Estatística Aplicada à Administração II

Tópico:

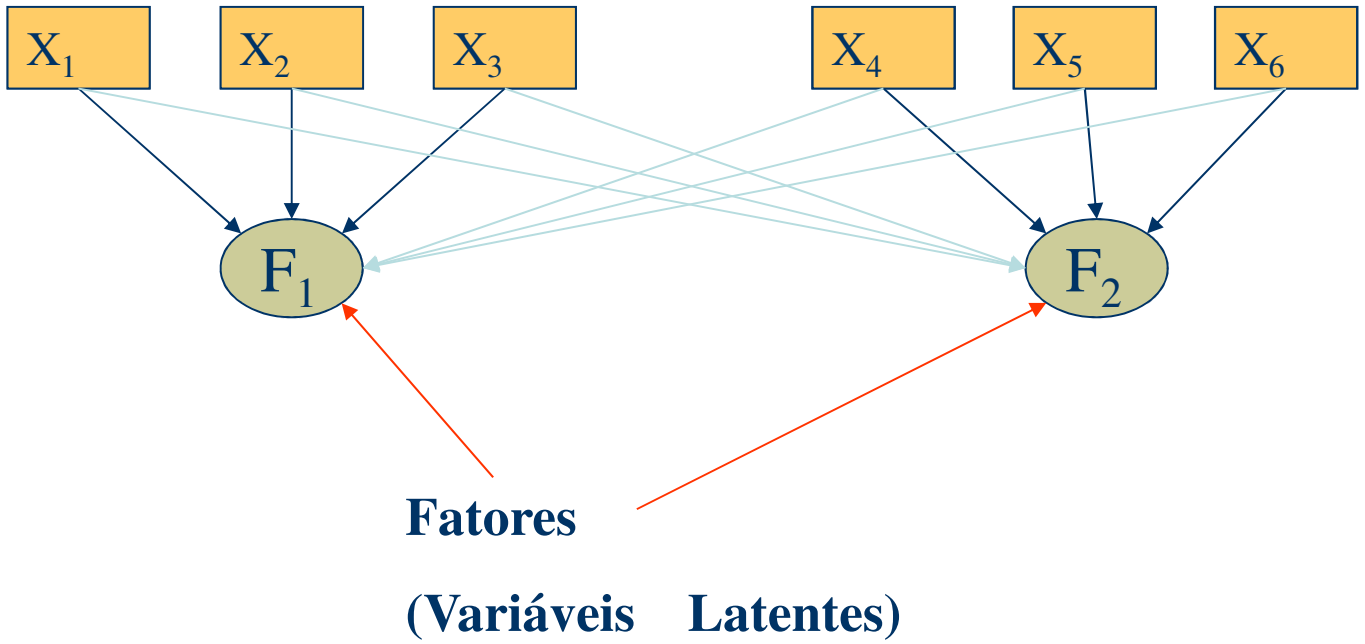
Análise Fatorial

Bibliografia:

- R.A. **Johnson**, Applied Multivariate Statistical Analysis, Prentice Hall, 1992
- L.J. **Corrar**; E. **Paulo**; J.M. **Dias Filho**, Analise multivariada ..., Atlas, 2007
- L.P. **Favero**; et al, Análise de dados - modelagem multivariada ..., Campus, 2009.

Análise Fatorial

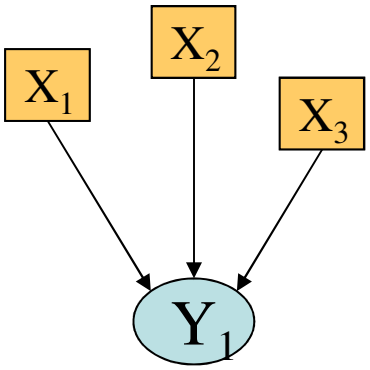
A Análise Fatorial tem como principal objetivo descrever um **conjunto de variáveis originais** através da criação de um número menor de variáveis (**fatores**). Os fatores são variáveis hipotéticas que explicam parte de variabilidade total dos dados.



Análise Fatorial – Técnica de interdependência

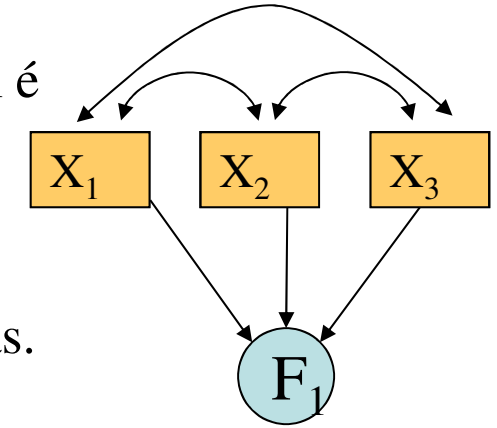
Técnicas de dependência × interdependência

• Qual a diferença entre Análise Fatorial e as técnicas de análise de dependência (regressão múltipla, discriminante, etc.)?



Técnica de dependência: Análise das variáveis independentes para determinar a capacidade de previsão da variável dependente

Cada variável é explicada levando em consideração todas as outras.



Técnica de interdependência: Análise das variáveis com o intuito de maximizar o poder de explicação do conjunto de variáveis

Análise Fatorial - Utilização

- Identificar dimensões latentes (fatores) que expliquem correlações entre um conjunto de variáveis.

- Identificar um conjunto menor de variáveis não correlacionadas para substituir um conjunto original de variáveis correlacionadas.

- Identificar em um conjunto maior, um conjunto menor de variáveis que se destacam para uso em um análise multivariada posterior.

Análise Fatorial - Charles Spearman

Análise fatorial: começou com Charles Spearman analisando correlações entre notas de várias disciplinas cursadas por estudantes:

	Cl (X1)	**Fr** (X2)	**In** (X3)	**Ma** (X4)	**DT** (X5)	**Um** (X6)
Cl	1	0,83	0,78	0,70	0,66	0,63
Fr	0,83	1	0,67	0,67	0,65	0,57
In	0,78	0,67	1	0,64	0,54	0,51
Ma	0,70	0,67	0,64	1	0,45	0,51
DT	0,66	0,65	0,54	0,45	1	0,40
Mu	0,63	0,57	0,51	0,51	0,40	1

Cl: clássicos
Fr: Francês
In: Inglês
Ma: matemática
DT: Discriminação de tom
Mu: Música

Matriz de correlação (*C. Spearman, Am. J. Psychol. 15, 201, 1904.*)

Análise Fatorial - Charles Spearman

A idéia básica de Spearman era que a correlação entre as variáveis poderia ser explicada pela dependência comum delas **com uma variável latente** que ele chamou de fator, ou seja, no caso das seis disciplinas:

$$\mathbf{x}_i = a_i \mathbf{F} + \boldsymbol{\varepsilon}_i \qquad i = 1,2,...,6$$

Onde:

$\mathbf{x}_i$: variável analisada padronizada (**média zero e desvio-padrão 1**)

a_i: é uma constante (**carga fatorial**)

$\mathbf{F}$: **Fator** (com média zero e desvio-padrão 1) comum a todas as variáveis

$\boldsymbol{\varepsilon}_i$: erro (que é específico de cada variável)

Análise Fatorial - Charles Spearman

Conclusão de Spearman:

Cada resultado do teste (notas nas disciplinas) é composto de duas partes:

- Uma parte que é comum a todos os testes (disciplinas), contida no fator comum **F** ao qual foi dado o nome de "inteligência geral".
- Uma parte que é específica de uma dada disciplina.

Análise Fatorial – Modelo ortogonal

Caso geral de **p variáveis** $\mathbf{x}_i$ $(i = 1, 2, 3,..,p)$ padronizadas e **m fatores** o modelo de análise de fatores é:

$$\mathbf{x}_i = a_{i1}\,\mathbf{F}_1 + a_{i2}\,\mathbf{F}_2 + a_{i3}\,\mathbf{F}_3 + ... + a_{im}\,\mathbf{F}_m + \boldsymbol{\varepsilon}_i \,, \qquad i = 1, 2, 3,..,p$$

Onde:

- a_{ij} são cargas fatoriais, medem o grau de correlação entre a variável original ($\mathbf{x}_i$) e os fatores.

- $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_m$ são os **m** fatores não correlacionados, cada um com média zero e variância unitária, ou seja,

$\mathrm{Var}(\mathbf{F}_i) = 1$ e $\mathrm{Corr}(\mathbf{F}_i\ \mathbf{F}_j) = 0\ (i \neq j)$.

- $\boldsymbol{\varepsilon}_i$: é um fator específico para a i-ésima variável, tem média zero e não é correlacionado com qualquer um dos fatores comuns, ou seja, $\mathrm{Corr}(\mathbf{F}_i\ \boldsymbol{\varepsilon}_j) = 0$

Análise Fatorial – A variância de x_i

$$\mathbf{x}_i = a_{i1}\,\mathbf{F}_1 + a_{i2}\,\mathbf{F}_2 + a_{i3}\,\mathbf{F}_3 + \ldots + a_{im}\,\mathbf{F}_m + \boldsymbol{\varepsilon}_i$$

$$\mathrm{Var}(\mathbf{x}_i) = a_{i1}^2\,\underbrace{\mathrm{Var}(\mathbf{F}_1)}_{1} + a_{i2}^2\,\underbrace{\mathrm{Var}(\mathbf{F}_2)}_{1} + \ldots + a_{im}^2\,\underbrace{\mathrm{Var}(\mathbf{F}_m)}_{1} + \mathrm{Var}(\boldsymbol{\varepsilon}_i)$$

$$\mathrm{Var}(\mathbf{x}_i) = a_{i1}^2 + a_{i2}^2 + \ldots + a_{im}^2 + \mathrm{Var}(\boldsymbol{\varepsilon}_i)$$

Comunalidade de $\mathbf{x}_i$: h_i^2, parte da variância relacionada aos fatores comuns (soma dos quadrados das cargas fatoriais)

Especificidade: parte da variância **não relacionada** aos fatores comuns

$$h_i^2 = a_{i1}^2 + a_{i2}^2 + \ldots + a_{im}^2$$

Análise Fatorial – Correlações

$$\mathbf{x}_i = a_{i1}\,\mathbf{F}_1 + a_{i2}\,\mathbf{F}_2 + a_{i3}\,\mathbf{F}_3 + \ldots + a_{im}\,\mathbf{F}_m + \boldsymbol{\varepsilon}_i$$

Correlação entre $\mathbf{x}_i$ e $\mathbf{x}_j$:

$$\text{Corr}(\mathbf{x}_i\ \mathbf{x}_j) = r_{ij} = a_{i1}\,a_{j1} + a_{i2}\,a_{j2} + \ldots + a_{im}\,a_{jm}$$

Portanto a correlação entre duas variáveis é alta se elas tem altas cargas no mesmos fatores.

Correlação entre $\mathbf{x}_i$ e $\mathbf{F}_j$: $\text{Corr}(\mathbf{x}_i\ \mathbf{F}_j) = a_{ij}$ (carga fatorial)

Como na análise de componentes principais, é interessante obter um número de fatores menor que o número de variáveis. Veremos então o procedimento para realizar uma análise fatorial.

Análise Fatorial (AF) X Componentes Principais (CP)

Semelhança: ambas buscam uma estrutura mais simples de variáveis (menor número de variáveis).

Diferenças:

- **CP** são combinações lineares das variáveis originais. Na **AF**, as variáveis originais são combinações lineares dos fatores.
- Na análise de **CP**, busca-se explicar a variância total dos dados. **Em AF, busca-se explicar as correlações** (ou covariância) entre variáveis.

Uma das maneiras de encontrar os fatores é por componentes principais

Análise Fatorial – x_i explicada em termos de F_j

Na AF, após analisar a correlação entre variáveis, é feito um agrupamento das variáveis num número menor de variáveis (denominadas fatores)

Modelo matemático: Spearman - As variações em $\mathbf{x}_i$ podem ser explicadas a partir de um conjunto de fatores $\mathbf{F}_j$

$$\mathbf{x}_i = a_{i1}\,\mathbf{F}_1 + a_{i2}\,\mathbf{F}_2 + a_{i3}\,\mathbf{F}_3 + \ldots + a_{im}\,\mathbf{F}_m + \boldsymbol{\varepsilon}_i$$

- a_{ij} são as cargas fatoriais que medem o grau de correlação entre a variável original e os fatores.

- a carga fatorial ao quadrado representa o quanto da variação de uma variável é explicado pelo fator:

a_{ij}^2 : variação de $\mathbf{x}_i$ explicado pelo fator $\mathbf{F}_j$

Análise Fatorial – $\mathbf{F}_i$ escrito em termos de $\mathbf{x}_i$

Os fatores, por sua vez, poderiam ser estimados por uma combinação linear das variáveis originais.

$$\mathbf{F}_j = b_{j1}\mathbf{x}_1 + b_{j2}\mathbf{x}_2 + \ldots + b_{ji}\mathbf{x}_i$$

Sendo.

- $\mathbf{F}_j$ os fatores comuns não relacionados entre si;

- b_{ji} os **coeficientes dos escores fatoriais** e

- $\mathbf{x}_i$ as variáveis originais envolvidas no estudo.

O **escore fatorial** é um número resultante da multiplicação dos coeficientes b_{ji} pelo valor das variáveis originais .

Análise Fatorial – O escore fatorial (nova variável)

Considerando a expressão para $\mathbf{F}_j$:

$$\mathbf{F}_j = b_{j1}\mathbf{x}_1 + b_{j2}\mathbf{x}_2 + \ldots + b_{ji}\mathbf{x}_i$$

podemos obter o escore fatorial para todas as observações

Ou seja, para uma determinada observação, o **escore fatorial** é um número resultante da multiplicação dos coeficientes b_{ji} pelo valor das variáveis originais desta observação.

Análise Fatorial – Exemplo clientes de um banco

Na Análise de Componentes Principais vimos um exemplo:

SSE OC

$$\mathbf{X} = \begin{bmatrix} 7 & 6 \\ 1 & 4 \\ 3 & 4 \\ 3 & 5 \\ 6 & 6 \\ 4 & 5 \end{bmatrix}$$

Matriz Covariância

$$\mathbf{S} = \begin{bmatrix} 4{,}8 & 1{,}8 \\ 1{,}8 & 0{,}8 \end{bmatrix}$$

Matriz Correlação

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}919 \\ 0{,}919 & 1 \end{bmatrix}$$

Variância Total = $\mathrm{Tr}(\mathbf{S})$ = 4,8 + 0,8 = **5,6**

Autovalores de $\mathbf{S}$

$$\lambda_1 = 5{,}491$$

$$\lambda_2 = 0{,}109$$

Autovetores

$$\mathbf{e}_1 = \begin{bmatrix} 0{,}9336 \\ 0{,}3583 \end{bmatrix} \quad \mathbf{e}_2 = \begin{bmatrix} 0{,}3583 \\ -0{,}9336 \end{bmatrix}$$

Componentes Principais:

$$CP_1 = 0{,}9336\ X_1 + 0{,}3583\ X_2$$

$$CP_2 = 0{,}3583\ X_1 - 0{,}9336\ X_2$$

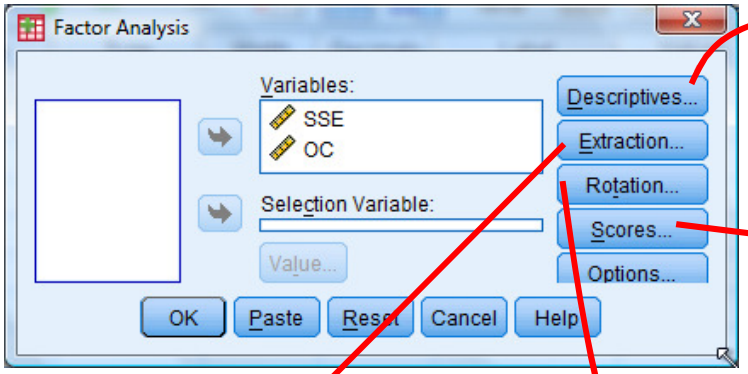
Correlações entre CP e x

	SSE	OC
CP_1	0,9985	0,9387
CP_2	0,0540	-0,3446

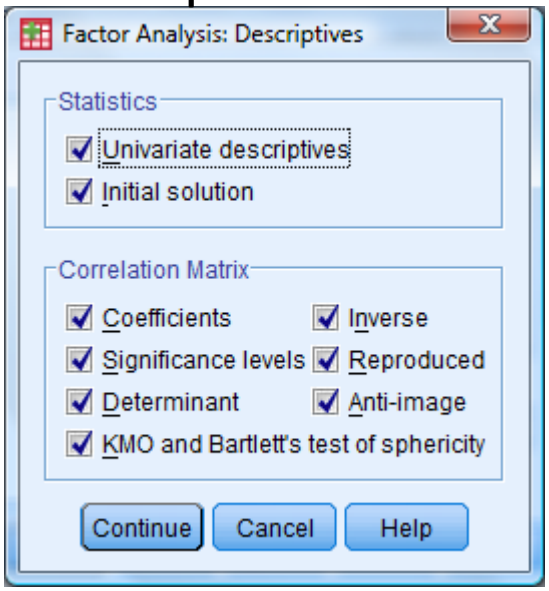
Análise Fatorial – Exemplo clientes de um banco (SPSS)

Na Análise Fatorial, utilizaremos o SPSS para este exemplo:

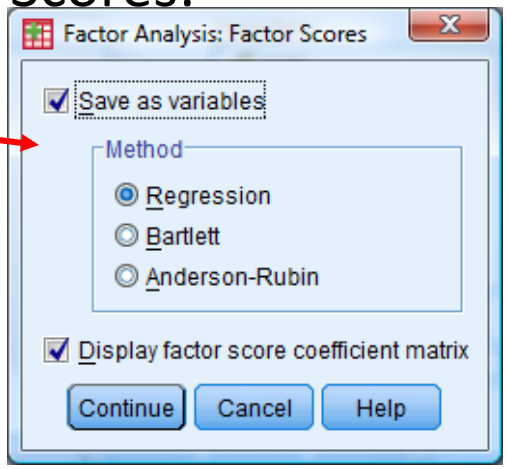
Analyze > Dimension Reduction > Factor:



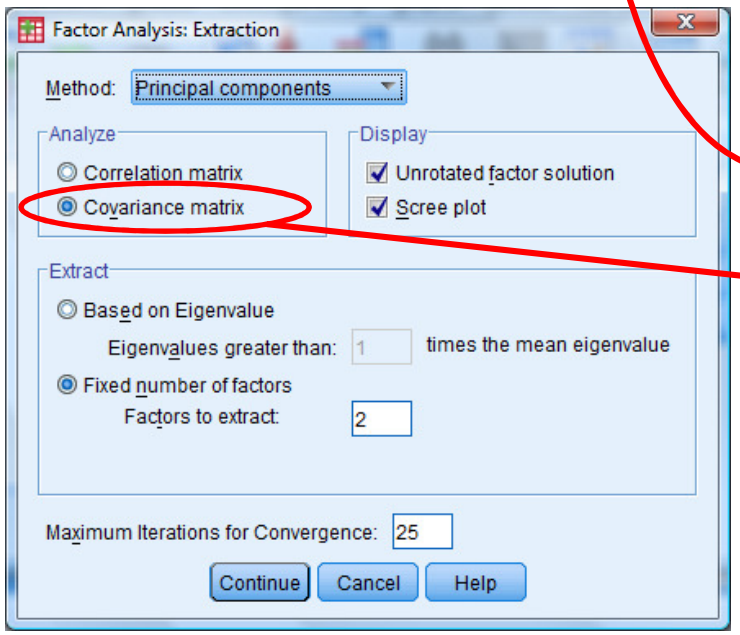
Descriptives:



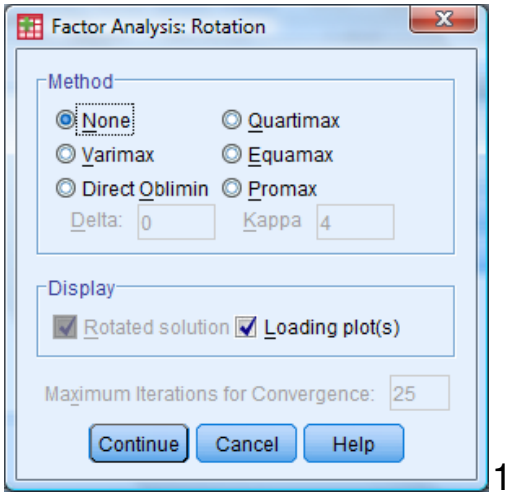
Scores:



Extraction:



Rotation:



Importante:
Em geral é analisada a **matriz de correlação**!
Neste exemplo utilizei a análise pela **matriz de covariância**.

Análise Fatorial – Exemplo clientes de um banco (SPSS)

Resultado AF no SPSS:

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
SSE	4,00	2,191	6
OC	5,00	,894	6

Inverse of Covariance Matrix

	SSE	OC
SSE	1,333	-3,000
OC	-3,000	8,000

Correlation Matrix

		SSE	OC
Correlation	SSE	1,000	,919
	OC	,919	1,000
Sig. (1-tailed)	SSE		,005
	OC	,005	

Communalities

	Raw		Rescaled	
	Initial	Extraction	Initial	Extraction
SSE	4,800	4,800	1,000	1,000
OC	,800	,800	1,000	1,000

Extraction Method: Principal Component Analysis.

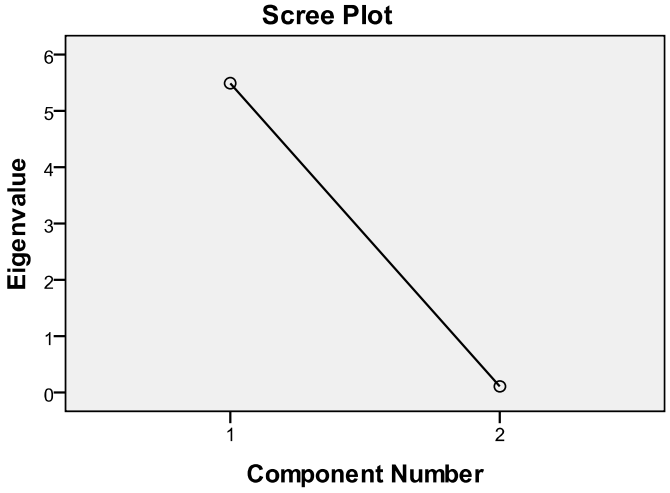
Total Variance Explained

Component		Initial Eigenvalues[a]			Extraction Sums of Squared Loadings		
		Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
Raw	1	5,491	98,049	98,049	5,491	98,049	98,049
	2	,109	1,951	100,000	,109	1,951	100,000
Rescaled	1	5,491	98,049	98,049	1,878	93,901	93,901
	2	,109	1,951	100,000	,122	6,099	100,000

Extraction Method: Principal Component Analysis.

SPSS – Variáveis em termos de Fatores

Resultado AF no SPSS:



Component Matrix[a]

	Raw		Rescaled	
	Component		Component	
	1	2	1	2
SSE	2,188	-,118	,999	-,054
OC	,839	,309	,939	,345

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

	Componente	
	Fator 1	Fator 2
SSE	2,188	-0,118
OC	0,839	0,309

Variáveis em termos de fatores:

$$SSE = 2{,}188\,F_1 - 0{,}118\,F_2 + \varepsilon_1$$

$$OC = 0{,}839\,F_1 + 0{,}309\,F_2 + \varepsilon_2$$

Cargas fatoriais a_{ij}

	Variação Explicada		Comunalidade h_i^2	
SSE	4,787	0,014	4,80	h_1^2
OC	0,704	0,095	0,80	h_2^2
Total	5,491	0,109	5,60	
	λ_1	λ_2		

Autovalores

Cálculo das comunalidades

Para a variável SSE:

$h_1^2 = (2{,}188)^2 + (-0{,}118)^2 = 4{,}80$

Para a variável OC:

$h_2^2 = (0{,}839)^2 + (0{,}095)^2 = 0{,}80$

SPSS – Variáveis em termos de Fatores

Quais variáveis são mais relacionadas a quais fatores?
Olhar a maior correlação (cargas mais altas em módulo) entre uma dada variável e os fatores.

Variáveis em termos de fatores:

$$SSE = 2{,}188\ F_1 - 0{,}118\ F_2$$

$$OC = 0{,}839\ F_1 + 0{,}309\ F_2$$

As duas variáveis são mais relacionadas ao Fator 1.
Obs. A expressão vale para variáveis normalizadas

SPSS – Fatores em termos de Variáveis

Resultado AF no SPSS:

Escore Fatorial (novas variáveis)

$$F_1 = 0{,}873\ Z_{SSE} + 0{,}137\ Z_{OC}$$

$$F_2 = -2{,}374\ Z_{SSE} + 2{,}526\ Z_{OC}$$

	SSE	OC	FAC1_1	FAC2_1
1	7	6	1,34819	-,42706
2	1	4	-1,34819	,42706
3	3	4	-,55133	-1,74051
4	3	5	-,39843	1,08378
5	6	6	,94976	,65673
6	4	5	,0	,0
7				

Component Score Coefficient Matrix[a]

	Component	
	1	2
SSE	,873	-2,374
OC	,137	2,526

Extraction Method: Principal Component Analysis.
Component Scores.

a. Coefficients are standardized.

SSE	OC	Z_SSE	Z_OC	F1	F2
7	6	1,3693	1,1180	1,3482	-0,4271
1	4	-1,3693	-1,1180	-1,3482	0,4271
3	4	-0,4564	-1,1180	-0,5513	-1,7405
3	5	-0,4564	0,0000	-0,3984	1,0838
6	6	0,9129	1,1180	0,9498	0,6567
4	5	0,0000	0,0000	0,0000	0,0000
Média	4	5			
Desv. Pad	2,1909	0,8944			