Análise Multivariada

Lupércio França Bessegato
Dep. Estatística/UFJF

Roteiro

Componentes Principais

Exemplo 8.3

- Pesquisa com 5 variáveis sócio-econômicas
  - √ X_1: população total (milhares)
  - √ X_2: Escolaridade mediana (anos concluídos)
  - √ X_3: Emprego total (milhares)
  - √ X_4: Empregos na área da saúde (centenas)
  - √ X_5: Valor mediano da habitação (x $10.000)
- Dados: *BD_multivariada.xls/pesquisa*

- Vetor de médias amostral ($\bar{x}$)

Variable	Mean
X1_Pop	4,323
X2_escol	14,014
X3_empregos	1,952
X4_saude	2,171
X5_habitacao	2,454

- Matriz de covariâncias amostral (S)

Covariances: X1_Pop; X2_escol; X3_empregos; X4_saude; X5_habitacao

	X1_Pop	X2_escol	X3_empregos	X4_saude	X5_habitacao
X1_Pop	4,307556				
X2_escol	1,683680	1,767473			
X3_empregos	1,802776	0,588026	0,800669		
X4_saude	2,155326	0,177978	1,064828	1,969475	
X5_habitacao	-0,253474	0,175549	-0,158339	-0,356807	0,504380

- A variação amostral pode ser resumida por uma ou duas componentes principais?

	Componentes Principais						
	1		2		3	4	5
	e1	r(y1,xk)	e2	r(y2,xk)	e3	e4	e5
População Total	0,781	0,99	0,671	-0,04	-0,004	-0,542	0,302
Escolaridade Mediana	0,306	0,61	0,764	-0,76	0,162	0,545	0,009
Total de Empregos	0,334	0,98	-0,883	0,12	-0,015	-0,051	-0,937
Empregos Área Saúde	0,426	0,86	-0,579	0,55	-0,220	0,636	0,172
Valor Mediano Habitação	0,054	0,29	0,262	0,49	0,962	0,051	0,025
Variância	6,931		1,785		0,350	0,230	0,014
% Variância Total (acumulada)	74,1		93,2		97,4	99,8	100,0

- Variância amostral é bem resumida por 2 componentes
  - √ redução de 14 observações de 5 variáveis para 14 observações de 2 variáveis
  - √ 1ª. componente: média ponderada de 4 variáveis
  - √ 2ª. componente: contraste entre empregos saúde com média ponderada da escolaridade com valor habitação

- Correlação mede unicamente importância de uma variável individual sem considerar a influência das demais
  - √ No exemplo, os coeficientes de correlação confirmam a interpretação fornecida pelos coeficientes das componentes

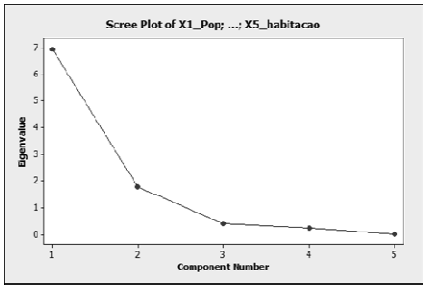
Número de Componentes Principais

- Quantas componentes principais devem ser retidas?
  - √ Não há resposta definitiva
- Considerações a serem tomadas:
  - √ Quantidade explicada de variância amostral total
  - √ Tamanho relativo dos autovalores
    (variância das componentes amostrais)
  - √ Interpretação das componentes

Scree Plot

- Gráfico λ_i *vs. i*
  - √ Procura-se um 'cotovelo' no gráfico
  - √ São consideradas as componentes até o ponto em que os autovalores remanescentes são relativamente pequenos e todos aproximadamente do mesmo valor

- Exemplo 8.3



Exemplo 8.4

- Relação entre tamanho e forma de cascos de tartaruga
  - √ Comprimento
  - √ Largura
  - √ Espessura
  - √ Gênero: male/female
- Análise para as tartarugas macho
- Literatura sugere transformação logarítmica em estudos de relação entre tamanho e forma
- Dados: *BD_multivariada.xls/tartarugas*

- Vetor de médias amostral ($\bar{x}$)

Descriptive Statistics: log_comp_male; log_larg_male; log_esp_male

Variable	Mean
log_comp_male	4,7254
log_larg_male	4,4776
log_esp_male	3,7032

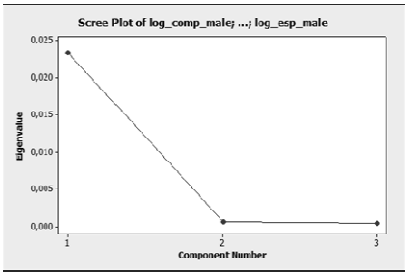
- Matriz de covariâncias amostral (S)

Covariances: log_comp_male; log_larg_male; log_esp_male

	log_comp_male	log_larg_male	log_esp_male
log_comp_male	0,01107200		
log_larg_male	0,00801914	0,00641673	
log_esp_male	0,00815965	0,00600527	0,00677276

- A variação amostral pode ser resumida por uma principal?

- Scree Plot



√ Uma componente principal é claramente dominante

- Componentes principais:

Principal Component Analysis: log_comp_male; log_larg_male; log_esp_male

Eigenanalysis of the Covariance Matrix

Eigenvalue	0,023303	0,000598	0,000360
Proportion	0,961	0,025	0,015
Cumulative	0,961	0,985	1,000

Variable	PC1	PC2	PC3
log_comp_male	0,683	-0,159	-0,713
log_larg_male	0,510	-0,594	0,622
log_esp_male	0,523	0,788	0,324

- Componente adotada:

$$\begin{aligned}\hat{y}_1 &= 0,683\ln(comp) + 0,510\ln(larg) + 0,523\ln(espes) \\ &= \ln\left[(comp)^{0,683}(larg)^{0,510}(esp)^{0,523}\right]\end{aligned}$$

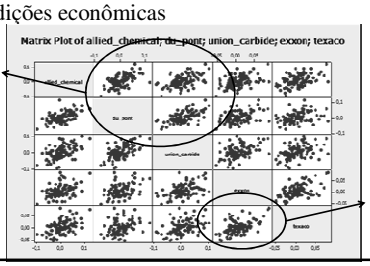
√ ln(volume) de uma caixa com dimensões ajustadas

Exemplo 8.5

- Taxas de retorno de 5 ações negociadas na Bolsa de New York
  - √ Período: jan/75 a Dez/76
  - √ Ações:
    - Allied Chemical
    - du Pont
    - Union Carbide
    - Exxon
    - Texaco
  - √ Dados: BD_multivariada.xls/

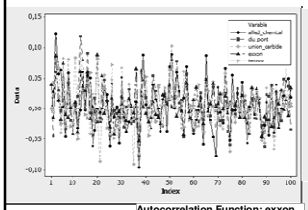
- Taxa de retorno: $\left(\begin{matrix}\text{taxa de}\\ \text{retorno}\end{matrix}\right) = \dfrac{\left(\begin{matrix}\$\text{ fechamento}\\ \text{sexta atual}\end{matrix}\right) - \left(\begin{matrix}\$\text{ fechamento}\\ \text{sexta anterior}\end{matrix}\right)}{\left(\begin{matrix}\$\text{ fechamento}\\ \text{sexta anterior}\end{matrix}\right)}$
- As taxas de retorno entre ativos estão correlacionadas
  - √ ações tendem a se mover juntas em resposta às condições econômicas

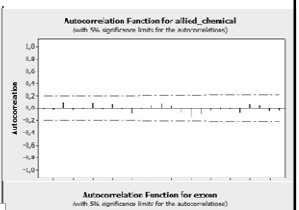
indústria química



indústria óleo & gás

- Observações de 100 semanas aparentam estar distribuídas independentemente





Autocorrelation Function: exxon

Lag	ACF	T	LBQ
1	-0,239413	-2,39	5,91
2	0,233674	2,21	11,59
3	-0,000760	-0,01	11,59
4	-0,086924	-0,79	12,39
5	-0,068469	-0,62	12,90
6	-0,084782	-0,76	13,67
7	-0,012353	-0,11	13,69
8	-0,017693	-0,16	13,73
9	-0,093155	-0,83	14,70
10	0,134497	1,19	16,75
11	-0,087232	-0,76	17,62
12	-0,022943	-0,20	17,68
13	0,012750	0,11	17,70
14	0,109989	0,95	19,14

LFB1

- Vetor de médias amostral ($\bar{x}$)

Descriptive Statistics: allied_chemical; du_pont; union_carbide; exxon; texaco

Variable	Mean
allied_chemical	0,00543
du_pont	0,00483
union_carbide	0,00565
exxon	0,00629
texaco	0,00371

- Matriz de correlação amostral (S)

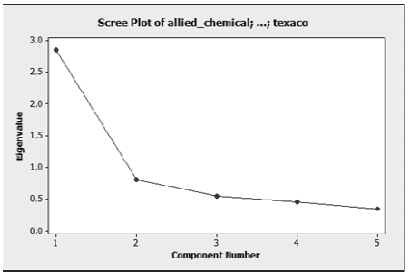
Correlations: allied_chemical; du_pont; union_carbide; exxon; texaco

	allied_chemical	du_pont	union_carbide	exxon
du_pont	0,577			
union_carbide	0,509	0,598		
exxon	0,387	0,390	0,436	
texaco	0,462	0,322	0,426	0,524

- A variação amostral pode ser resumida por uma ou duas componentes principais?

LFB1 Calcular matriz de covariâncias amostral
Há domínio de variabilidade?
Lupércio Bessegato; 20/02/2013

- Scree Plot



√ Aparentemente duas componentes principais resumem bem os dados

- Componentes principais:

Principal Component Analysis: allied_chemi; du_pont; union_carbid; exxon; texac

Eigenanalysis of the Correlation Matrix

Eigenvalue	2,8565	0,8091	0,5400	0,4513	0,3430
Proportion	0,571	0,162	0,108	0,090	0,069
Cumulative	0,571	0,733	0,841	0,931	1,000

Variable	PC1	PC2	PC3	PC4	PC5
allied_chemical	0,464	0,241	0,613	-0,381	-0,453
du_pont	0,457	0,509	-0,178	-0,211	0,675
union_carbide	0,470	0,261	-0,337	0,664	-0,396
exxon	0,422	-0,525	-0,539	-0,473	-0,179
texaco	0,421	-0,582	0,434	0,381	0,387

√ Duas primeiras componentes com 73% da variabilidade amostral padronizada total

- 1ª. componente principal:

$\hat{y}_1 = 0,464z_1 + 0,457z_2 + 0,470z_3 + 0,421z_4 + 0,421z_5$

√ Variáveis:

- z_1: retorno padronizado – Allied Chemical
- z_1: retorno padronizado – du Pont
- z_1: retorno padronizado – Union Carbide
- z_1: retorno padronizado – Exxon
- z_1: retorno padronizado – Texaco

√ Interpretação:

- soma ponderada (índice) das 5 ações
- pesos aproximadamente iguais
- Componente geral do mercado de ações (componente do mercado)

- 2ª. componente principal:

$$\hat{y}_1 = 0,240z_1 + 0,509z_2 + 0,260z_3 - 0,526z_4 - 0,582z_5$$

  √ Interpretação:
  - contraste entre ações de indústrias químicas e de óleo & gás
  - Componente industrial

- Comentários:
  √ A maioria das variações dos ativos devem-se às atividades de mercado (1ª. componente) e atividades industriais não correlacionadas (2ª. componente)
  √ As componente remanescentes não são de simples interpretação
  - coletivamente , representam variação que é provavelmente específica de cada ação

Variáveis Padronizadas – Regra Empírica

- Reter apenas as componentes cujas variâncias (λ_i) são maiores que a unidade
  √ componente que explicam individualmente pelo menos 1/*p* da variância amostral padronizada total
- No caso do exemplo anterior (8.6), pareceu-se sensível reter uma componente (y_2) associada à autovalor menor que a unidade

Gráfico dos Componentes Principais

- Podem:
  - √ revelar observações suspeitas
  - √ fornecer verificações da hipótese de normalidade

- As componentes principais são combinações das variáveis originais:
  - √ Se as obeservações provém de população normal multivariada, é razoável esperar que as componentes sejam aproximadamente normais
- Pode ser necessário verificar se as 1ª.s componentes são aproximadamente normais se eles forem usadas como entrada em análises adicionais
- As últimas componentes principais podem ajudar a apontar observações suspeitas

- Cada observação pode ser expressa como uma combinação linear

Resumo

1. Construa diagrama de dispersão para os pares dos primeiros componentes principais
   - √ Faça também Q-Q plots para os valores amostrais gerados por cada componente principal
   - √ Para ajudar a testar a hipótese de normalidade
2. Construir diagramas de dispersão e Q-Q plots para ás últimas componentes principais.
   - √ Isso ajudará a identificar observações suspeitas

Exemplo 8.7

- Plotando os Componentes Principais dos dados das tartarugas macho:
  - √ $x_1 = \ln(\text{comp})$
  - √ $x_2 = \ln(\text{larg})$
  - √ $x_3 = \ln(\text{esp})$
- Componentes:

$$\begin{aligned}
\hat{y}_1 &= 0.683\ln(x_1 - 4.725) + 0.510\ln(x_2 - 4.478) + 0.523\ln(x_3 - 3.703) \\
\hat{y}_2 &= -0.159\ln(x_1 - 4.725) - 0.594\ln(x_2 - 4.478) + 0.788\ln(x_3 - 3.703) \\
\hat{y}_3 &= -0.713\ln(x_1 - 4.725) + 0.622\ln(x_2 - 4.478) + 0.324\ln(x_3 - 3.703)
\end{aligned}$$

- Comandos Minitab para Q-Q Plot

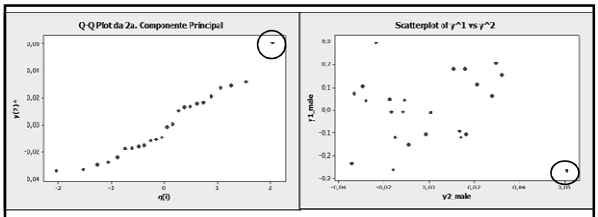
```
Name C30 "(j-1/2)/n"
Set C30
1( 1 : 24 / 1 )1
End.

Let C30 = (C30-0,5)/24     # Cálculo percentagens

Name C31 "q(i)"
Invcdf c30 c31;            # Cálculo quantis
Normal 0 1.

Name C32 "y(2)^"
Sort c25 c32               # Ordenação vetor de dados

Plot C32*C31;              # Scatter plot
Title "Q-Q Plot da 2ª. Componente Principal";
Symbol.
```



- Observação da 1ª. tartaruga é supeita.
  - √ Checar registros ou verificar anomalias na tartaruga
- Excetuado esse dado o scatter plot aparenta estar razoavelmente elíptico
- Verificar os plots dos outros conjunto de componentes principais.

Exercício – Solo

- Análise de solo
  - √ 20 amostras
  - √ Variáveis:
    - areia (%)
    - sedimentos (%)
    - argila (%)
    - qte. material orgânico (%)
    - acidez do solo (pH)
  - √ Banco de dados: *BD_multivariada.xls/solo*

- Matriz de covariâncias amostral (S)

Covariances: areia; sedimentos; argila; morganico; ph

	areia	sedimentos	argila	morganico	ph
areia	138,32674				
sedimentos	-102,12274	79,73818			
argila	-36,20400	22,38455	13,81945		
morganico	-0,94221	1,52661	-0,58439	0,64345	
ph	-0,13579	0,11079	0,02500	0,03237	0,26263

- Autovalores de S

Eigenvalues

223,841	8,218	0,472	0,258	0,000

√ S é singular pois $\lambda_5 = 0$ ($|S| = 0$)

($X_1 + X_2 + X_3 = 100\%$)

- Componentes principais (p=5)

Principal Component Analysis: areia; sedimentos; argila; morganico; ph

Eigenanalysis of the Covariance Matrix

Eigenvalue	223,84	8,22	0,47	0,26	0,00
Proportion	0,962	0,035	0,002	0,001	0,000
Cumulative	0,962	0,997	0,999	1,000	1,000

Variable	PC1	PC2	PC3	PC4	PC5
areia	-0,785	0,223	-0,027	-0,004	-0,577
sedimentos	0,587	0,561	-0,086	-0,010	-0,577
argila	0,198	-0,784	0,113	0,014	-0,577
morganico	0,007	0,146	0,980	0,136	0,000
ph	0,001	0,002	0,137	-0,991	-0,000

√ y5 é constante para qualquer observação j

y5 = 0,577 (100)

√ Qualquer das três variáveis poderia ser eliminada

- Eliminada X_1 (areia)
  - √ maior variância amostral
    - tenderia dominar primeira componente
- Matriz de covariâncias amostral (S)

Covariances: sedimentos; argila; morganico; ph

79,7382	22,3846	1,52661	0,110789
22,3846	13,8194	-0,58439	0,025000
1,5266	-0,5844	0,64345	0,032368
0,1108	0,0250	0,03237	0,262632

- Autovalores de S

Eigenvalues

86,6403	7,0936	0,4714	0,2584

- Componentes principais (p = 4 – eliminada X_1)

Principal Component Analysis: sedimentos; argila; morganico; ph

Eigenanalysis of the Covariance Matrix

Eigenvalue	86,640	7,094	0,471	0,258
Proportion	0,917	0,075	0,005	0,003
Cumulative	0,917	0,992	0,997	1,000

Variable	PC1	PC2	PC3	PC4
sedimentos	0,956	-0,288	0,059	0,006
argila	0,294	0,945	-0,142	-0,018
morganico	0,015	-0,154	-0,979	-0,136
ph	0,001	-0,002	-0,137	0,991

√ Duas primeiras componentes explicam 99,2% da variância total

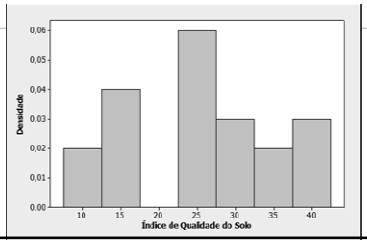
- 1ª. Componente: Índice de qualidade do solo em termos de % sedimentos e argila
  - sedimentos é a variável mais importante
- 2ª. Componente: Comparação entre % de sedimentos e % de argila
  - argila tem peso maior na componente
- 3ª. Componente: variável material orgânico

- Scores da 1ª. Componente
  - √ Índice de qualidade do solo em termos de sedimentos e areia

    (material orgânico tem pouco participação)

Descriptive Statistics: Scores y1

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Scores Y1	20	0	24,86	2,08	9,31	11,57	14,77	24,29	32,43	39,70



- Diferença de escala e unidades da variáveis
  - √ Recomendável padronização para análise de componentes

- Componentes principais (p=4) – Matriz de correlação

Principal Component Analysis: sedimentos; argila; morganico; ph

Eigenanalysis of the Correlation Matrix

Eigenvalue	1,6757	1,1461	0,9601	0,2181
Proportion	0,419	0,287	0,240	0,055
Cumulative	0,419	0,705	0,945	1,000

Variable	PC1	PC2	PC3	PC4
sedimentos	0,710	0,182	-0,147	-0,664
argila	0,702	-0,241	0,111	0,661
morganico	0,025	0,836	-0,423	0,349
ph	0,042	0,459	0,887	-0,026

Referências

Bibliografia Recomendada

- JOHNSON, R. A.; WINCHERN, D. W. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1998
- MINGOTI, D.C. *Análise de Dados através de Métodos de Estatística Multivariada*. Ed. UFMG, 2005.
- LATTIN, J.; CARROLL, J. D.; GREEN, P. E. *Análise de Dados Multivariados*. Cengage, 2011.