0,05	-0,72	0,55	-0,32	-0,02	4,76
highjump	-0,05	0,01	0,05	-0,04	0,03	0,00	-0,38
shot	-0,72	0,05	2,23	-0,99	0,53	1,42	-5,19
run200m	0,55	-0,04	-0,99	0,94	-0,38	-1,14	4,96
longjump	-0,32	0,03	0,53	-0,38	0,22	0,11	-2,75
javelin	-0,02	0,00	1,42	-1,14	0,11	12,57	0,59
run800m	4,76	-0,38	-5,19	4,96	-2,75	0,59	68,74

Tabela 9: Matriz de correlações

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
hurdles	1,00	-0,81	-0,65	0,77	-0,91	-0,01	0,78
highjump	-0,81	1,00	0,44	-0,49	0,78	0,00	-0,59
shot	-0,65	0,44	1,00	-0,68	0,74	0,27	-0,42
run200m	0,77	-0,49	-0,68	1,00	-0,82	-0,33	0,62
longjump	-0,91	0,78	0,74	-0,82	1,00	0,07	-0,70
javelin	-0,01	0,00	0,27	-0,33	0,07	1,00	0,02
run800m	0,78	-0,59	-0,42	0,62	-0,70	0,02	1,00

As Figuras 10, 11 e 12 (estão no Apêndice) contém os boxplots, os histogramas e os QQplots dos resultados de cada prova, respectivamente. Podemos perceber, pelas figuras, a existência de pelo menos uma variável que não segue uma distribuição normal. Sendo assim, rejeita-se a normalidade multivariada. Já a Figura 13 (Apêndice) contém o correlograma dos resultados de cada prova. Nele,

reforça-se a ideia de que as provas de corrida pertencem a um grupo e as provas de desempenho medidas em distância a outro. Ou seja, as correlações entre provas de grupos distintos são negativas e, correlações do mesmo grupo, positivas.

2.3 Análise inferencial

O objetivo principal da análise é classificar os atletas, com relação às variáveis medidas, bem como reduzir a dimensão dos dados, para futuras análises. A metodologia utilizada para a devida classificação foi a Análise de Componentes Principais, veja mais em Azevedo, C. L. N. (2020).

Podemos ver pela Figura 9 e Tabela 10 que duas componentes principais explicam, aproximadamente, 80% da variabilidade dos dados. Com isso, como podemos ver na Tabela 11, foram retidas apenas duas componentes. Nessa tabela, também é possível ver que para criar o rank, a componente principal mais indicada é a primeira, pois é um contraste entre os resultados das provas de tempo e distância, o que reforça as correlações encontradas anteriormente. Os pesos das provas de corrida são positivos, já os pesos das provas de desempenho medidas em distância são negativos. Como estamos utilizando a matriz de correlações, os resultados do desempenho de cada atleta estão padronizados. Ou seja, para variáveis de tempo, atletas com valores negativos (tempo abaixo da média) possuem um desempenho melhor. Para as variáveis de distância, atletas com valores negativos (distância abaixo da média) possuem um desempenho pior. Com isso, os atletas que se destacaram no heptatlo teriam um escore menor. Portanto, para melhor interpretação do escore, o mesmo foi multiplicado por -1, pois, desta forma, atletas com escore mais alto tiveram um desempenho melhor.

Os resultados estão na Tabela 12. Podemos ver que o rank para algumas atletas foi alterado com a análise de componentes principais.

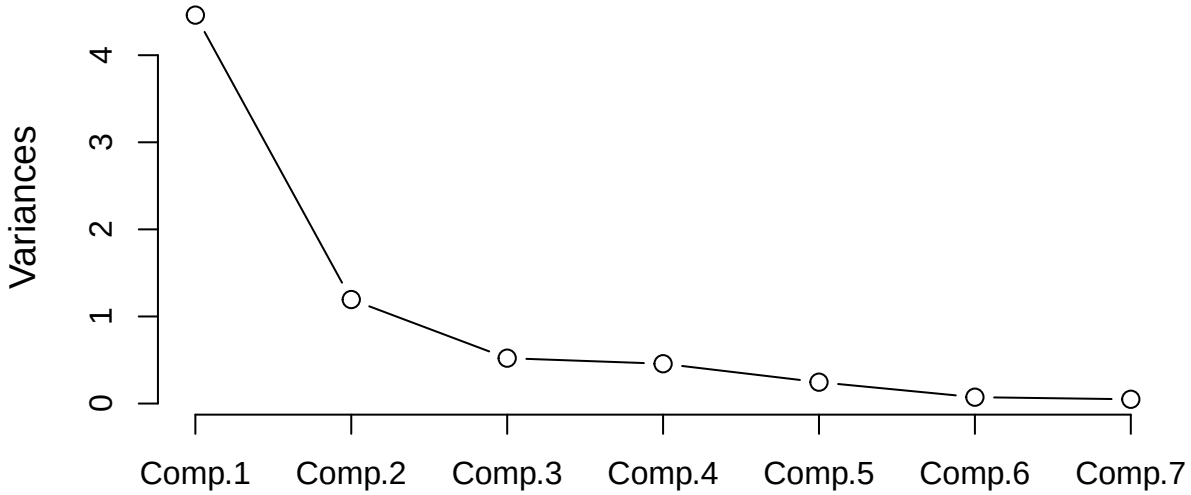

Figura 9: Screeplot das variáveis.

Tabela 10: Variância Explicada

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
PVE (%)	63,72	17,06	7,44	6,53	3,50	1,04	0,70
PVEA (%)	63,72	80,78	88,22	94,75	98,26	99,30	100,00

Tabela 11: Componentes principais

	Comp.1	Comp.2
hurdles	0,45	0,16
highjump	-0,38	-0,25
shot	-0,36	0,29
run200m	0,41	-0,26
longjump	-0,46	-0,06
javelin	-0,08	0,84
run800m	0,37	0,22

Tabela 12: Rankings e escores dos atletas

Atleta	Rank	Escore PCA	Novo rank	Atleta	Rank	Escore PCA	Novo rank
Joyner-Kersee (USA)	1	4,12	1	Braun (FRG)	14	-0,00	14
John (GDR)	2	2,88	2	Ruotsalainen (FIN)	13	-0,02	15
Behmer (GDR)	3	2,65	3	Yuping (CHN)	15	-0,09	16
Sablovskaite (URS)	5	1,36	4	Hagger (GB)	17	-0,17	17
Choubenkova (URS)	4	1,34	5	Brown (USA)	18	-0,52	18
Schulz (GDR)	12	1,19	6	Mulliner (GB)	20	-1,09	19
Fleming (AUS)	7	1,10	7	Hautenauve (BEL)	19	-1,13	20
Greiner (USA)	6	1,04	8	Kytola (FIN)	21	-1,45	21
Lajbnerova (CZE)	8	0,92	9	Geremias (BRA)	22	-2,01	22
Bouraga (URS)	10	0,76	10	Hui-Ing (TAI)	23	-2,88	23
Wijnsma (HOL)	11	0,56	11	Jeong-Mi (KOR)	24	-2,97	24
Dimitrova (BUL)	9	0,53	12	Launa (PNG)	25	-6,27	25
Scheider (SWI)	16	0,14	13				

2.4 Conclusão

A partir da classificação pela Análise de Componentes Principais (ACP), foi possível notar que o ranking gerado pelo escore da ACP foi diferente comparado com o ranking original do escore total. Como por exemplo, a atleta Schulz, da República Democrática Alemã, tinha ficado em 12º lugar. Com o ranking feito a partir da primeira componente principal, a mesma atleta ficou em 6º lugar.

Referências

Azevedo, C. L. N. (2020). Notas de aula sobre análise fatorial. https://www.ime.unicamp.br/~cnaber/Material_AM_2S_2020.htm.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686.

Taiyun Wei and Viliam Simko (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). https://github.com/taiyun/corrplot.

Apêndice

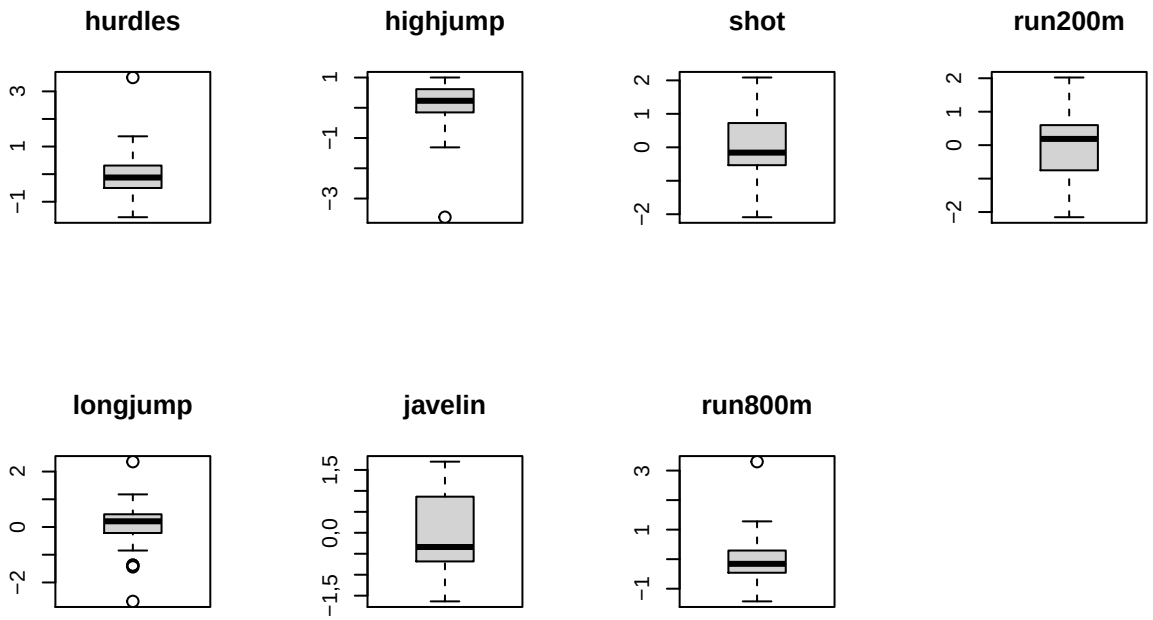

Figura 10: Boxplots das variáveis.

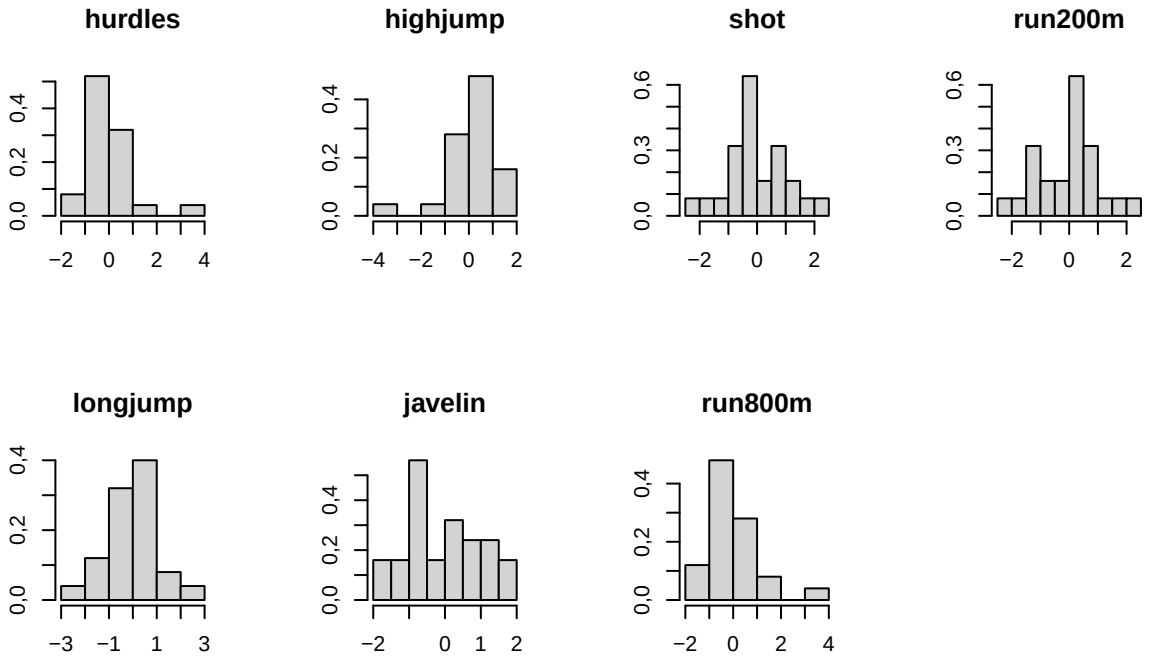

Figura 11: Histogramas das variáveis.

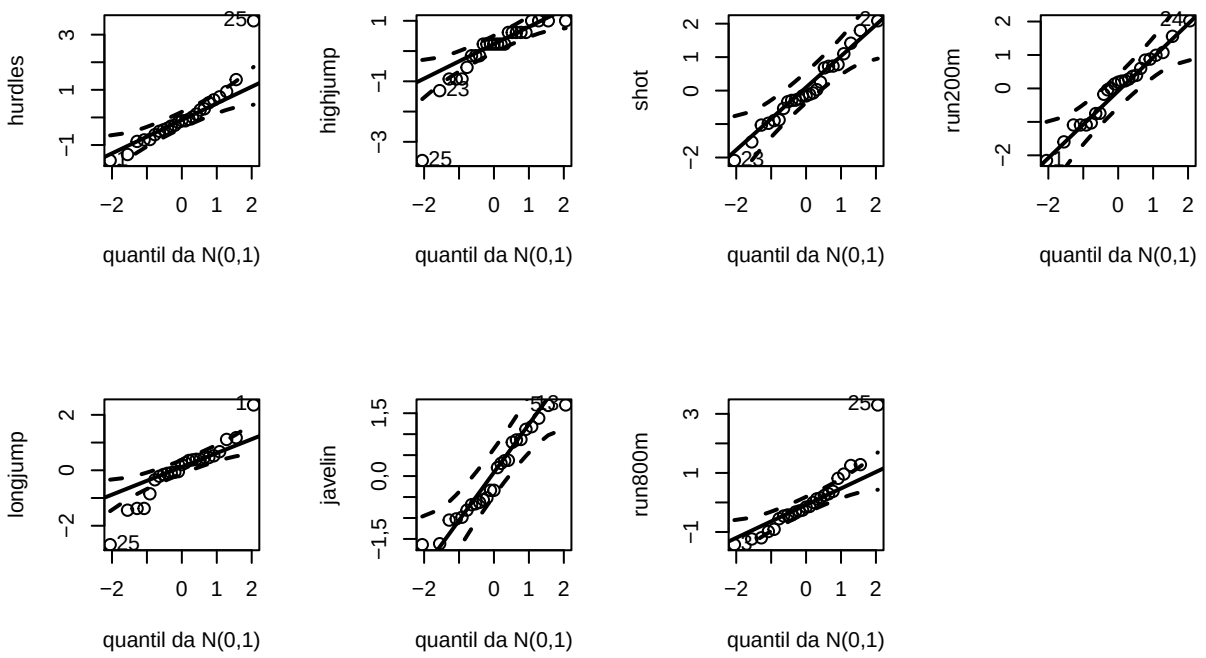

Figura 12: QQplots das variáveis.

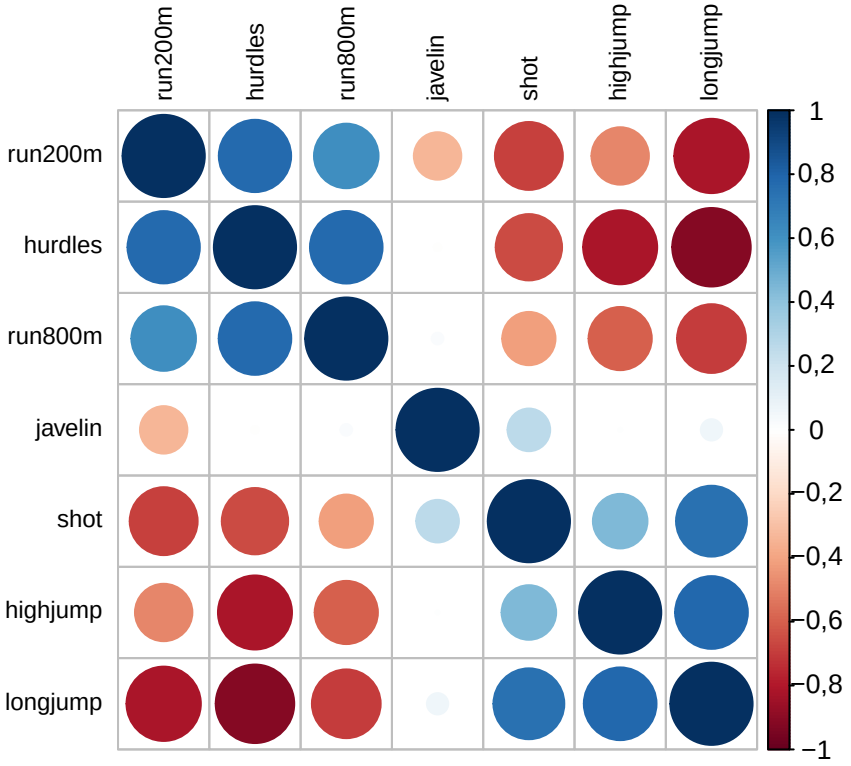

Figura 13: Correlograma das variáveis.

3 Questão 3

3.1 Introdução

Os dados foram retirados de um estudo feito para comparar a relação entre status socioeconômico dos pais com a saúde mental dos estudantes. As classificações de status foram de A (alto) até E (baixo) e para saúde foram boa, presença fraca de sintomas, presença moderada de sintomas e debilitado. O número de indivíduos testados foi de 1760 indivíduos, distribuídos da seguinte forma:

Tabela 13: Saúde Mental x Status Socioeconômico dos pais

	Status Socioeconômico dos pais					
Saúde Mental	**A**	**B**	**C**	**D**	**E**	**Total**
Boa	121,00	57,00	72,00	36,00	21,00	307,00
Presença fraca de sintomas	188,00	105,00	141,00	97,00	71,00	602,00
Presença moderada de sintomas	112,00	65,00	77,00	54,00	54,00	362,00
Debilitado	186,00	60,00	94,00	78,00	71,00	489,00
Total	607,00	287,00	384,00	265,00	217,00	1760,00

Para realizar a análise desse banco, foi utilizado o software estatístico R (R Core Team (2020)), realizando uma análise descritiva com o auxílio do pacote *Tidyverse* (Wickham et al. (2019)) e aplicando as metodologias necessárias.

3.2 LETRA A

O modelo probabilístico gerador da tabela de contingência é uma multinomial de tamanho 1760 e 20 categorias. Considerando p_{ij} a probabilidade de pertencer às categorias i de Saúde Mental e j de Status socioeconômico dos pais, em que $p_{i.} = \sum_{j=1}^{J} p_{ij}$ e $p_{.j} = \sum_{i=1}^{I} p_{ij}$ são as probabilidades marginais de cada unidade amostral pertencer, respectivamente, à categoria i de Saúde Mental e à categoria j de Status socioeconômico dos pais. Há então o interesse de testar a seguinte hipótese:

$$\begin{cases} H_0 : p_{ij} = p_{i.}p_{.j}, \forall i, j \\ H_1 : p_{ij} \neq p_{i.}p_{.j}, \textit{para pelo menos um par (i,j)} \end{cases} \quad (1)$$

Ou seja, H_0: as variáveis são estatisticamente independentes e H_1: caso contrário. Deste modo, sob H_0, temos a estatística do teste $Q = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(X_{ij}-E_{ij})^2}{E_{ij}}$, em que $E_{ij} = \frac{X_{i.}X_{.j}}{X_{..}}$. Sob a suposição de independência e com tamanho amostral suficientemente grande, temos que $Q \approx \chi^2_{(I-1)(J-1)}$. Mais informações sobre a metodologia podem ser encontradas em Azevedo (2020).

Levando em conta um nível de significância de 5%, obteve-se uma estatística do teste $Q = 31,17$ e um p-valor = 0,0002, e rejeitamos então a hipótese de independência entre Saúde Mental e Status Socioeconômico dos pais.

3.3 LETRA B

Observando a Tabela 14, com o perfil de linhas, vemos que em relação as pessoas com saúde mental boa, 39,41%, 18,57% e 23,45% são por terem pais com condições socioeconômicas mais altas A, B e C, respectivamente, que são maiores do que as respectivas proporções na população geral. Do mesmo modo, pessoas com presença fraca de sintomas se destacam em status medianos (B, C e D), pessoas com sintomas moderados nos status B e E, e por fim, pessoas debilitadas são maioria no maior status (A) e nos menores (D e E).

Tabela 14: Perfil de Linhas ($\times 100$).

	Status Socioeconômico dos pais					
Saúde Mental	**A**	**B**	**C**	**D**	**E**	**Total**
Boa	**39,41**	**18,57**	**23,45**	11,73	6,84	100,00
Presença fraca de sintomas	31,23	**17,44**	**23,42**	**16,11**	11,79	100,00
Presença moderada de sintomas	30,94	**17,96**	21,27	14,92	**14,92**	100,00
Debilitado	**38,04**	12,27	19,22	**15,95**	**14,52**	100,00
Total	34,49	16,31	21,82	15,06	12,33	100,00

Analisando o perfil de colunas na Tabela 15, vemos que a maioria das pessoas com pais nos status A possuem saúde mental boa ou debilitada, em relação ao status B com saúde mental boa, sintomas fracos ou sintomas moderados, o status C com saúde mental boa ou sintomas fracos, o status D com sintomas fracos ou debilitado, e por fim, no status E a maioria tem sintomas moderados ou estão debilitadas.

Tabela 15: Perfil de Colunas ($\times 100$).

	Status Socioeconômico dos pais					
Saúde Mental	**A**	**B**	**C**	**D**	**E**	**Total**
Boa	**19,93**	**19,86**	**18,75**	13,58	9,68	17,44
Presença fraca de sintomas	30,97	**36,59**	**36,72**	**36,60**	32,72	34,20
Presença moderada de sintomas	18,45	**22,65**	20,05	20,38	**24,88**	20,57
Debilitado	**30,64**	20,91	24,48	**29,43**	**32,72**	27,78
Total	100,00	100,00	100,00	100,00	100,00	100,00

3.4 LETRA C

A Tabela 16 apresenta a inércia e a proporção de variabilidade explicada das duas primeiras componentes. Vemos que o percentual de variabilidade explicada pelas duas é de quase 95%, ou seja, explicam muito sobre os dados.

Tabela 16: Inércia e Proporção de Variabilidade Explicada

Componentes	**Valor singular**	**Inércia Princ.**	**Percentual**	**Percentual acum.**
1	0,1023	0,0105	59,08	59,08
2	0,0797	0,0064	35,87	94,96
3	0,0299	0,0009	5,03	100,00

Vemos na Figura 14 o gráfico Bi-plot. Podemos dizer que o status A está mais relacionado com saúde mental debilitada que as demais, os status B e C estão mais relacionados com presença fraca, o status D está mais relacionado à presença moderada e o status E está mais longe da saúde mental boa.

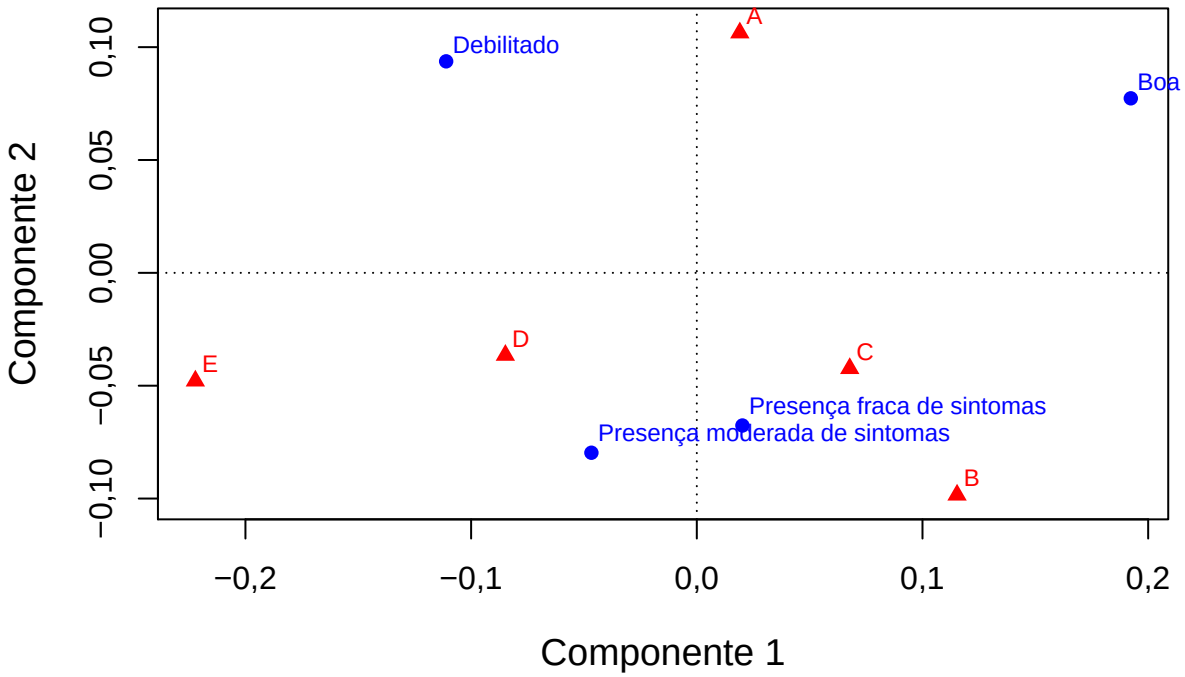

Figura 14: Bi-Plot das duas primeiras componentes.

Referências

Azevedo, C. L. N. (2020). Notas de aula sobre análise de correspondência. https://www.ime.unicamp.br/~cnaber/Material_AM_2S_2020.htm.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686.

4 Questão 4

4.1 Introdução

O problema a ser resolvido diz respeito a características relativas à saúde de 768 mulheres indianas, portadoras de uma herança genética chamada prima. O banco de dados contém 768 observações (mulheres) e 9 variáveis (medidas). Como existem muitos dados faltantes, foram retiradas todas as observações que continham pelo menos um dado faltante. Com isso, o número de observações diminuiu para 392. O objetivo é criar uma regra de classificação de diabetes em função das outras variáveis. Para realizar a análise desse banco, foi utilizado o software estatístico R (R Core Team (2020)), realizando uma análise descritiva com o auxílio do pacote Tidyverse (Wickham et al. (2019)) e aplicando as metodologias necessárias com o auxílio do pacote MASS (Venables, W. N. & Ripley, B. D. (2002)).

4.2 Análise descritiva

A Tabela 17 contém as medidas resumo das portadoras da herança genética “prima”. Podemos notar que, com exceção de “Gravidez” e “Tríceps” para o caso de diabetes positiva, a maioria das variáveis não tem uma curtose ideal (próxima de 3) e o coeficiente de assimetria da maioria das variáveis também não foram ideais.

Tabela 17: Medidas resumo das variáveis dos pacientes com diabetes positivo e negativo

	Média	Var.	DP	CV(%)	Min.	Med.	Max.	CA	Cur.
Positivo									
Gravidez	4,47	15,34	3,92	87,62	0,00	3,00	17,00	0,81	2,91
Glicose	145,19	890,39	29,84	20,55	78,00	144,50	198,00	-0,14	2,08
Pressão	74,08	169,56	13,02	17,58	30,00	74,00	110,00	-0,13	3,96
Tríceps	32,96	92,98	9,64	29,25	7,00	33,00	63,00	0,10	3,01
Insulina	206,85	17609,26	132,70	64,15	14,00	169,50	846,00	1,86	7,31
Massa	35,78	45,36	6,73	18,82	22,90	34,60	67,10	1,36	6,78
Idade	35,94	113,10	10,63	29,59	21,00	33,00	60,00	0,57	2,20
Negativo									
Gravidez	2,72	6,85	2,62	96,20	0,00	2,00	13,00	1,49	5,23
Glicose	111,43	607,23	24,64	22,11	56,00	107,50	197,00	0,72	3,51
Pressão	68,97	141,44	11,89	17,24	24,00	70,00	106,00	-0,16	3,68
Tríceps	27,25	108,87	10,43	38,29	7,00	27,00	60,00	0,35	2,49
Insulina	130,85	10532,13	102,63	78,43	15,00	105,00	744,00	2,50	11,81
Massa	31,75	46,17	6,79	21,40	18,20	31,25	57,30	0,43	3,13
Idade	28,35	80,80	8,99	31,71	21,00	25,00	81,00	2,20	9,13

Podemos notar correlações consideráveis para os pares “Gravidez” e “Idade”, “Massa” e “Tríceps” e “Insulina” e “Glicose” sendo que este último par só possui valores consideráveis para o grupo das pacientes sem diabetes. A Tabela 18 contém a matriz de correlação das pacientes com e sem diabetes.

As Figuras 18 e 20, que estão no Apêndice e a Figura 15 contêm, respectivamente, os boxplots, os QQplots e os histogramas de cada variável das pacientes com diabetes. Além disso, as Figuras 19 e 21, que estão no Apêndice e a Figura 16 contém, respectivamente, os boxplots, os QQplots e os histogramas de cada variável das pacientes sem diabetes. Podemos perceber, pelas figuras, a existência de pelo menos uma variável que não segue uma distribuição normal. Sendo assim, rejeita-se a normalidade multivariada.

Tabela 18: Matriz de correlações dos pacientes com diabetes positivo

Positivo							
	Gravidez	Glicose	Pressão	Tríceps	Insulina	Massa	Idade
Gravidez	1,00	0,02	0,18	-0,07	-0,00	-0,19	0,60
Glicose	0,02	1,00	0,09	0,04	0,40	-0,03	0,19
Pressão	0,18	0,09	1,00	0,16	-0,07	0,19	0,27
Tríceps	-0,07	0,04	0,16	1,00	0,05	0,59	-0,12
Insulina	-0,00	0,40	-0,07	0,05	1,00	-0,02	0,22
Massa	-0,19	-0,03	0,19	0,59	-0,02	1,00	-0,18
Idade	0,60	0,19	0,27	-0,12	0,22	-0,18	1,00
Negativo							
	Gravidez	Glicose	Pressão	Tríceps	Insulina	Massa	Idade
Gravidez	1,00	0,14	0,17	0,10	0,01	-0,04	0,70
Glicose	0,14	1,00	0,16	0,11	0,62	0,15	0,21
Pressão	0,17	0,16	1,00	0,21	0,12	0,31	0,24
Tríceps	0,10	0,11	0,21	1,00	0,15	0,66	0,20
insulin	0,01	0,62	0,12	0,15	1,00	0,27	0,05
Massa	-0,04	0,15	0,31	0,66	0,27	1,00	0,06
Idade	0,70	0,21	0,24	0,20	0,05	0,06	1,00

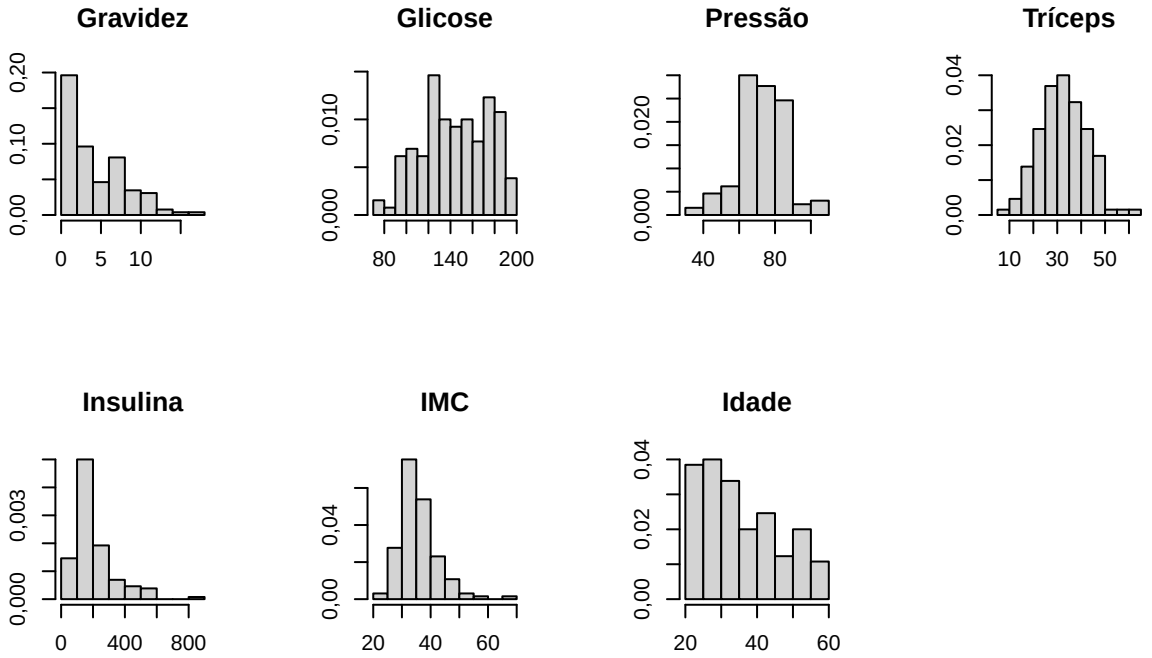

Figura 15: Histogramas das variáveis dos pacientes com diabetes positivo.

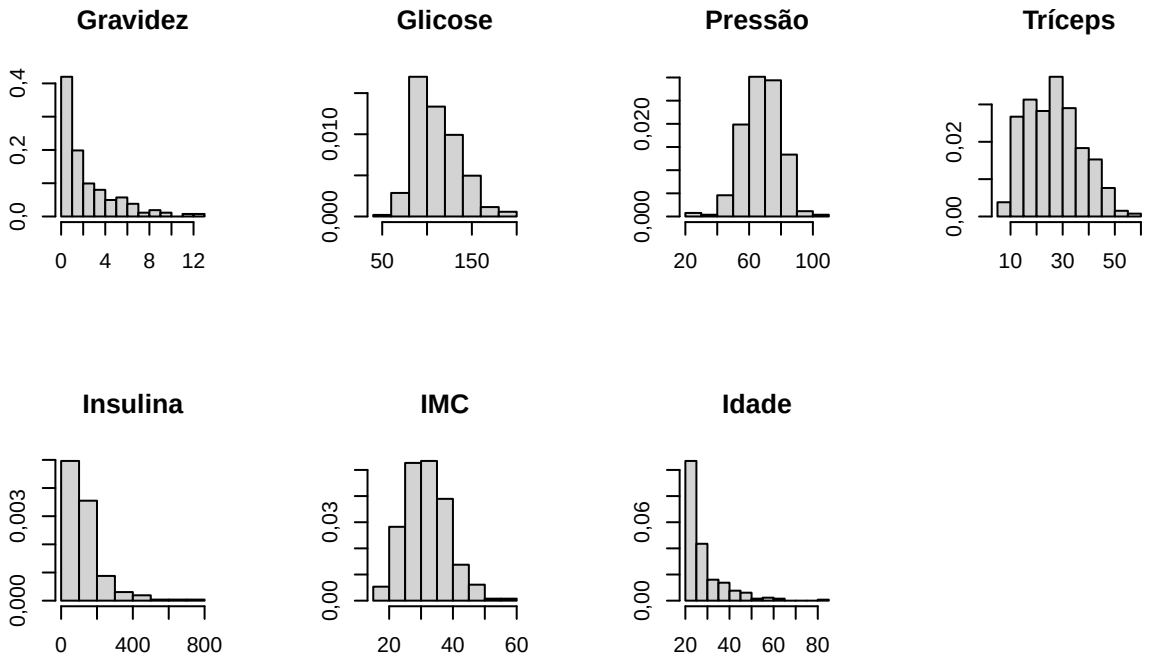

Figura 16: Histogramas das variáveis dos pacientes com diabetes negativo.

4.3 Análise Inferencial

A análise discriminante para duas populações foi a metodologia escolhida para a classificação entre diabetes positiva e diabetes negativa. Sabe-se que existem suposições para o modelo. Dois métodos equivalentes, um que supõe normalidade multivariada e homocedasticidade entre as duas populações e outro que somente supõe homocedasticidade entre as duas populações (Método de Fisher).

Um teste de homocedasticidade, com nível de significância de 0,05, retornou um p-valor de, aproximadamente, 0,0228. Portanto, rejeitamos a hipótese de homocedasticidade entre as duas populações.

Como nenhum dos dois modelos tem suas suposições respeitadas, não há diferença entre eles nesse problema.

A probabilidade à priori foi definida como a proporção de observações de cada categoria do banco de dados (cerca de 67% sem diabetes e 33% com diabetes). Mais informações sobre a metodologia podem ser encontradas em Azevedo (2020).

O banco de dados foi separado em treino e teste. No conjunto de treino, a análise discriminante foi aplicada. A Tabela 19 contém a matriz de confusão da análise discriminante no conjunto de teste. Com isso, a taxa de erro aparente foi de, aproximadamente, 18% e a taxa de erro ótimo foi de, aproximadamente, 25%.

A Tabela 20 contém as medidas resumo das funções discriminantes. O grupo positivo tem a média negativa para a função discriminante, enquanto que o grupo negativo tem a média positiva. Na

Tabela 19: Matriz de confusão do teste

	Negativo	Positivo
Negativo	119	12
Positivo	24	41

Figura 17 temos os boxplots das funções discriminantes (A) e as densidades estimadas das funções discriminantes (B), respectivamente. Através delas podemos notar que grande parte dos valores das funções discriminantes estão sobrepostos. Com isso, a classificação se torna confusa, obtendo a taxa de erro aparente e taxa de erro ótimo altos.

Tabela 20: Medidas resumo das funções discriminantes

	Grupo	Média	DP	Var.	Mínimo	Mediana	Máximo	n
1	Negativo	0,20	1,27	1,61	-2,45	0,09	2,91	131
2	Positivo	-0,57	0,90	0,82	-2,49	-0,71	1,96	65

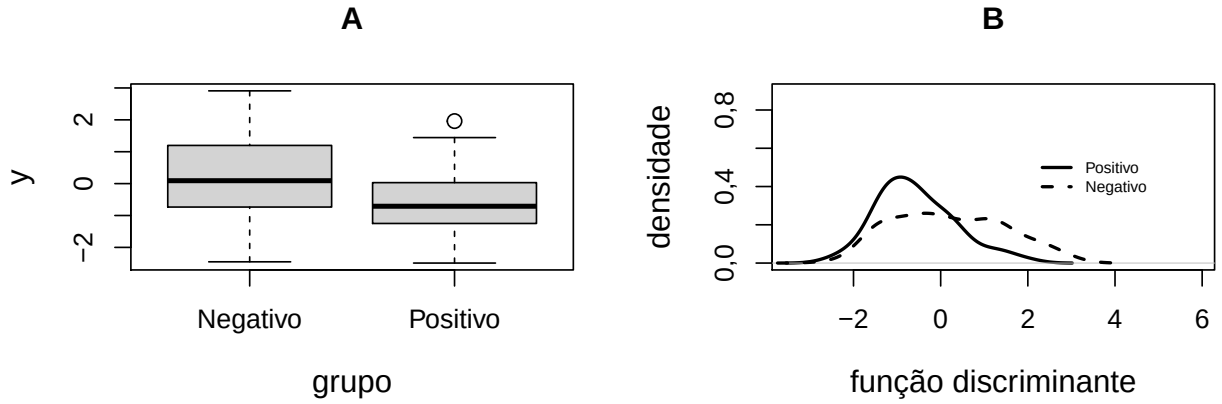

Figura 17: Boxplots (A) e Densidades estimadas (B) das funções discriminantes.

4.4 Conclusão

Com a função discriminante calculada, para ambos os grupos, é possível realizar a classificação entre os mesmos. Como visto em Azevedo (2020), quanto maior forem as diferenças entre os vetores de médias dos grupos, temos que a probabilidade total de classificação incorreta (PTCI) diminuirá e quanto menor a diferença entre o vetor de médias, maior será a PTCI.

Portanto, para o conjunto de dados em questão, como os vetores de médias são semelhantes, como visto na Tabela 17, a classificação não ficará muito alta e apresentará TEA e TOE relativamente altas.

Referências

Azevedo, C. L. N. (2020). Notas de aula sobre análise discriminante. https://www.ime.unicamp.br/~cnaber/Material_AM_2S_2020.htm.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686.

Venables, W. N. & Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York.

Apêndice

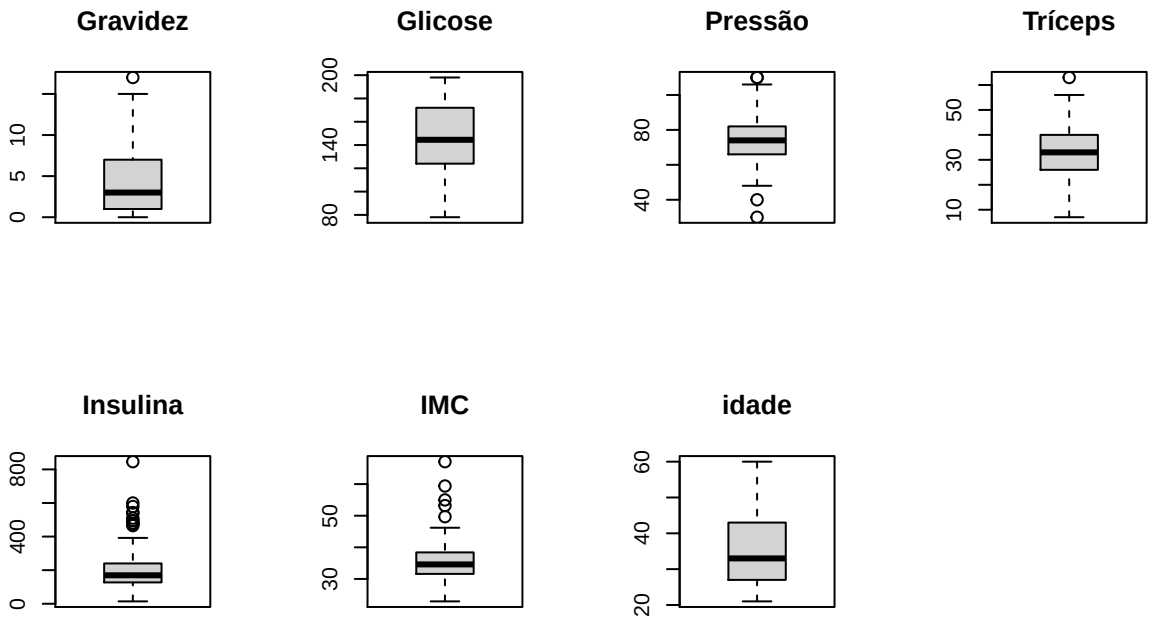

Figura 18: Boxplots das variáveis dos pacientes com diabetes positivo.

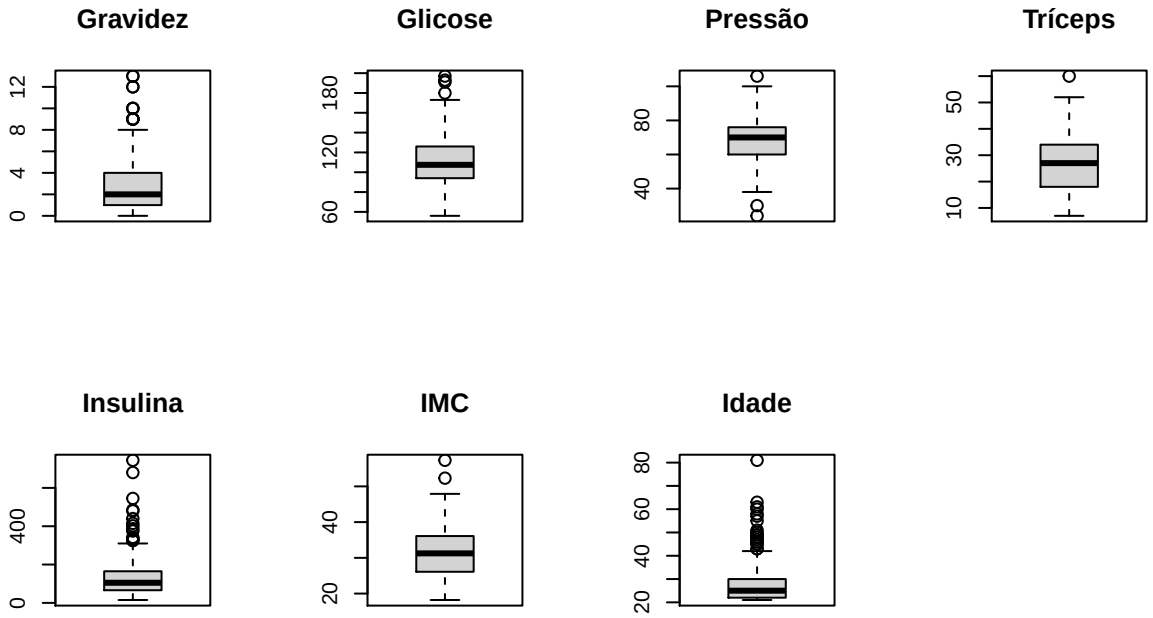

Figura 19: Boxplots das variáveis dos pacientes com diabetes negativo.

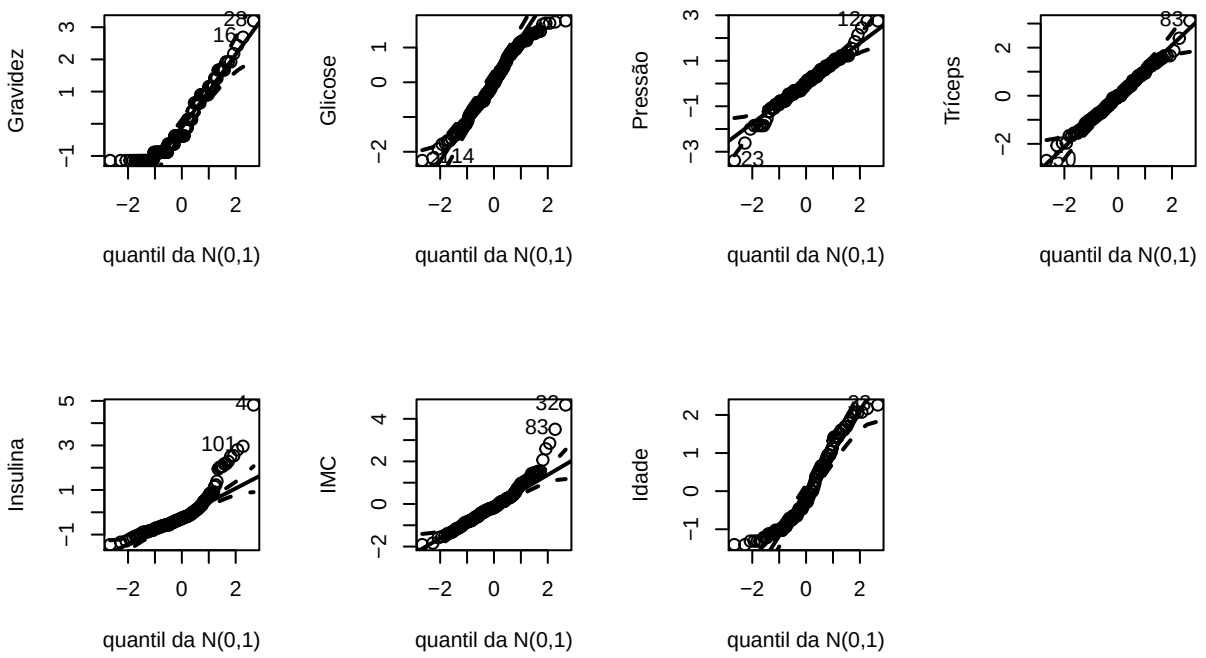

Figura 20: QQplots das variáveis dos pacientes com diabetes positivo.

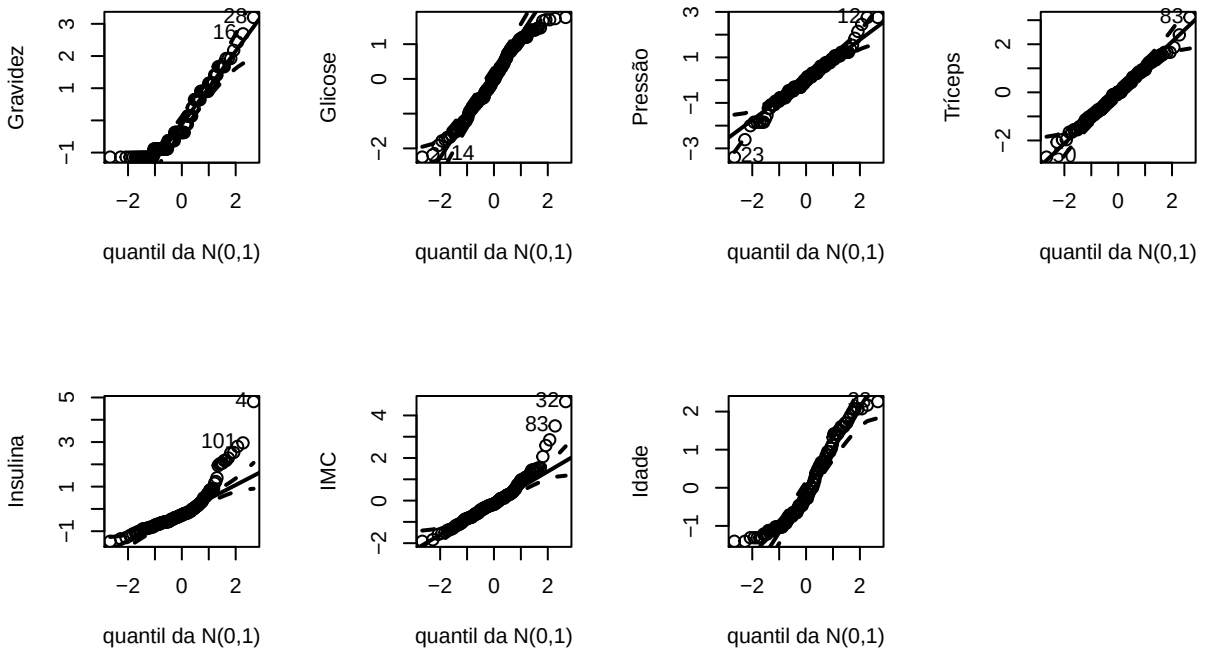

Figura 21: QQplots das variáveis dos pacientes com diabetes negativo.