Análise de Componentes Principais

Caroline Vasconcelos

Relatório

Dados Hematológicos

A análise de componentes principais é um método estatístico versátil para reduzir uma tabela de dados, por variável, às suas características essenciais, chamadas componentes principais. Os componentes principais são combinações lineares das variáveis originais que explicam ao máximo a variância de todas as variáveis. No processo, o método fornece uma aproximação da tabela de dados original usando apenas poucos componentes principais. Utilizaremos esta técnica para analisar um conjunto de dados hematológicos de 51 trabalhadores, as variáveis são as seguintes:

y_1 = concentração de hemoglobina

y_2 = volume globular

y_3 = contagem de glóbulos brancos

y_4 = contagem de linfócitos

y_5 = contagem de neutrófilos

y_6 = concentração sérica de chumbo

Para iniciar a análise dos dados, é de fundamental importância investigar como cada uma dessas variáveis se comporta, o que nos conduz para uma análise descritiva univariada de cada uma delas, ou mesmo bivariada (Figure 1), para então partimos para uma análise multivariada desses dados.

Para se chegar a uma conclusão sobre a normalidade dos dados, além da leitura gráfica, faz-se necessário realizar também os testes de normalidade. Para tanto, foi usado o pacote MVN, que é um pacote do software R que avalia não somente a normalidade multivariada, bem como, traz ferramentas para analisar as variáveis individualmente.

Foram feitos os testes de Shapiro-Wilk Univariado, um dos mais utilizados devido ao fato dele ser o mais sensível para diversas distibuições, e o teste de Anderson-Darling, que é baseado na função de distribuição empírica, para distribuições com parâmetros desconhecidos.

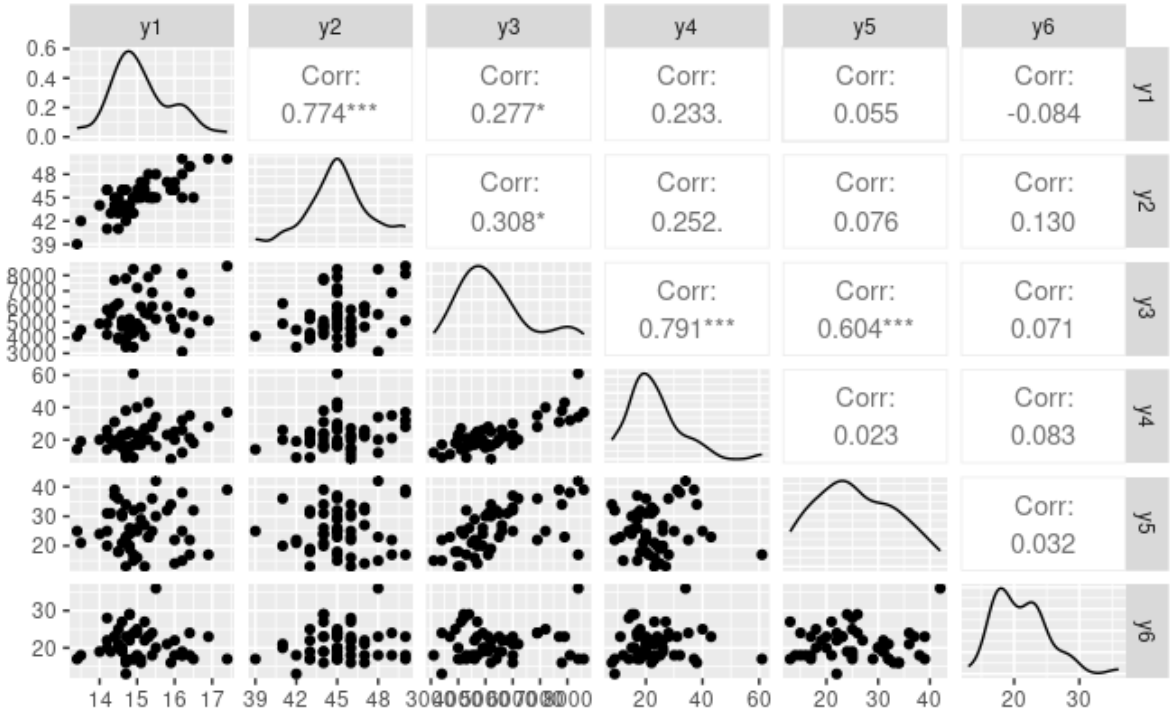

Figure 1: Gráficos bivariados e correlações

Table 1: Shapiro-Wilk Univariado

Variáveis	p-valor	Média	Desvio-Padrão
y_1	0.1203	15.1	0.83
y_2	0.1427	45.2	2.32
y_3	0.0027	53.8	1416.57
y_4	0.008	23.0	9.68
y_5	0.2076	25.6	7.62
y_6	0.0061	21.0	4.26

Table 2: Anderson-Darling

Variável	p-valor	Assimetria	Curtose
y_1	0.0330	0.535	0.037
y_2	0.0415	0.029	0.147
y_3	0.0013	0.759	-0.335
y_4	0.0078	1.350	2.948
y_5	0.2885	0.222	-0.957
y_6	0.01530	0.937	1.281

O teste de Shapiro-Wilk Univariado nos forneceu não somente o p-valor, que indicou normalidade para três variáveis (y_1, y_2 e y_5), como também as médias e os desvio-padrão para cada uma delas. No teste de Anderson-Darling, além do p-valor, que indicou normalidade para apenas uma variável (y_5), ainda observamos dados de assimetria e curtose.

Para os testes multivariados foram usados dois pacotes, o pacote MVN, usado anteriormente, e o pacote mvnormtest que faz o teste de Shapiro-Wilk para verificar normalidade multivariada, assim como o MVN também faz.

Table 3: Valor-p para teste de normalidade multivariados.

Pacote	p-valor
mvnormtest	0,00001
MVN	0,000003

Os testes multivariados apresentam evidências para rejeitar a hipóstese nula, ou seja, existem evidências para rejeitar a hipótese de que o conjunto de dados apresenta normalidade.

Dessa forma, a **Transformação Box-cox** se mostra uma excelente técnica para estabilizar a variância, tornar os dados mais semelhantes à distribuição normal, melhorar a validade das

medidas de associação (correlação de Pearson) e outros procedimentos de estabilização de dados. A transformação de Box-Cox é bastante conhecida no meio econométrico, e é usada para enfrentar problemas de heterocedasticidade e/ou falta de normalidade, como é o caso.

Tanto a forma linear quanto a logarítmica são dois casos particulares de uma família mais extensa de transformações não-lineares. A tranformação de potência é definida como uma função de variação contínua, em relação ao parâmetro de potência λ, ou seja, x^{λ} . É definida por:

$$f_{\lambda}(x) = \frac{x^{\lambda} - 1}{\lambda}, \ \lambda \neq 0$$

$$f_{0}(x) = log(x), \ \lambda = 0$$

Para isto, foi usada a função boxcox do pacote MASS, afim de estimar o parâmetro de transformação λ , usando estimativa de máxima verossimilhança. Seguem os resultados após as transformações:

Table 4: Shapiro-Wilk Univariado.

Variável	p-valor
y_1	0.519
y_2	0.144
y_3	0.451
y_4	0.835
y_5	0.336
y_6	0.352

Table 5: Normalidade Multivariada.

Teste	p-valor
Royston	0.443

Componentes Principais

Table 6: Componentes principais para matriz de covariância

Componentes	λ	Porcentagem explicada	Porcentagem explicada acumulada
Y_1	2.052613e+00	5.887383e-01	58.87383
Y_2	1.796210e+00	3.136980e-01	90.24363
Y_3	2.190240e-01	9.502305e-02	99.74594
Y_4	1.012642e-03	2.540639e-03	100.00000
Y_5	8.027965e-07	3.828063e-09	100.00000
Y_6	2.034894e-08	6.058567e-33	100.00000

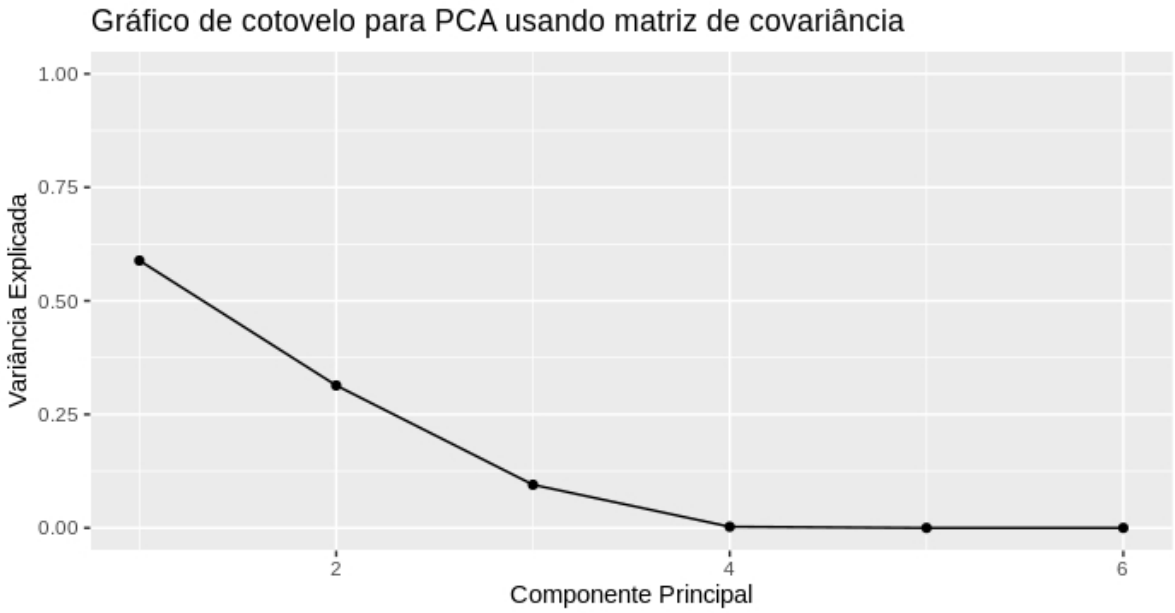

Figure 2: Gráfico de cotovelo usando matriz de covariâncias.

Componentes	λ	Porcentagem explicada	Porcentagem Acumulada
Y_1	2.3829152	4.312625e-01	43.12625
Y_2	1.4256192	3.278497e-01	75.91121
Y_3	1.1051018	2.183552e-01	97.74673
Y_4	0.8488797	1.783034e-02	99.52977
Y_5	0.2005533	4.702324e-03	100.00000
Y_6	0.0369308	1.616938e-32	100.00000

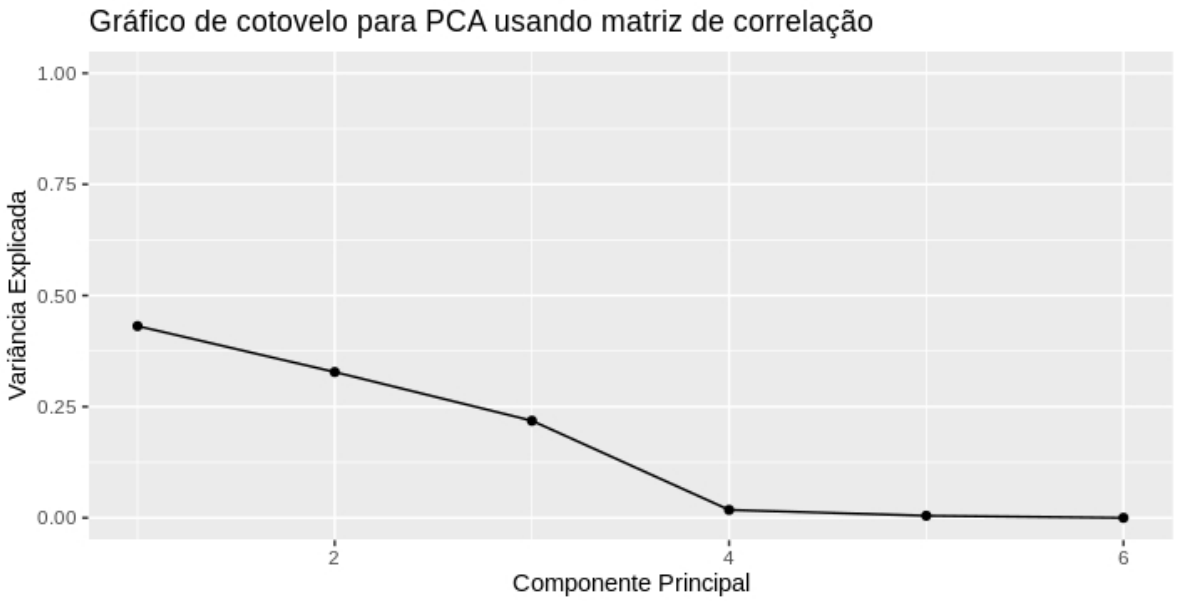

Figure 3: Gráfico de cotovelo usando matriz de correlação.