As componentes principais no descarte de variáveis em um modelo de regressão múltipla

The principal components in the reduction of variables in a multiple regression model

*Jair Mendes Marques**
*Marcos Augusto Mendes Marques***

Resumo

A Análise de Componentes Principais é uma metodologia da Análise Estatística Multivariada cujos principais objetivos são: reduzir o número de dados de um problema e explicar a estrutura da matriz variância-covariância pelas poucas combinações lineares das variáveis originais. Neste artigo, procurou-se utilizar a Análise de Componentes Principais para reduzir o número de variáveis explicativas (independentes) em um modelo de regressão linear múltipla. O método foi aplicado para um conjunto de dados envolvendo seis variáveis explicativas da economia brasileira para os anos de 1980 a 2003. A análise foi realizada com uso do software Matlab e o método utilizado proporcionou a redução das seis variáveis explicativas para apenas uma componente principal.

Palavras-chave: descarte de variáveis; regressão linear múltipla; análise de componentes principais.

Abstract

The Principal Components Analysis is a Multivariate Statistical Analysis methodology whose main objectives are: data reduction and explaining variance-covariance structure through a few linear combinations of the original variables. This paper uses the Principal Components Analysis to reduce the number of predictor (independent) variables in a multiple regression model. The use of this method involved six predictor variables of the Brazilian economy for the years 1980 to 2003. The software Matlab was used in this study and of the six predictor variables just one principal component remained.

Key words: discarding of variables; multiple linear regression; principal components analysis.

* Engenheiro químico e matemático pela Universidade Federal do Paraná - UFPR e doutor em Ciências Geodésicas pela UFPR. Professor do Centro Universitário - FAE Business School.
jair.marques@utp.br
** Engenheiro eletricista pela UFPR e mestrando da UFPR em Métodos Numéricos Aplicados à Engenharia. Bolsista da CAPES.
mmarques@brturbo.com.br

Introdução

Nos modelos de regressão múltipla, é muito comum a ocorrência de variáveis independentes altamente correlacionadas, resultando em coeficientes de regressão estimados com baixa precisão. Nesses casos, é vantajoso o descarte de algumas variáveis com o objetivo de aumentar a estabilidade dos coeficientes de regressão estimados.

Entre as várias alternativas que existem para reduzir a dimensionalidade do modelo, uma delas consiste na utilização de componentes principais. Como nos modelos de regressão, cujo propósito é a explicação da variável dependente, devem-se reter aquelas componentes principais que têm altas correlações com a variável dependente. No caso de um modelo de regressão multivariada, analisam-se as correlações das variáveis independentes com cada uma das variáveis dependentes. Existe uma tendência para os dados com componentes de grandes variâncias de melhor explicar as variáveis dependentes (MARDIA, KENT e BIBBY, 1982).

Quando, no modelo de regressão adotado, as componentes principais tiverem um significado natural e intuitivo, talvez seja melhor expressar o modelo de regressão em termos das componentes principais, caso contrário, é mais conveniente retornar às variáveis originais.

1 Componentes principais

Para investigar as relações entre um conjunto de p variáveis correlacionadas pode ser útil transformar o conjunto de variáveis originais em um novo conjunto de variáveis não-correlacionadas chamadas componentes principais, tendo propriedades especiais em termos de variâncias.

As novas variáveis, as componentes principais, são combinações lineares das variáveis originais e derivadas em ordem decrescente de importância tal que, por exemplo, a primeira componente principal é a combinação linear normalizada com variância máxima (JOHNSON e WICHERN, 1988).

A reprodução da variabilidade total do sistema requer as p variáveis, porém, freqüentemente, a maior parte dessa variabilidade pode ser explicada por um número pequeno $k < p$, de componentes principais. Nesse caso, existe praticamente a mesma quantidade de informações nas k componentes principais que nas p variáveis originais. As k componentes principais podem então substituir as p variáveis originais.

Os principais objetivos das componentes principais são: a) reduzir o número de variáveis; b) analisar quais as variáveis ou quais conjuntos de variáveis explicam a maior parte da variabilidade total, revelando que tipo de relacionamento existe entre eles (BOUROCHE e SAPORTA, 1982).

1.1 Componentes principais populacionais

Algebricamente, as componentes principais são combinações lineares de p variáveis originais:

$X_1, X_2, \ldots, X_p$.

Geometricamente, as combinações lineares representam a seleção de um novo sistema de coordenadas, obtido por rotação do sistema original com $X_1, X_2, \ldots, X_p$ como eixos. Os novos eixos, Y_1, $Y_2, \ldots, Y_p$, representam as direções com variabilidade máxima, permitindo uma interpretação mais simples da estrutura da matriz de covariância.

Por exemplo, para $p = 2$

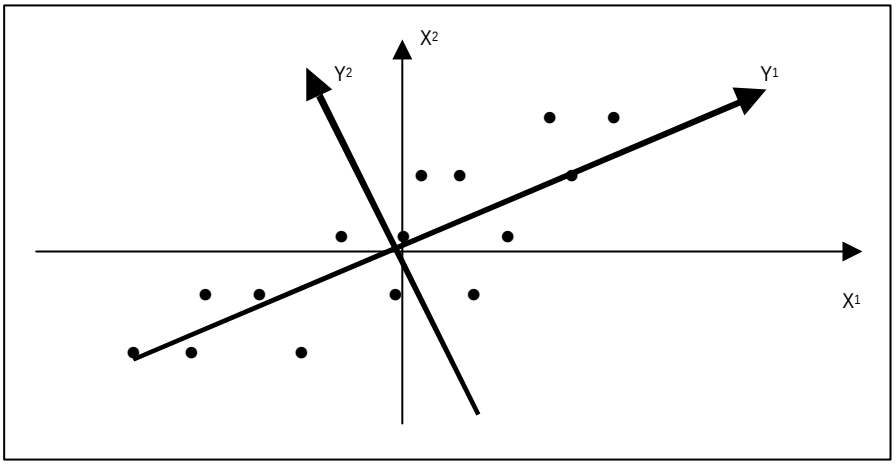

FIGURA 1 - SIGNIFICADO GEOMÉTRICO DAS COMPONENTES PRINCIPAIS PARA p=2

Seja $X'=[X_1,X_2,...,X_p]$ um vetor aleatório p-dimensional com vetor de médias μ, matriz de covariância Σ e autovalores: $\lambda_1 \geq \lambda_2 \geq , ... , \geq \lambda_p$.

Considere as combinações lineares:

$$Y_1 = c'_1 X = c_{11}X_1 + c_{21}X_2 + ... + c_{p1}X_p$$
$$Y_2 = c'_2 X = c_{12}X_1 + c_{22}X_2 + ... + c_{p2}X_p$$
$$..... ....$$
$$Y_j = c'_j X = c_{1j}X_1 + c_{2j}X_2 + ... + c_{pj}X_p$$
$$..... ....$$
$$Y_p = c'_p X = c_{1p}X_1 + c_{2p}X_2 + ... + c_{pp}X_p$$

ou:

$$Y = C'X \qquad (2.1)$$

onde:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ M \\ Y_p \end{bmatrix} \quad \text{e} \quad C = \begin{bmatrix} c_{11} & c_{12} & \Lambda & c_{1p} \\ c_{21} & c_{22} & \Lambda & c_{2p} \\ M & M & \Lambda & M \\ c_{p1} & c_{p2} & \Lambda & c_{pp} \end{bmatrix}$$

com:

$$E(Y_j) = E(c'_j X) = c'_j E(X) = c'_j \mu \qquad (2.2)$$
$$V(Y_j) = V(c'_j X) = c'_j V(X) c_j = c'_j \Sigma c_j \qquad (2.3)$$
$$Cov(Y_i, Y_j) = V(c'_i X, c'_j X) = c'_i \Sigma c'_j \qquad (2.4)$$
$$i \neq j = 1, 2, ... , p.$$

tendo solução normalizada com:

$$c'_j c_j = \begin{bmatrix} c_{1j} & c_{2j} & \Lambda & c_{pj} \end{bmatrix} \begin{bmatrix} c_{1j} \\ c_{2j} \\ M \\ c_{pj} \end{bmatrix} = \sum_{i=1}^{p} c_{ij}^2 = 1$$

Define-se, de acordo com Chatfield e Collins (1992):

- a 1.ª componente principal populacional como a combinação linear $c'_1 X$ que maximiza $V(c'_1 X)$ sujeita à restrição $c'_1 c_1 = 1$.
- a 2.ª componente principal populacional como a combinação linear $c'_2 X$ que maximiza $V(c'_2 X)$ sujeita às restrições $c'_2 c_2 = 1$ e $Cov(c'_1 X, c'_2 X) = 0$.

⋮

- a j-ésima componente principal populacional como a combinação linear $c'_j X$ que maximiza $V(c'_j X)$ sujeita às restrições $c'_j c_j = 1$ e $Cov(c'_j X, c'_i X) = 0$ para todo $i < j$.

1.2 Propriedades das componentes principais populacionais

(1) Seja o vetor aleatório $X'=[X_1, X_2,..., X_p]$ com matriz covariância Σ e pares de autovalores-autovetores (λ_1,e_1), (λ_2,e_2), ... ,(λ_p,e_p), onde $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. A j-ésima componente principal é dada por:

$$Y_j = e'_j X = e_{1j}X_1 + e_{2j}X_2 + ... + e_{pj}X_p \text{ , } j = 1, 2,..., p \quad (2.5)$$

onde:

$$V(Y_j) = e'_j \Sigma e_j = \lambda_j \text{ e } Cov(Y_i, Y_j) = e'_i \Sigma e_j = 0 \text{ , } i \neq j \quad (2.6)$$

(2) Variância total:

$$\sum_{i=1}^{p} V(X_i) = \sigma_1^2 + \sigma_2^2 + ... + \sigma_p^2 = \lambda_1 + \lambda_2 + ... + \lambda_p = \sum_{j=1}^{p} V(Y_j) \quad (2.7)$$

(3) Se $Y_1 = e'_1 X$, $Y_2 = e'_2 X$, ..., $Y_p = e'_p X$ são as componentes principais de Σ então:

$$\rho_{Y_j X_i} = \frac{e_{ij}\sqrt{\lambda_j}}{\sigma_i}, \quad i, j = 1, 2, ... , p \quad (2.8)$$

são os coeficientes de correlação entre as componentes principais Y_j e as variáveis X_i, onde (λ_1,e_1), (λ_2,e_2), ... ,(λ_p,e_p) são os pares de autovalores-autovetores de Σ (ANDERSON, 1958).

(4) A proporção da variância total explicada pela j-ésima componente principal é:

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + ... + \lambda_p} \text{ , } j = 1, 2, ... , p \quad (2.9)$$

Cada autovetor $e'_j = [e_{1j}, e_{2j}, ... , e_{pj}]$ pode auxiliar na interpretação da componente principal Y_j. A magnitude de e_{ij} mede a importância da i-ésima variável X_i para a j-ésima componente principal Y_j. Na realidade, e_{ij} é proporcional ao coeficiente de correlação entre Y_j e X_i.

1.3 Componentes principais populacionais de variáveis padronizadas

A j-ésima componente principal y_j das variáveis padronizadas:

$$z' = [z_1, z_2,..., z_p] = \left[\frac{X_1 - \mu_1}{\sigma_1}, \frac{X_2 - \mu_2}{\sigma_2}, \Lambda , \frac{X_p - \mu_p}{\sigma_p}\right] \quad (2.10)$$

ou, em notação matricial:

$$z' = \left(V^{1/2}\right)^{-1}(X - \mu) \quad (2.11)$$

onde: $V^{1/2} = \begin{bmatrix} \sigma_1 & 0 & \Lambda & 0 \\ 0 & \sigma_2 & \Lambda & 0 \\ M & M & M & M \\ 0 & 0 & \Lambda & \sigma_p \end{bmatrix}$ $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ M \\ \mu_p \end{bmatrix}$ $X = \begin{bmatrix} X_1 \\ X_2 \\ M \\ X_p \end{bmatrix}$

com $Cov(z) = \rho$ é dada por:

$$y_j = e'_j . z = e'_j (V^{1/2})^{-1}(X - \mu) \text{ , } j = 1, 2, ... , p \quad (2.12)$$

1.4 Propriedades das componentes principais populacionais de variáveis padronizadas

(1) $\sum_{j=1}^{p} V(y_j) = \sum_{i=1}^{p} V(z_i) = p$. (2.13)

(2) $?_{y_j z_i} = e_{ij}\sqrt{\lambda_j}$, $i, j = 1, 2, ... , p,$

onde: (λ_1,e_1), (λ_2,e_2), ... ,(λ_p,e_p) são os pares de autovalores-autovetores de ρ com $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. (2.14)

(3) A proporção da variância total explicada pela j-ésima componente principal de **z** é dada por $\frac{\lambda_j}{p}$. (2.15)

1.5 Componentes principais amostrais

Na prática, os parâmetros μ e Σ são desconhecidos e devem ser estimados. Suponha-se que $x_1, x_2, \ldots, x_n$, com $n > p$ são vetores $p \times 1$ de observações independentes de X.

As estimativas de μ e Σ são, respectivamente:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{e} \quad S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})' \qquad (2.16)$$

A j-ésima componente amostral é dada por:

$$\hat{Y}_j = \hat{e}_j' X = \hat{e}_{1j}X_1 + \hat{e}_{2j}X_2 + \ldots + \hat{e}_{pj}X_p,\ j = 1, 2, \ldots, p \qquad (2.17)$$

onde: $(\hat{\lambda}_1, \hat{e}_1), (\hat{\lambda}_2, \hat{e}_2), \ldots, (\hat{\lambda}_p, \hat{e}_p)$ são os autovalores-autovetores de S com $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$.

1.6 Propriedades das componentes principais amostrais

(1) $V(\hat{Y}_j) = \hat{\lambda}_j,\ j = 1, 2, , \ldots, p.$ (2.18)

(2) $\mathrm{Cov}(\hat{Y}_i, \hat{Y}_j) = 0,\ i \neq j.$ (2.19)

(3) $$\sum_{i=1}^{p} s_i^2 = s_1^2 + s_2^2 + \ldots + s_p^2 = \sum_{j=1}^{p} \hat{\lambda}_j = \hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_p \qquad (2.20)$$

(4) A proporção da variância total explicada pela j-ésima componente principal estimada é:

$$\frac{\hat{\lambda}_j}{\hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_p},\quad j = 1, 2, \ldots, p. \qquad (2.21)$$

(5) A correlação amostral entre $\hat{Y}_j$ e X_i é:

$$r_{\hat{Y}_j X_i} = \frac{\hat{e}_{ij}\sqrt{\hat{?}_j}}{s_i},\ i, j = 1, 2, \ldots, p. \qquad (2.22)$$

1.7 Componente principal amostral de variáveis padronizadas

Para um vetor de observações padronizadas $\hat{z} = [\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_p]$ a matriz covariância será:

$$S_Z = R = \begin{bmatrix} 1 & \hat{\rho}_{12} & \Lambda & \hat{\rho}_{1p} \\ \hat{\rho}_{21} & 1 & \Lambda & \hat{\rho}_{2p} \\ M & M & \Lambda & M \\ \hat{\rho}_{p1} & \hat{\rho}_{p2} & \Lambda & 1 \end{bmatrix}$$

A j-ésima componente principal das variáveis padronizadas será:

$$\hat{y}_j = \hat{e}'_j z_j = \hat{e}_{1j}z_1 + \hat{e}_{2j}z_2 + \ldots + \hat{e}_{pj}z_p,\ j = 1, 2, \ldots, p \qquad (2.23)$$

onde: $(\hat{\lambda}_j, \hat{e}_j)$ é o j-ésimo par autovalor-autovetor de R com $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$.

1.8 Propriedades das componentes principais amostrais de variáveis padronizadas

(1) $V(\hat{y}_j) = \hat{\lambda}_j,\ j = 1, 2, \ldots, p.$ (2.24)

(2) $\mathrm{Cov}(\hat{y}_i, \hat{y}_j) = 0,\ i \neq j.$ (2.25)

(3) Variância total amostral:

$$= \mathrm{tr}® = p = \sum_{j=1}^{p} \hat{\lambda}_j = \hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_p\ . \qquad (2.26)$$

(4) $r_{\hat{y}_j z_i} = \hat{e}_{ij}\sqrt{\hat{\lambda}_j},\ j = 1, 2, \ldots, p.$ (2.27)

(5) A proporção da variância total amostral, explicada pela j-ésima componente, será dada por $\frac{\hat{\lambda}_j}{p},\ j = 1, 2, \ldots, p.$ (2.28)

2 Modelo de Regressão Múltipla

Considere as n observações de uma variável dependente y $(n\times1)$ e p variáveis independentes X $(n\times p)$, sendo as observações centradas tais que $\bar{y}=\bar{x}_i=0$, $i=1, 2, ..., p$, então a equação de regressão é (MARDIA, KENT e BIBBY, 1982):

$$y=X\beta+\varepsilon \text{, onde } \varepsilon\sim N_n(0,\sigma^2 I). \quad (3.1)$$

Considerando que $E=XG$ denota a transformação de $\mathbf{X}$ para as componentes principais E, então a equação de regressão pode ser escrita como:

$$y=E\alpha+\varepsilon, \quad (3.2)$$

onde $\alpha=G'\beta$. Como a matriz E representa as componentes principais, suas colunas são ortogonais, e os estimadores $\hat{\alpha}_i$ permanecem inalterados se alguma coluna de E for eliminada do modelo de regressão (HAIR et al., 2005). Os estimadores de mínimos quadrados dos vetores α e ε são dados, respectivamente, por:

$$\hat{\alpha}=(E'E)^{-1}E'y \text{ e } \hat{\varepsilon}=y-E\hat{\alpha}, \quad (3.3)$$

ou

$$\hat{\alpha}_i=n^{-1}l_i^{-1}e'_{(i)}y, \quad i=1, 2, ..., p, \quad (3.4)$$

onde l_i é o i-ésimo autovalor da matriz covariância $n^{-1}XX'$. O estimador $\hat{\alpha}_i$ tem esperança e variância dadas, respectivamente, por $E(\hat{\alpha}_i)=\alpha$ e $V(\hat{\alpha}_i)=\sigma^2/nl_i$. A covariância entre e_i e y,

$$Cov(e_i, y)=n^{-1}e'_{(i)}y=n^{-1}g'_{(i)}X'y, \quad (3.5)$$

pode ser usada para testar se a contribuição de e_i é significativa para a regressão. Sob a hipótese nula, H_0: $\alpha_i=0$, a estatística (MARDIA, KENT e BIBBY, 1982):

$$\frac{\hat{\alpha}_i(nl_i)^{1/2}}{(\hat{\varepsilon}'\hat{\varepsilon}/(n-p-1))^{1/2}}=\frac{e'_{(i)}y}{(l_i n\hat{\varepsilon}'\hat{\varepsilon}/(n-p-1))^{1/2}}\sim t_{n-p-1}. \quad (3.6)$$

Como as componentes principais são ortogonais, pode-se escrever:

$$y'y=\sum_{i=1}^{p}\hat{\alpha}_i^2 e'_{(i)}e_{(i)}+\hat{\varepsilon}'\hat{\varepsilon}=\sum_{i=1}^{p}nl_i\hat{\alpha}_i^2+\hat{\varepsilon}'\hat{\varepsilon}. \quad (3.7)$$

Então, $nl_i\hat{a}_i^2/y'y=corr^2(y,e_{(i)})$ representa a proporção da variância de y explicada pela i-ésima componente.

Selecionar aquelas componentes para as quais a estatística em (3.6) é significativa constitui um método consistente de escolha das componentes principais a serem retidas.

3 Resultados

Considere as variáveis explicativas (independentes): receita com impostos, receita com contribuições, gasto com pessoal, gasto com juros, gasto com investimentos, transferências para estados e municípios e a variável resposta (dependente) benefícios previdenciários, para o Brasil, de 1980 a 2003, conforme mostra a tabela 1.

Denomine as variáveis de: X_1 = receita com impostos, X_2 = receita com contribuições, X_3 = gasto com pessoal, X_4 = gasto com juros, X_5 = gasto com investimentos, X_6 = transferências para estados e municípios e Y = benefícios previdenciários.

O desenvolvimento computacional foi todo realizado com uso do software Matlab, versão 5.3.

Deve-se ressaltar a limitação do método, quando aplicado aos dados da tabela 1, a qual envolve dados de uma série temporal, não havendo uma maior preocupação com os fenômenos econômicos envolvidos.

TABELA 1 - RECEITAS, GASTOS, TRANSFERÊNCIAS E BENEFÍCIOS - BRASIL - 1980-2003

(Em R$ bilhões)

ANO	X_1	X_2	X_3	X_4	X_5	X_6	Y
1980	108,6	4,8	13,7	6,7	10,0	12,3	9,9
1981	99,6	5,0	15,7	2,9	18,6	11,4	9,6
1982	103,0	5,9	13,8	4,2	12,5	23,9	11,0
1983	91,9	25,0	11,7	5,1	8,5	21,4	9,2
1984	88,6	25,8	9,9	3,7	6,1	22,1	8,2
1985	107,9	21,8	13,8	10,7	8,8	32,3	10,9
1986	120,3	23,3	14,4	35,2	18,3	40,5	11,8
1987	109,8	21,0	14,8	10,0	21,7	40,0	12,6
1988	105,0	25,8	18,9	38,3	18,9	37,2	14,4
1989	93,4	32,4	28,7	89,8	9,0	36,0	21,3
1990	103,2	114,2	77,6	35,8	13,5	40,8	94,1
1991	81,0	97,4	59,2	6,1	15,4	34,8	85,1
1992	85,9	97,7	55,6	27,5	12,5	36,6	83,2
1993	92,2	111,6	61,4	35,0	17,4	42,8	64,5
1994	111,3	124,9	73,1	39,4	15,3	68,0	52,3
1995	119,8	143,3	95,0	42,0	12,0	54,2	81,6
1996	120,8	178,9	92,3	44,0	13,0	57,9	93,6
1997	122,6	194,1	93,1	44,0	15,8	97,1	61,4
1998	136,6	188,7	96,5	62,0	16,7	75,4	107,7
1999	135,0	205,6	93,2	82,0	12,6	75,7	105,1
2000	125,0	222,0	92,5	61,7	16,0	81,9	102,8
2001	131,9	232,0	94,2	76,0	21,0	86,1	107,7
2002	136,8	245,5	95,2	70,0	12,8	93,3	111,0
2003	124,0	243,4	85,0	70,7	6,9	86,3	116,8

FONTES: Tesouro Nacional e Ministério da Fazenda

NOTA: Valores atualizados com base no IGP-DI.

A matriz de correlações resultante para as variáveis explicativas foi:

$$R = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \end{matrix} \\ \begin{bmatrix} 1{,}00 & 0{,}70 & 0{,}63 & 0{,}64 & 0{,}22 & 0{,}78 \\ 0{,}70 & 1{,}00 & 0{,}95 & 0{,}73 & 0{,}08 & 0{,}93 \\ 0{,}63 & 0{,}95 & 1{,}00 & 0{,}69 & 0{,}13 & 0{,}86 \\ 0{,}64 & 0{,}73 & 0{,}69 & 1{,}00 & 0{,}05 & 0{,}76 \\ 0{,}22 & 0{,}08 & 0{,}13 & 0{,}05 & 1{,}00 & 0{,}18 \\ 0{,}78 & 0{,}93 & 0{,}86 & 0{,}76 & 0{,}18 & 1{,}00 \end{bmatrix} \end{array} \begin{matrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{matrix}$$

tendo autovetores normalizados associados dados pela matriz G, onde a primeira coluna representa o autovetor associado ao maior autovalor, a segunda coluna representa o autovetor associado ao segundo maior autovalor e, assim, sucessivamente.

$$G = \begin{bmatrix} 0{,}83 & 0{,}13 & 0{,}45 & -0{,}29 & -0{,}09 & 0{,}00 \\ 0{,}96 & -0{,}12 & -0{,}22 & -0{,}08 & 0{,}02 & 0{,}13 \\ 0{,}92 & -0{,}07 & -0{,}34 & -0{,}05 & -0{,}17 & -0{,}08 \\ 0{,}84 & -0{,}15 & 0{,}22 & 0{,}47 & -0{,}04 & 0{,}00 \\ 0{,}19 & 0{,}97 & -0{,}08 & 0{,}10 & -0{,}00 & 0{,}01 \\ 0{,}94 & 0{,}01 & -0{,}03 & -0{,}06 & 0{,}25 & -0{,}06 \end{bmatrix}$$

Os autovalores associados e respectivas variações explicadas das componentes principais estão resumidos na tabela 2.

TABELA 2 - PERCENTUAL DA VARIACÃO TOTAL EXPLICADA PELAS COMPONENTES PRINCIPAIS

COMPONENTE PRINCIPAL	AUTOVALOR	% EXPLICADA	% EXPLICADA ACUMULADA
1ª	4.1086	68.48	68.48
2ª	1.0056	16.76	85.24
3ª	0.4248	7.08	92.32
4ª	0.3316	5.53	97.84
5ª	0.1032	1.72	99.56
6ª	0,0261	0,44	100,00

FONTE: Os autores

As correlações resultantes entre as componentes principais e as variáveis originais estão resumidas na tabela 3.

TABELA 3 - CORRELACÕES ENTRE AS COMPONENTES PRINCIPAIS E AS VARIÁVEIS ORIGINAIS

VARIÁVEL ORIGINAL	COMPONENTES PRINCIPAIS					
	1ª	2ª	3ª	4ª	5ª	6ª
X_1	0,83	0.13	0.45	-0.29	-0.09	0,00
X_2	0,96	-0.12	-0.22	-0.08	0.02	0.13
X_3	0,92	-0.07	-0.34	-0.05	-0.17	-0.08
X_4	0,84	-0.15	0.22	0.47	-0.04	0,00
X_5	0,19	0.97	-0,08	0.10	-0.00	0,01
X_6	0,96	0,01	-0,03	-0,06	0,25	-0,06

FONTE: Os autores

Os escores correspondentes às componentes principais estão resumidos na tabela 4, gerando esses dados a matriz E.

TABELA 4 - ESCORES PARA AS COMPONENTES PRINCIPAIS

e_1	e_2	e_3	e_4	e_5	e_6
-2.3558	-0.5149	0.7009	-0.6091	-0.4159	0.0667
-2.4320	1.3687	0.0050	-0.1081	-0.3406	0.1811
-2.2583	0.0182	0.3420	-0.4617	0.0204	-0.0598
-2.5715	-0.9995	-0.0340	-0.2757	0.1834	0.1079
-2.7313	-1.5571	-0.0980	-0.3138	0.2996	0.0875
-1.8866	-0.8317	0.6573	-0.6213	0.1888	-0.0868
-0.8466	1.2779	1.1733	0.0791	0.0903	-0.0217
-1.4196	2.0972	0.3350	-0.2135	0.3288	-0.0162
-1.1488	1.2619	0.4873	0.6541	0.1609	-0.0075
-0.7435	-1.3655	0.7515	2.1090	-0.0026	-0.1752
-0.0686	-0.1914	-0.6543	0.1257	-0.4899	-0.1121
-1.4403	0.2757	-1.6568	0.0805	0.0711	0.0519
-1.0769	-0.4460	-1.0606	0.4544	0.0227	0.0521
-0.4441	0.6404	-0.9938	0.6247	-0.0199	0.0998
0.7127	0.2597	-0.3429	-0.0503	0.2410	-0.3149
1.0120	-0.5013	-0.2275	-0.3887	-0.6217	-0.2772
1.3130	-0.3208	-0.2938	-0.3818	-0.4718	0.0449
2.2163	0.3107	-0.4294	-0.5016	0.6527	-0.3522
2.4684	0.5259	0.3526	-0.2671	-0.3508	-0.0912
2.6926	-0.5343	0.6296	0.1924	-0.3200	0.0529
2.4126	0.2410	-0.1945	-0.0217	0.1204	0.1495
3.0550	1.3316	0.0546	0.3604	0.0396	0.2523
3.1191	-0.4657	0.3431	-0.3454	0.2217	0.1264
2.4224	-1.8807	0.1535	-0.1206	0.3918	0.2419

FONTE: Os autores

Os parâmetros estimados pela expressão (3.3) resultaram em:

$$\hat{\alpha}_1 = 17{,}8649 \;\; ; \;\; \hat{\alpha}_2 = -5{,}3395 \;\; ; \;\; \hat{\alpha}_3 = -24{,}7003 \;\; ;$$
$$\hat{\alpha}_4 = 0{,}0101 \;\; ; \;\; \hat{\alpha}_5 = -25{,}0773 \;\; \text{e} \;\; \hat{\alpha}_6 = 44{,}7734 \;\; .$$

O cálculo da estatística dada pela expressão (3.6) permite testar qual ou quais componentes principais têm contribuição significativa para a regressão. O resultado das estatísticas estão resumidos na tabela 5, de acordo com a qual se verifica que apenas a primeira componente principal é significativa para a regressão estudada.

TABELA 5 - ESTATÍSTICA PARA TESTAR A SIGNIFICÂNCIA DAS COMPONENTES PRINCIPAIS PARA A REGRESSÃO

COMPONENTE PRINCIPAL	ESTATÍSTICA t	VALOR-p
1ª	4,9635	*0.0000
2ª	-0,2568	0,4002
3ª	-0,4097	0,3436
4ª	0,0001	0,5000
5ª	-0,0783	0,4693
6ª	-0,0323	0,4873

FONTE: Os autores

Considerando-se a regressão com apenas a primeira componente principal, que é a única significativa para a regressão, como variável explicativa, resulta em um coeficiente de determinação $R^2 = 0{,}7285$. Suponha-se que se deseja retornar às variáveis explicativas originais, e, nesse caso, de acordo com a tabela 3, as variáveis que se correlacionam mais fortemente com a primeira componente principal são: X_1, X_2, X_3, X_4 e X_6, ficando de fora apenas a variável X_5, resultando em um coeficiente de determinação $R^2 = 0{,}9496$. Se ainda fossem consideradas apenas as variáveis com correlação acima de 0,90, ou seja, retendo apenas as variáveis explicativas: X_2, X_3 e X_6, o coeficiente de determinação resultante seria $R^2 = 0{,}9463$.

Conclusão

O método de descarte de variáveis explicativas pelo uso de componentes principais não é o único existente, por ser comum a aplicação dos métodos de regressão que envolvem a análise do coeficiente de determinação. No entanto, o método das componentes principais permite uma redução significativa no número de variáveis, fundamentalmente quando se tem um significado adequado para a componente retida, a qual pode ser tratada como a nova variável explicativa.

A aplicação desse método é adequado principalmente nos casos envolvendo um número muito grande de variáveis explicativas em que as componentes principais têm uma interpretação significativa para o pesquisador. A substituição das variáveis explicativas originais pelas componentes principais retidas proporciona um modelo com uma redução substancial no número de variáveis explicativas.

As componentes principais no descarte de variáveis em um modelo de regressão múltipla

- Recebido em: 07.04.2005
- Aprovado em: 30.05.2005

Referências

ANDERSON, T. W. **An introduction to multivariate statistical analysis**. New York: John Wiley, 1958.

BOUROCHE, J.-M.; SAPORTA, G. **Análise de dados**. Rio de Janeiro: Zahar Editores, 1982.

CHATFIELD, C.; COLLINS, A. J. **Introduction to multivariate analysis**. London: Chapman & Hall, 1992.

HAIR JR., J. F. et al. **Análise multivariada de dados**. Porto Alegre: Bookman, 2005.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. Englewood Cliffs: Prentice-Hall, 1988.

MARDIA, K.; KENT, J.; BIBBY, J. **Multivariate analysis**. London: Academic Press, 1982.

TESOURO NACIONAL E MINISTÉRIO DA FAZENDA. **Receitas, gastos, transferências e benefícios**. Brasil, 1980-2003.